ENCYKLOPEDIA FIZYKI WSPÓŁCZESNEJ

ENCYKLOPEDIA FIZYKI WSPÓŁCZESNEJ

PANSTWOWE WYDAWNICTWO NAUKOWE · WARSZAWA

ISBN 83-01-00391

Państwowe Wydawnictwo Naukowe
Wydanie pierwsze. Nakład 59.750+250 egzemplarzy
Arkuszy wydawniczych 181,0; drukarskich 126,0
Tablice: barwne 8 str., czarne 56 str.
Tekst złożono na monotypie krojem pisma Times
Druk z form fotopolimerowych na maszynach typu „Rotafolio X"
Papier na tekst drukowy satynowany klasy III 80 g, 86 × 112 cm
Papier na wkładki barwne kredowany klasy III 90 g, 61 × 86 cm
Papier na wkładki czarne rotograwiurowy klasy III 90 g, 61 × 86 cm
Oddano do składania w styczniu 1979 roku
Podpisano do druku w lipcu 1982 roku
Druk ukończono w styczniu 1983 roku
Zamówienie nr 569/79.
Zakłady Graficzne „Dom Słowa Polskiego" w Warszawie

Przedmowa

Encyklopedia fizyki współczesnej jest pierwszą w Polsce próbą przeglądu zagadnień fizyki współczesnej na poziomie, który umożliwi korzystanie z niej tym wszystkim czytelnikom, którzy fizykę i matematykę znają w zakresie szkoły średniej. Nie jest to encyklopedia w tradycyjnym znaczeniu tego słowa. Nie ma w niej bowiem haseł definicyjnych dotyczących poszczególnych pojęć, zjawisk i praw, są natomiast dłuższe artykuły przeglądowe, obejmujące pewną całość tematyczną (np. Cząstki elementarne, Magnetooptyka, Kwasy nukleinowe). Artykuły są pogrupowane tematycznie w działy. Czytelnik, który chciałby sobie wyjaśnić jakieś pojęcie czy termin, może korzystać ze szczegółowego skorowidza terminów, pojęć i wzorów, ułatwiającego znalezienie odpowiedniego miejsca w artykule przeglądowym.

Fizykę współczesną stanowi zespół zagadnień, które w obecnej dobie są tematem zainteresowań i badań fizyków. W *Encyklopedii* zostały omówione przede wszystkim zagadnienia najciekawsze, bądź ze względu na ich wartość poznawczą, bądź też ze względu na nowe i ważne zastosowania w fizyce szeroko pojętej, w technice i w życiu codziennym.

Spotykamy się często z poglądem, że na pograniczu fizyki i innych nauk przyrodniczych powstało szereg specjalności, jak astrofizyka, biofizyka, geofizyka, chemia kwantowa itd. Piszącemu te słowa wydaje się, że są to właściwie nowe gałęzie fizyki, która w postępującym od dawna procesie integracji wchłania dyscypliny uważane przez jakiś czas za niezależne gałęzie wiedzy. I tak np., gdy w latach trzydziestych naszego stulecia dzięki rozwojowi fizyki jądrowej udało się wskazać źródło energii gwiazd, częścią fizyki stała się astrofizyka, natomiast część astronomii zajmująca się badaniem ruchu ciał niebieskich stała się częścią fizyki jeszcze w czasach Newtona. Podobnie, gdy dzięki mechanice kwantowej udało się zrozumieć budowę atomów i cząsteczek oraz istotę wiązań chemicznych przez sprowadzenie procesów chemicznych do kwantowych elektromagnetycznych oddziaływań elektronów, gałęzią fizyki stała się w znacznej mierze chemia.

Wracamy więc powoli do pierwotnego znaczenia fizyki jako wszechogarniającej nauki o całej przyrodzie (gr. *physis* 'natura'). W taki właśnie sposób pojmowano fizykę przez stulecia. Na przykład typowy siedemnastowieczny podręcznik fizyki *Traité de physique* Jacquesa Rohaulta, wydany w Paryżu w 1671 r., zawiera nie tylko fizykę sensu stricto, lecz także kosmologię, astronomię, meteorologię, geografię, fizjologię i medycynę. Z piśmiennictwa polskiego przypominamy piękną definicję fizyki z gimnazjalnego podręcznika fizyki wydanego w Wilnie w 1825 r., napisanego przez Feliksa Drzewińskiego, profesora Uniwersytetu Wileńskiego:

„Fizyka nazwisko nauki pochodzące od wyrazu greckiego *Physis*, natura, przyrodzenie, oznacza *naukę poznawania natury*. Przyrodzeniem, albo naturą zmysłową, zowiemy to wszystko cokolwiek działa na zmysły nasze, i sprawuje w nas czucie. Wszystkie więc rzeczy nas otaczające, których się dotykamy, na które patrzymy, to co słyszymy, co działa na zmysły smaku, i powonienia, wszystko to stanowi naturę zmysłową, i poznawanie tego wszystkiego do Fizyki należy. Przeto Fizyka jest nauką poznawania rzeczy świat składających, badania i dochodzenia ich własności.

Rzeczy składające świat fizyczny, to jest zmysłom naszym dostępny, są niezmiernie liczne, i rozmaitémi własnościami obdarzone. Cała ziemia i części ją składające, istoty na niey i wewnątrz jey umieszczone, niebo i to co na niém postrzegamy, powietrze w którém żyjemy, budowa nas samych i innych podobnych nam jestestw, to wszystko stanowi świat zmysłowy. W epokach zaczęcia nauk, badania tych wszystkich rzeczy stanowiły jedną naukę, mającą ogólne nazwisko Fizyki: taką miały fizykę starożytne narody Egipcyan, Greków i Rzymian.

Poznawaniem coraz bliżey tylu rozlicznych przedmiotów, i odkrywaniem w nich coraz więcej nowych własności, gdy nauka z czasem uzrosła, i stała się bardzo obszérną, uczeni wiekow późnieyszych podzielili ją na wiele odnog, albo części, z których głównieysze są następujące: lód Nauka uważania nieba, albo raczey położeń, ruchow, i postaci brył świątłych widzianych, w przestrzeni niebios; tę część fizyki Astronomiją nazywamy. 2re Oznaczanie kształtów istot znaydujących się na ziemi i w ziemi, w wodzie i powietrzu, należy do oddziału Fizyki, zwanego *Historyą nataralną*. 3cie Uważanie wewnętrzney budowy tych istot, stanowi naukę zwaną *Anatomiją*. 4te Dochodzenie pierwiastków, czyli nayprostszych części, z jakich się składają istoty w naturze, i na jakie ostatecznie rozebrać się, rozdzielić, albo rozłożyć mogą, jest przedmiotem *Chemii*."

Wzorem naszych przodków sprzed 150 lat przedstawiamy zatem Czytelnikom encyklopedię traktującą o tak szeroko pojętej fizyce, a także o jej wielorakich związkach z techniką.

Jak powiedzieliśmy wyżej, zakładamy u Czytelnika w zasadzie znajomość fizyki i matematyki na poziomie szkoły średniej. Nie ma więc w *Encyklopedii* wyjaśnienia takich pojęć i zagadnień, jak np. siła, prędkość, prawo Ohma, prawo Coulomba, co nie oznacza, że elementarne pojęcia nie są czasem przypominane w kontekście omawiania innych zagadnień. *Encyklopedia fizyki współczesnej* stanowi w pewnym sensie uzupełnienie trzytomowej *Ekcyklopedii fizyki* w tradycyjnej formie, jaką wydało Państwowe Wydawnictwo Naukowe przed kilku laty.

Encyklopedię otwiera kilka artykułów wprowadzających w podstawowe pojęcia, prawa i strukturę fizyki oraz charakteryzujących fundamentalne wielkie teorie fizyki współczesnej. Następne artykuły są uporządkowane w działy traktujące o obiektach fizycznych w porządku wzrastających rozmiarów i złożoności, od cząstek elementarnych, przez jądro atomowe, atom, cząsteczkę do złożonych struktur biologicznych, za którymi następują działy obejmujące artykuły o Ziemi jako planecie oraz o Wszechświecie.

Pomysł opracowania *Encyklopedii fizyki współczesnej* narodził się w końcu 1973 r. We wstępnym stadium opracowania założeń programowych Państwowe Wydawnictwo Naukowe uzyskało obszerne wypowiedzi ankietowe od kilkudziesięciu wybitnych specjalistów z różnych dziedzin fizyki. Zebrane cenne uwagi pozwoliły na opracowanie wstępnego zestawu artykułów hasłowych, który w trakcie pracy nad *Encyklopedią* nabrał ostatecznego kształtu. Dodajmy, że w ciągu kilku lat przygotowywania artykułów następował postęp w wielu działach fizyki tu omawianych, co nakładało na autorów poważny obowiązek uaktualniania tekstów. *Encyklopedia fizyki współczesnej*, pierwsza publikacja tego typu na naszym rynku księgarskim, mogła powstać tylko dzięki życzliwemu stosunkowi i pomocy wielu osób, których nie sposób wymienić, a którym składamy w tym miejscu gorące podziękowanie. Ostateczny osąd wartości dzieła należy do Czytelników, którym życzymy przyjemnej i pożytecznej lektury.

Andrzej Kajetan Wróblewski

Spis treści

CZYM JEST FIZYKA — Józef Werle

PODSTAWOWE POJĘCIA I TEORIE

CZĄSTKI ELEMENTARNE I FIZYKA WIELKICH ENERGII

FIZYKA JĄDRA ATOMOWEGO

FIZYKA ATOMU, CZĄSTECZKI I CIAŁA STAŁEGO

ELEKTRONIKA WSPÓŁCZESNA

AKUSTYKA

BIOFIZYKA I FIZYKA MEDYCZNA

FIZYKA ZIEMI

ASTROFIZYKA I KOSMOCHEMIA

Autorzy

Antoni Adamczyk
Tadeusz Bartczak
Czesław Bazan
Grzegorz Białkowski
Oskar Chomicki
Jacek Chrostowski
Stefan Czarnecki
Dymitr Czernawski
Jerzy Czerwonko
Edward Czuryło
Piotr Decowski
Marek Demiański
Jerzy Dera
Renata Dmowska
Czesław Druet
Zbigniew Fajklewicz
Kazimierz Fidelus
Tadeusz Figielski
Magdalena Fikus
Leszek Filipczyński
Jacek Furdyna
Robert R. Gałązka
Ryszard Gawroński
Waldemar Gorzkowski
Andrzej Graja
Marek Gutowski
Krzysztof Haman
Michał Heller
Marian A. Herman
Bożena Hilczer
Andrzej Hrynkiewicz
Lech Jakubowski
Wiktor Jungowski
Andrzej J. Kalestyński
Magdalena Kądziałko-Hofmokl
Krzysztof Kędzior
Tadeusz Kłopotowski
Janusz Konopka
Wojciech Kopczyński
Marek Kosmal
Zofia Kosturkiewicz
Marcin Kubiak
Bronisław Kuchowicz
Adam Kujawski
Marian Kupczyński
Józef Stanisław Kwiatkowski
Henryk Lachowicz
Janusz Leciejewicz
Czesław Lewa
Bogumił Linde
Jacek Łagowski
Ignacy Malecki
Jerzy Mielnicki
Janusz Mika
Andrzej Morawski
Adam Morecki
Jan Mostowski
Bogusław Mrygoń
Wacław Nazarewicz
Romuald Pawluczyk
Stanisław Piasecki
Jerzy Pniewski
Jerzy Prochorow
Stanisław Przestalski
Michał Różyczka
Marek Sadowski
Stanisław Sikorski
Tadeusz Skaliński
Ewa Skrzypczak
Adam Sobiczewski
Zdzisław Sołtys
Ryszard Sosnowski
Jan Stankowski
Andrzej Staruszkiewicz
Kazimierz Stępień
Lech Stolarczyk
Hanna Strzelecka-Gołaszewska
Tadeusz Suski
Krystyna Szczepaniak
Mieczysław Szustakowski
Andrzej Szymacha
Zdzisław Szymański
Henryk Szymczak
Rita Szymczak
Antoni Śliwiński
Jarosław Świderski
Michał Święcki
Eugeniusz Trojnar
Zygmunt Trzaska Durski
Bohdan Utrysko
Wiesław Wardzyński
Tadeusz Warmiński
Jerzy Wehr
Józef Werle
Andrzej Wernik
Kazimierz Wierzchowski
Iwona Wojciechowska
Wojciech Wójcik
Janusz Zakrzewski
Kazimierz Zakrzewski
Andrzej Zawadzki
Przemysław Zieliński
Stanisław Zięba

Fotografie

Marek Bączkowski
il. 14, 16 i 17
Zygmunt Fiałkowski
il. 42–45, 54–56 i 104
Aleksander Kuc
il. 140–144
Stefan Nargiełło
il. 63b
Tadeusz Warmiński
il. 84–86
Krzysztof Wojciechowski
il. 1, 13 i 23
Tomasz Zaremba
il. 66b i rys. 7 str. 502

Rysunki

Krzysztof Figielski
str. 70 (margines), rys. 3 str. 72
i rys. 3 str. 461
Edward Lutczyn
str. 689–694 (marginesy)

CZYM JEST FIZYKA

Józef Werle

Fundamentalne zasady przyrodoznawstwa · Metoda naukowa fizyki · Zarys rozwoju fizyki · Praktyczne i kulturotwórcze znaczenie fizyki

Fundamentalne zasady przyrodoznawstwa

Fizyka jest najbardziej podstawową nauką przyrodniczą, zajmującą się badaniem strukturalnych i dynamicznych praw przyrody. Jednym z najważniejszych zadań fizyki jest poznanie fundamentalnych i uniwersalnych własności materii wspólnych wszystkim lub przynajmniej wielu różnym jej formom. Z zadania tego wynika podstawowa, integrująca całe przyrodoznawstwo rola fizyki.

Możliwość wykonania tego zadania przez fizykę opiera się na dwóch niezwykle ważnych i płodnych ogólnych zasadach: jednolitości materii i powszechności praw fizyki. W rozumieniu większości fizyków pierwsza zasada odnosi się do strukturalnych, druga — do bardziej dynamicznych własności materii. Zgodnie ze współczesną fizyką można by im nadać następujące brzmienie:

jednolitość materii

I. Wszystkie ciała materialne, nieożywione i ożywione, na Ziemi i gdziekolwiek indziej we Wszechświecie są zbudowane z takich samych elementarnych składników (np. odpowiednich cząstek i pól).

powszechność praw fizyki

II. Oddziaływania między elementarnymi składnikami materii i wynikające z nich własności materii oraz prawa fizyki są wszędzie takie same.

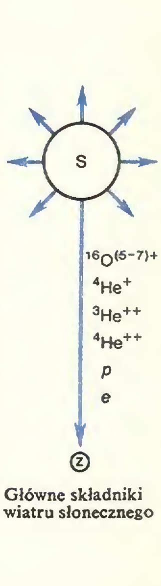
Obie te zasady są dziś wszechstronnie potwierdzone przez doświadczenie. Oczywiście nie oznaczają one bynajmniej, że świat jest pod każdym względem jednorodny (co byłoby jawnie sprzeczne z naszym doświadczeniem), zasady te nie wykluczają bowiem występowania materii w różnych formach i stanach.

Zilustrujemy to na przykładzie Ziemi i Słońca. Materia we wnętrzu Słońca i na Ziemi występuje niewątpliwie w bardzo różnych stanach. Różnice własności materii ziemskiej i słonecznej wynikają z ogromnych różnic wartości takich parametrów stanu jak temperatura, ciśnienie, natężenie pola grawitacyjnego, pola promieniowania itd. Niemniej jest to taka sama materia, podlegająca tym samym prawom. Należy to rozumieć tak, że np. materia ziemska o odpowiednim składzie chemicznym, poddana takiemu samemu jak we wnętrzu Słońca ciśnieniu, temperaturze, promieniowaniu itp., będzie się zachowywała identycznie jak materia słoneczna. Weźmy inny przykład: Organizmy żywe zbudowane są z takich samych atomów, między którymi występują takie same oddziaływania i reakcje jak te, które badają fizycy i chemicy w laboratoriach w warunkach wprawdzie sztucznych, ale za to dobrze kontrolowanych. Oczywiście występujące w żywych komórkach bardzo duże drobiny przysparzają kłopoty wskutek skomplikowanej budowy i wynikających stąd specyficznych własności, ale jest to komplikacja w pełni zrozumiała na gruncie fizyki.

Całe współczesne przyrodoznawstwo opiera się więc na założeniu, że wszelkie, nieraz bardzo skomplikowane struktury i procesy chemiczne, biologiczne, geologiczne i astronomiczne są wynikiem działania uniwersalnych praw fizyki. Fundamentalne prawa fizyki są względnie proste i ogólne, ale zastosowane do różnych złożonych układów i sytuacji mogą implikować bardzo różnorodne i skomplikowane własności szczególne.

Weryfikacja tego podstawowego założenia współczesnego przyrodoznawstwa może być w pewnych wypadkach dość prosta, ale na ogół jest bardzo trudna, gdyż wymaga długiego łańcucha obserwacji, wnioskowań i rachunków. Bezpośrednie sprawdzenie pewnych konsekwencji jest czasem praktycznie niemożliwe. Nie potrafimy np. przenieść badanego kawałka materii ziemskiej do wnętrza Słońca w celu sprawdzenia, czy będzie się ona tam rzeczywiście zachowywała tak jak materia słoneczna. Jednakże identyczność materii na Ziemi i na Słońcu można potwierdzić w inny sposób, który wprawdzie nie jest bezpośredni, ale mimo to jest przekonywający. A mianowicie — w widmie promieniowania Słońca występują linie absorpcyjne charakterystyczne dla znanych również na Ziemi pierwiastków. Ba, jeden z nich, hel, został właśnie w ten sposób odkryty najpierw na Słońcu, a dopiero potem został znaleziony na Ziemi. Dalej — Słońce promieniuje nie tylko fale elektromagnetyczne, lecz także różne cząstki, które można dziś dobrze zidentyfikować (np. przy użyciu odpowiednich detektorów umieszczonych na sputnikach). Okazuje się, że są one identyczne z cząstkami poznanymi przez fizyków w ich laboratoriach na Ziemi.

Można podać tysiące faktów, zarówno z zakresu właściwej fizyki, jak i chemii, biologii, geologii i astronomii, potwierdzających zasady jednolitości materii i powszechności praw fizyki, natomiast nie znamy naukowo potwierdzonych faktów sprzecznych z nimi. Zasady te okazały się niezmiernie ważne, płodne i stymulujące dla wszystkich nauk przyrodniczych. Bez nich wiele tych nauk nie mogłoby po prostu istnieć albo ograniczyłoby się do bardzo powierzchownych obserwacji i czysto opisowego podejścia. Na przykład dopiero przyjęcie zasady jednolitości materii i powszechności praw przyrody pozwala na wykorzystanie licznych wiadomości z zakresu fizyki jądrowej i cząstek elementarnych, teorii promieniowania, termodynamiki, teorii grawitacji, mechaniki itd. w celu wyjaśnienia obserwowanych zjawisk astronomicznych.

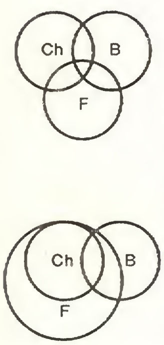


Główne składniki wiatru słonecznego

Warto zwrócić uwagę na zmiany, jakie w pojmowaniu obu zasad zachodziły w toku rozwoju nauki. Rodowód tych zasad jest niewątpliwie bardzo stary i sięga z pewnością starożytnych greckich filozofów

przyrody, a może czasów jeszcze dawniejszych. Nie wdając się tu w szczegółowe historyczne rozważania, należy zwrócić uwagę na bardzo istotny przełom, który się dokonał w ostatnich 100 latach. Otóż ugruntowany już bardzo dawno i właściwie nie zmieniany przez długie wieki sens obu zasad był pierwotnie znacznie skromniejszy. Polegały one właściwie tylko na stwierdzeniu, że wszystkie ciała materialne mają pewne wspólne cechy, zwane fizycznymi, np. rozciągłość, położenie, prędkość, masę, pęd, temperaturę, barwę. Było oczywiste, że ze względu na występowanie tych cech wszystkie ciała materialne muszą oczywiście podlegać odpowiednim prawom fizycznym. Uważano jednak, że oprócz cech fizycznych ciała materialne mogą posiadać zupełnie niezależne od praw fizyki własności chemiczne, biologiczne (witalne), astralne czy spirytualne. Przez długie wieki uważano, że zjawiska fizyczne, chemiczne, biologiczne, geologiczne i astronomiczne są wywoływane przez specyficzne, odrębne i niezależne siły fizyczne, chemiczne, witalne, geologiczne, astralne itd., podlegające niezależnym prawom. Dziś jesteśmy głęboko przekonani, że nie ma potrzeby przyjmowania istnienia takich odrębnych sił i praw. Wszystkie te zjawiska staramy się dziś sprowadzić — na ogół z dobrym skutkiem — do odpowiednich zjawisk i praw czysto fizycznych, zachodzących w specyficznych dla danej gałęzi nauki układach (obiektach, procesach, sytuacjach, stanach). Współczesne przyrodoznawstwo bardzo pogłębiło więc znaczenie obu zasad, rozciągając je na wszystkie niemal obiekty i zjawiska przyrody.

Współczesna fizyka łączy i zespala w jedną całość wszystkie nauki przyrodnicze. Chemia, geologia i astronomia oraz pokrewne, bardziej szczegółowe nauki przyrodnicze, jak np. krystalografia, mineralogia, meteorologia, są dziś uważane po prostu za działy fizyki zajmujące się badaniem pewnych szczególnych klas obiektów za pomocą uniwersalnych metod, pojęć i praw fizyki. W naszych oczach również biologia staje się coraz bardziej „fizyką materii ożywionej", czy krócej — „fizyką życia". Postępująca coraz głębiej i szerzej integracja wszystkich nauk przyrodniczych na bazie fizyki jest jedną z najbardziej charakterystycznych cech nowożytnej nauki.



Relacje między chemią, biologią i fizyką dawniej i dziś

Metoda naukowa fizyki

Zasady jednolitości materii i powszechności praw fizyki otwierają wprawdzie bardzo szerokie horyzonty, ale są same w sobie zbyt ogólnikowe. Konkretne zastosowanie tych zasad w pracy badawczej wymaga — z jednej strony — znajomości odpowiednich zjawisk, praw i teorii fizycznych, z drugiej — stosowania w całym przyrodoznawstwie tego samego języka i tej samej metody badawczej. Ta ostatnia sprawa jest szczególnie ważna, warto się więc zastanowić nieco dokładniej nad metodą naukową fizyki, ponieważ stała się ona dziś metodą wszystkich nauk przyrodniczych.

Otóż najważniejsze i najbardziej charakterystyczne cechy metody naukowej fizyki są następujące:
— możliwie szerokie stosowanie eksperymentu,
— ścisłe, ilościowe formułowanie wyników eksperymentu i wniosków ogólnych,
— poszukiwanie praw przyczynowych,
— silne obustronne sprzężenie doświadczenia (eksperymentu) z teorią.

Obserwacja i eksperyment

Fizyka jest nauką o realnych faktach przyrodniczych, które możemy poznać jedynie (bezpośrednio lub pośrednio) za pomocą naszych zmysłów. Fakty takie nazywamy w fizyce zjawiskami. Badanie zjawisk przyrody może polegać na ich biernym oglądaniu w warunkach naturalnych, w których nie mamy wpływu na oglądane zjawiska. Ograniczenie się do biernego oglądania, zwanego często obserwacją, jest konieczne, jeśli obiekty obserwacji są bardzo wielkie lub bardzo odległe w przestrzeni lub czasie, jak to ma miejsce np. w astronomii lub geologii.

obserwacja

Wiele zjawisk zachodzących w mniejszej skali można jednak wywołać sztucznie, umyślnie, w warunkach laboratoryjnych. Takie postępowanie nazywamy eksperymentowaniem. Czynny eksperyment ma oczywiście wiele zalet. Przede wszystkim jest znacznie bardziej elastyczny niż bierna obserwacja, da się powtórzyć na życzenie i łatwiej go kontrolować ilościowo. Możliwość powtórzenia zjawiska dowolną liczbę razy, w ilościowo identycznych warunkach, pozwala na usunięcie przypadkowych lub nawet systematycznych błędów, na zwiększenie dokładności i pewności wyników badań, na uściślenie, sprawdzenie i korygowanie dotychczasowej wiedzy. Wykonanie eksperymentów w różnych, celowo zmienionych warunkach pozwala na bardziej wszechstronne, pełniejsze i głębsze poznanie badanych zjawisk. Zastąpienie biernej obserwacji przez czynny eksperyment przynosi więc zawsze istotny postęp w pracy badawczej.

eksperymentowanie

Zjawiska powtarzające się często, ale tylko w warunkach naturalnych, trudniej jest kontrolować, gdyż na ogół nie możemy być pewni identyczności tych warunków, nie możemy też dobrze określić wartości parametrów, zmieniających się za każdym razem.

Jeszcze gorzej przedstawia się sprawa ze zjawiskami niepowtarzalnymi lub bardzo rzadkimi. Można je również badać za pomocą odpowiednich metod fizycznych, ale wiedzy o takich zjawiskach nie możemy pogłębiać, sprawdzać ani korygować przez powtórzenie ich i wykonanie nowych badań.

Fizyka właściwa zajmuje się niemal wyłącznie zjawiskami, które można badać eksperymentalnie w warunkach laboratoryjnych. Inne nauki przyrodnicze, jak np. astronomia, geologia oraz niektóre działy biologii, muszą się z natury rzeczy ograniczać do biernych obserwacji. W astronomii układu planetarnego era czynnego eksperymentowania zaczęła się bardzo niedawno, dopiero wówczas, gdy zastosowano rakiety kosmiczne.

Wielkości fizyczne

Metoda naukowa fizyki jest ścisła, tj. ilościowa. Oznacza to przede wszystkim, że każde zjawisko starają się fizycy opisać za pomocą odpowiednich mierzalnych cech, zwanych wielkościami fizycznymi. Metoda ta okazała się niezwykle skuteczna i płodna. Stopniowo przyjmowały ją i nadal w większym lub mniejszym zakresie przyjmują inne nauki empiryczne. Wiele ludzi, w tym spora liczba naukowców, broni się przed wprowadzaniem metod ilościowych. Uważają oni, że ilościowe podejście gubi istotne różnice jakościowe, które nie dają się zmierzyć, ponieważ możemy porównywać ilościowo tylko te same jakości, a więc ilościowe podejście rzekomo zuboża ludzkie poznanie. To obronne stanowisko wynika z niezrozumienia znaczenia metod ilościowych i możliwości pomiaru oraz matematycznego zapisu.

jakość a ilość w fizyce

Otóż trzeba podkreślić z naciskiem, że pomiar nie gubi żadnych różnic jakościowych. Przeciwnie — uściśla je i wzbogaca, czyni je komunikatywnymi, obiektywnymi i odtwarzalnymi. Jest faktem, że niektóre różnice, które naszym zmysłom wydają się czysto jakościowe, udało się fizykom sprowadzić do różnic czysto ilościowych. Tak np. barwa żółta jest dla naszego oka jakościowo odmienna od barwy zielonej. Fizycy odkryli jednak, że światło jest falą elektromagnetyczną i że każdej czystej barwie odpowiada określona długość fali. Jakościowe różnice

między barwami zostały wyrażone przez ilościowe różnice tej samej jakości, którą jest długość fali elektromagnetycznej. Ludzkie doznania i wrażenia odbierane przy oglądaniu barwnych przedmiotów nie zostały przez to odkrycie zubożone, lecz wręcz przeciwnie, wzbogacone przez możliwość dokładnej fizycznej analizy i reprodukcji barwnych obrazów.

Najczęściej jednak różnice jakościowe nie dają się sprowadzić do różnic często ilościowych w ramach tej samej jakości. Dźwięki np. mają też naturę falową. Wysokości czystego tonu przyporządkowana jest znowu określona długość fali, ale tym razem jest to fala akustyczna, która ma zupełnie inne własności niż fala elektromagnetyczna. Chociaż czysta barwa i czysty ton określone są przez długość fali, wiadomo jednak, że chodzi o zupełnie odmienne rodzaje fal (które mierzymy w zupełnie inny sposób), więc stwierdzenie takie bynajmniej nie lekceważy zasadniczych różnic jakościowych. Zatem ilościowe podejście nie zuboża, lecz wzbogaca, uściśla i upraszcza nasze poznanie, stwarza możliwości zupełnie nowego, syntetycznego spojrzenia na świat zjawisk, daje nowe możliwości rozlicznych praktycznych zastosowań.

Analiza badanego zjawiska przy użyciu adekwatnych wielkości fizycznych bywa nieraz zadaniem nader trudnym, polega ona bowiem na konkretnej odpowiedzi na 3 zasadnicze pytania eksperymentatora: co, jak i czym należy mierzyć? Trzeba więc wybrać takie ilościowe cechy badanego zjawiska, które są istotne i nieodzowne, a jednocześnie możliwie wygodne do jego zrozumienia i opisu. Następnie trzeba się zastanowić, jak mierzyć wybrane cechy. Wreszcie trzeba skonstruować odpowiednie przyrządy pomiarowe, wybrać jednostki, wycechować przyrządy, no i wreszcie wykonać pomiary z pożądaną dokładnością. Nie są to bynajmniej proste zadania, o czym świadczą liczne przykłady z historii fizyki.

co, jak i czym mierzyć?

Od niepamiętnych np. czasów ludzie obserwowali rozmaite ruchy mechaniczne polegające na zmianie położenia i prędkości ciał materialnych. Mieli też pewne — choć raczej niejasne — pojęcie o sile jako przyczynie sprawczej. A jednak bardzo długo nie potrafili znaleźć poprawnego związku między siłą a ruchem, ponieważ nie umieli wybrać odpowiednich wielkości do opisu ruchu. Największy uczony starożytności, Arystoteles, nie doceniał znaczenia pomiaru ani ilościowego, ścisłego opisu zjawisk. Opierając się na powierzchownych obserwacjach ruchów, w których występuje tarcie, doszedł do błędnego prawa ruchu, według którego stała siła wywołuje zawsze ruch ze stałą prędkością. Choć przez dwa tysiące lat fizyka Arystotelesa miała wielu krytyków, to jednak aż do XVII w. nikt nie potrafił podać właściwego rozwiązania tego problemu.

przykład: związek ruchu z siłą

Dopiero Galileusz zwrócił uwagę na przyspieszenie jako na wielkość bardzo istotną dla zrozumienia związku ruchu z siłą. Wreszcie Newton wprowadził pojęcie masy inercjalnej i podał poprawne sformułowanie prawa ruchu, według którego siła jest równa zmianie pędu ciała (punktu materialnego) w jednostce czasu. Przy pełnym, kinetycznym opisie ruchu punktu materialnego istotnymi wielkościami są składowe wektora położenia $\vec{r}(t)$, które są na ogół funkcjami czasu. Pomiary $\vec{r}(t)$ ciał na Ziemi są wykonywane za pomocą stosunkowo prostych przyrządów, a mianowicie miarki służącej do pomiaru położeń i zegara — do pomiaru czasu. Wyznaczenia położeń planet, komet, Słońca i gwiazd praktycznie nie da się wykonać jedynie za pomocą miarki i zegara. Wymaga ono pomiarów odpowiednich kątów i wykonania specjalnych rachunków. W celu zwiększenia ostrości widzenia i dokładności oraz w celu zobaczenia wielu niedostrzegalnych gołym okiem obiektów astronomicznych stosuje się ponadto lunety, teleskopy, refraktory, interferometry itp. Do obserwacji i pomiaru bardzo małych obiektów stosuje się lupy, mikroskopy optyczne, elektronowe itp. Liczne przyrządy fizyczne służą więc do zwiększenia dokładności pomiaru, a nawet rozszerzenia zasięgu naszych zmysłów poza obszar ich naturalnych możliwości postrzegania.

$$\vec{F} = \frac{d\vec{p}}{dt}$$

Detektory zjawisk

Istnieje wiele zjawisk (nawet makroskopowych), na które nasze zmysły wcale lub prawie wcale nie reagują. Nie odczuwamy np. w ogóle pola elektrycznego i magnetycznego, promieniowania elektromagnetycznego (poza wąskim zakresem widma widzialnego) ani promieniowań korpuskularnych (α, β, neutronowego), chociaż duże dawki promieniowania mogą nawet wywołać groźne schorzenia lub śmierć. Bez odpowiednich przyrządów fizycznych, zwanych detektorami, odkrycie i poznanie tego typu zjawisk byłoby w ogóle niemożliwe. Zostałyby one poza granicami naszego poznania, jako tajemne, nieznane moce „z innego świata", których istnienia człowiek mógłby się tylko domyślać, nie mając jednak żadnych możliwości ich zbadania, zrozumienia i wykorzystania.

Nasze zmysły są więc nie tylko bardzo niedokładne i subiektywne, a czasem wręcz zwodnicze (przywidzenia, złudzenia, halucynacje, majaki, stany otępienia itp.), ale również mają pod względem jakościowym bardzo wąski zakres. Detektory fizyczne pozwalają nam uwolnić się od subiektywności, niedokładności i zawodności naszych zmysłów, a ponadto umożliwiają nieograniczone rozszerzenie ludzkiego poznania na jakościowo nowe zjawiska, niedostępne bezpośredniej zmysłowej obserwacji. Potęga współczesnych nauk fizycznych polega na konstrukcji coraz bogatszego zestawu coraz bardziej wszechstronnych, coraz czulszych i wygodniejszych detektorów.

konwersja zjawisk

Ogólna zasada działania wszelkich detektorów fizycznych sprowadza się do zamiany (konwersji) zjawisk słabo dostrzegalnych lub wręcz bezpośrednio, zmysłowo niedostrzegalnych — na inne, łatwo obserwowalne przez nasze zmysły. Najdokładniejszym i najważniejszym narzędziem poznawczym człowieka jest z pewnością zmysł wzroku, który wsparty odpowiednimi urządzeniami w postaci podziałek, wskaźników i różnych wzmacniaczy — potrafi zauważyć nawet bardzo niewielkie zmiany wzajemnego położenia odpowiednio skonstruowanych części detektora (np. wskazówki przesuwające się względem nieruchomej skali, cyfry licznika automatycznego rosnące o jeden po każdym zarejestrowanym błysku optycznym czy innym sygnale). Bez żadnych istotnych ograniczeń możemy więc określić detektory fizyczne jako urządzenia (przyrządy) zamieniające różne zjawiska fizyczne na dające się łatwo obserwować zjawiska mechaniczno-wizualne: w końcu obserwujemy zawsze tylko wzajemne położenie pewnych części detektora. Detektorami są zarówno tak proste przyrządy, jak np. zegar i waga, jak i bardziej złożone, w rodzaju termoskopu, busoli, baroskopu, elektroskopu, galwanoskopu, spektroskopu, mikroskopu, teleskopu.

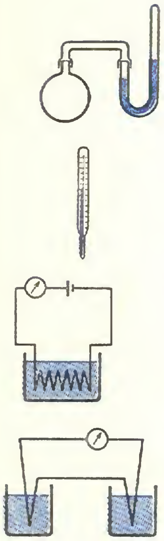
Przyrządy pomiarowe

Dobrze skonstruowany detektor reaguje na występowanie lub zmianę w badanym zjawisku jakiejś jednej wielkości fizycznej. Detektor taki można więc wywzorcować, tj. zaopatrzyć w skalę odpowiednią do wybranego układu jednostek. Z jakościowego wykrywacza danego zjawiska staje się on wówczas ilościowym przyrządem pomiarowym, a więc termoskop staje się termometrem, galwanoskop — woltomierzem lub amperomierzem, spektroskop — spektrometrem, wykrywacz promieniowania staje się licznikiem określonego rodzaju mikrocząsteczek itd. Zbudowanie i wywzorcowanie odpowiedniego zestawu detektorów stanowi definitywną i konkretną odpowiedź na trzy sakramentalne dla każdego kierunku badań pytania: co, jak i czym mierzyć?

wzorcowanie detektora

Problem wzorcowania detektora i przeistoczenia go w pełnosprawny przyrząd pomiarowy jest często bardzo trudny — nawet w pozornie prostych przypadkach. Weźmy za przykład termometr, tj. przyrząd do pomiaru temperatury. Nasz wrodzony zmysł temperatury jest wysoce niedokładny, subiektywny i ma ponadto mały zasięg, uwarunkowany chociażby faktem, że nasze ciało może żyć i funkcjonować tylko w bardzo wąskim przedziale temperatur. Termoskop wykrywający zmiany temperatury na podstawie zmiany objętości ustalonej ilości gazu skonstruował w końcu XVI w. Galileusz. Wybór najwygodniejszej substancji termometrycznej, wybór jednostki i punktu zerowego skali temperatur itp. trwał właściwie parę wieków. W połowie XVIII w. powstały względnie dokładne i wygodne termometry rtęciowe i wprowadzono do dziś stosowane skale, oparte na punktach krzepnięcia i wrzenia wody lub innych cieczy, a dopiero w 2. połowie XIX w. usunięto dowolność tego typu skal, wprowadzając pojęcie absolutnego zera i termodynamiczną skalę temperatur.

przykład: od termoskopu do termopary



Ponieważ jednak nawet najlepiej wzorcowany termometr rtęciowy nie działa ani w zbyt niskich, ani w zbyt wysokich temperaturach, powstał więc problem konstrukcji innych typów termometrów, działających poza zakresem termometru rtęciowego, problem rozszerzenia zakresu mierzonych temperatur. Skonstruowano termometry oporowe, radiacyjne, termopary itp., które pozwalają na ekstrapolację pojęcia temperatury daleko poza wąski obszar bezpośrednio uchwytny zmysłowo. Chociaż dzięki znajomości zasad termodynamiki nie mamy zasadniczych kłopotów z ekstrapolacją skali temperatur, to jednak trudności techniczne związane z konstrukcją coraz dokładniejszych i czulszych termometrów o coraz większym zakresie są bardzo duże i prace nad udoskonaleniem termometrów prowadzi się nadal.

Zagadnienie przygotowania dostosowanych do badanego zjawiska przyrządów pomiarowych wymaga nie tylko pracy techniczno-konstrukcyjnej, lecz także pracy koncepcyjnej, pojęciowej. Przyrząd pomiarowy jest ściśle związany z pojęciem wielkości fizycznej, którą mierzy. Na czym polega ten związek?

Wiele ludzi sądziło dawniej, że wprowadzenie adekwatnych do badanego zjawiska wielkości fizycznych jest przede wszystkim procesem koncepcyjnym, dokonującym się w umyśle badacza, który dopiero po wypracowaniu w drodze abstrakcyjnego rozumowania najstosowniejszych według niego pojęć powinien przystąpić do konstrukcji odpowiednich przyrządów pomiarowych. Jaką gwarancję może mieć tak postępujący badacz, że przyrząd pomiarowy mierzy rzeczywiście tę samą wielkość fizyczną, która została określona czysto abstrakcyjnie i *a priori* (np. przez podanie z góry logicznej definicji pojęcia)? W historii fizyki wiele pojęć wprowadzono *a priori*, a dopiero później przystąpiono do konstrukcji odpowiednich przyrządów pomiarowych. Nierzadko okazywało się potem, że postulowany obiekt, proces czy nawet pozornie prosta wielkość fizyczna wcale nie ma narzuconych *a priori* własności. Na przykład atomy wprowadzono do fizyki i chemii jako niezniszczalne, niepodzielne i najmniejsze cząstki materii, które mogą się łączyć w drobiny związków chemicznych. Wszechstronne badania i pomiary własności atomów wykazały, że nie są one ani niezniszczalne, ani niepodzielne, ani nie stanowią najmniejszych cegiełek składowych materii, że zatem pierwotne koncepcje można traktować jedynie jako przybliżenie przydatne w pewnym wąskim zakresie zjawisk.

określenie *a priori* a pomiary

Na początku XX w. Einstein wykazał, że newtonowskie definicje nawet tak podstawowych pojęć fizycznych jak odległość, czas, równoczesność, kolejność czasowa zdarzeń itd. nie zgadzają się z pomiarami wykorzystującymi własności światła. Implikują one zupełnie inne własności tych wielkości niż te, które postulował Newton.

Nauczeni tymi i wieloma innymi doświadczeniami fizycy przypisują obecnie znacznie większą rolę przyrządowi pomiarowemu. Przyjmuje się mianowicie, że wielkość fizyczna jest z reguły dobrze określona dopiero przez podanie pełnego przepisu jej pomiaru. Dopiero analiza sposobu i wyników pomiaru daje nam możliwość adekwatnej do rzeczywistości konstrukcji pojęciowej badanej wielkości czy złożonego zjawiska. Takie określenie wielkości fizycznej przez podanie sposobu jej pomiaru nazywa się określeniem operacyjnym. Prócz wielkości określonych operacyjnie, a więc związanych bezpośrednio z pomiarami, fizyka współczesna używa także pojęć, które mają charakter konstrukcji matematycznych, związanych z pomiarem tylko pośrednio (np. funkcja stanu w mechanice kwantowej) lub wymagających do pełnego określenia bardzo wielu, a nawet nieskończenie wielu pomiarów.

operacyjne określenie wielkości fizycznej

matematyczne określenie wielkości fizycznej

W każdym razie warto jeszcze raz podkreślić, że przyrząd pomiarowy odgrywa we współczesnej fizyce ważną rolę koncepcjotwórczą.

Szukanie związków (praw fizycznych)

Wynik pomiaru zawsze można zapisać w postaci liczby lub zbioru liczb, funkcji itp., dających się przedstawić w postaci wykresu, diagramu, tabeli itp. Trzeba przy tym pamiętać, że są to liczby obdarzone jakością, która jest określona przez przepis wykonywania pomiaru. To często nie uwidaczniane w zapisie „miano” wielkości jest bardzo ważne dla fizyka, choć może mało obchodzić matematyka.

Między różnymi wielkościami występującymi w danym zjawisku mogą występować pewne związki, zwane prawami fizycznymi. Ponieważ wielkości fizyczne wyrażają się przez liczby lub bardziej złożone twory matematyczne, związki między nimi dają się zapisać w postaci relacji matematycznych między symbolami reprezentującymi poszczególne wielkości. Najcenniejsze dla fizyka są związki mające postać matematycznych równań, choć nie do pogardzenia są również nierówności i inne relacje matematyczne. Znajdowanie praw fizyki i badanie na drodze matematycznej ich różnych konsekwencji jest głównym celem teorii fizycznych.

Przystępując do badania jakiegoś nowego, złożonego zjawiska, staramy się przede wszystkim dokładnie i wszechstronnie zmierzyć różne występujące w nim wielkości fizyczne. Lepiej jest zmierzyć raczej za dużo niż za mało różnych wielkości fizycznych. Posiadając bowiem za mało informacji, na pewno nie możemy w pełni opisać badanego zjawiska. Jeśli natomiast wprowadzimy za dużo wielkości, to dostatecznie dokładne pomiary wykażą, że nie są one niezależne i że można wszystkie wielkości wyrazić algebraicznie przez pewne wielkości podstawowe. Żadna z wielkości podstawowych nie da się — ogólnie biorąc — wyrazić jako funkcja pozostałych, ale mogą między nimi występować bardziej złożone związki, np. w postaci równań różniczkowych lub całkowych. Taką postać mają równania ruchu Newtona, równania Maxwella czy równania wielu innych teorii.

wielkości podstawowe i pochodne

Rozpatrzmy względnie prosty przykład klasycznej mechaniki punktu materialnego (cząstki). Podstawowymi wielkościami fizycznymi w tym problemie są masa cząstki m oraz wektory położenia $\vec{r}(t)$ i siły $\vec{F}(t)$. Jeśli znamy te wielkości w interesującym nas przedziale czasu, to znamy właściwie wszystko: wiemy, jak się cząstka porusza, wiemy, jaka jest jej masa i jaka siła na nią działa. Algebraicznie wielkości te są niezależne, tzn. nie ma ogólnego słusznego dla dowolnego ruchu związku funkcyjnego, który by wyrażał $\vec{r}(t)$ jako funkcję m i $\vec{F}(t)$ (lub $\vec{F}(t)$ jako funkcję m i $\vec{r}(t)$). Łatwo jednak zauważyć, że siła $\vec{F}(t)$ w jakiś sposób wpływa na ruch, choć nie określa go jednoznacznie. Na przykład stała siła grawitacyjna powoduje, że wszelkie możliwe tory cząstek są parabolami o osiach skierowanych zgodnie z kierunkiem siły.

przykład: mechanika punktu materialnego

Znając funcję wektorową $\vec{r}(t)$ w w pewnym przedziale czasu, możemy obliczyć jej kolejne pochodne:

$$\vec{v}=\frac{d\vec{r}}{dt},\quad \vec{a}=\frac{d^2\vec{r}}{dt^2},\quad \vec{\eta}=\frac{d^3\vec{r}}{dt^3}\quad \dots$$

Pierwsze dwie pochodne nazywamy odpowiednio prędkością i przyspieszeniem cząstki w chwili t. Z pierwotnych wielkości oraz powyższych pochodnych możemy utworzyć teraz dalsze bardziej złożone wielkości, jak np. pęd, moment pędu, moment siły:

$$\vec{p}=\frac{m\vec{v}}{\sqrt{1-v^2/c^2}},\quad \vec{j}=\vec{r}\times\vec{p},\quad \vec{D}=\vec{r}\times\vec{F},\quad \dots$$

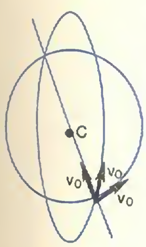
Szukany związek między ruchem, czyli funkcją $\vec{r}(t)$, a siłą znalazł Newton w postaci równości siły i pochodnej pędu:

$$\frac{d\vec{p}}{dt}=\vec{F}.$$

Czysto matematyczną konsekwencją tego prawa jest analogiczny związek między momentami:

$$\frac{d\vec{j}}{dt}=\vec{D}.$$

Prawo ruchu Newtona ma postać niezależną ani od konkretnej wartości siły, ani od konkretnego toru $\vec{r}(t)$, ani od wartości masy. Dotyczy więc dowolnych sił, dowolnych wartości masy i całej klasy torów dopuszczalnych przez to prawo. Z matematycznego punktu widzenia prawo dynamiki Newtona jest układem 3 równań różniczkowych drugiego rzędu na nieznane funkcje czasu $x(t)$, $y(t)$, $z(t)$. Nawet przy określonej wartości $\vec{F}(t)$ i m równania te dopuszczają jeszcze nieskończenie wiele rozwiązań. Rozwiązanie, tj. funkcja $\vec{r}(t)$, będzie jednoznacznie określone dopiero po sprecyzowaniu warunków początkowych, tj. po podaniu wektorów położenia $\vec{r}(t_0)$ i prędkości $\vec{v}_0=\vec{v}(t_0)$ w chwili początkowej t_0.



Zależność toru od $\vec{v}_0$ (C — centrum siły harmonicznej)

Prawa fizyki mają bardzo często różniczkową lub całkową postać, która jak widzieliśmy na powyższym, stosunkowo prostym przykładzie — nie dotyczy jednego szczególnego ruchu, ale podaje związki prawdziwe w odniesieniu do całej klasy fizycznie dopuszczalnych ruchów. Innymi słowy — różniczkowe prawa fizyki dopuszczają nieskończenie wiele rozwiązań. Wybór konkretnego rozwiązania jest zależny od interesującej nas konkretnej sytuacji fizycznej. Prawa fizyki opisują więc tylko ogólne związki, które odnoszą się do nieskończonej liczby fizycznie dopuszczalnych sytuacji, mają zatem bardzo syntetyczny, ogólny i powszechny charakter. Językiem adekwatnym do zapisu tych ogólnych związków oraz do ich analizy i wyciągania różnorodnych wniosków jest ścisły i ogólny język matematyki.

Teorie fizyczne

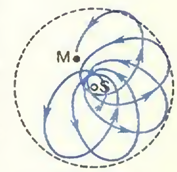
Fizyka interesują takie obiekty jak cząstki elementarne, jądra, atomy, jony, molekuły oraz makroskopowe układy takich cząstek jak np. gazy, ciecze, ciała stałe, planety i gwiazdy tudzież pola. Teorią fizyczną nazywamy zbiór zapisanych w postaci matematycznej praw fizycznych odnoszących się do obiektów określonego typu, uzupełniony warunkami stosowalności, warunkami wyboru fizycznie dopuszczalnych rozwiązań oraz omówieniem precyzującym znaczenie fizyczne występujących symboli matematycznych. Teoria powinna podawać związki przyczynowe i opisywać występujące w rozważanym obiekcie (układzie) zjawiska i prawa możliwie kompletnie, ogólnie i z wystarczającą dokładnością. Kryterium przydatności teorii fizycznej jest zawsze ilościowa zgodność jej przewidywań z faktami doświadczalnymi. Każda

co to jest teoria fizyczna

teoria fizyczna ma ograniczony zasięg stosowalności i ograniczoną dokładność. Jednym z głównych celów badań fizycznych jest więc konstrukcja coraz doskonalszych teorii, tzn. teorii, które są coraz głębsze i ogólniejsze, mają coraz większy zasięg stosowalności i coraz lepszą zgodność z doświadczeniem.

przykład: teoria ruchu planet

Rozpatrzymy dla przykładu sukcesywne teorie ruchów planet. Teoria geocentryczna Ptolemeusza była czysto opisowym, kinematycznym, niedynamicznym i nieprzyczynowym schematem rachunkowym. Znacznie uprościł i ufizycznił opis ruchów planet Kopernik, wprowadzając konsekwentnie układ heliocentryczny. Stosując ten układ, Kepler odkrył 3 ważne kinematyczne prawa ruchu. Pierwszą ogólną, przyczynową i dynamiczną teorię podał Newton. Składają się na nią uniwersalne prawa dynamiki Newtona oraz prawo powszechnego ciążenia. Następnym udoskonaleniem było uwzględnienie na początku XX w. relatywistycznej zależności masy od prędkości. Najdokładniejszą w tej chwili teorią ruchów planet jest teoria grawitacji Einsteina (tzw. ogólna teoria względności), która przewiduje drobne odstępstwa od wyników teorii Newtona (np. ruch perihelionowy Merkurego). Teoria Einsteina jest ponadto najgłębszą — w sensie przyczynowego tłumaczenia zjawisk — teorią grawitacji i ruchów planetarnych. Paradoksalnie wręcz brzmi stwierdzenie, że dokładność teorii Einsteina przekracza ciągle doświadczalne możliwości. Wiele przewidywanych przez tę teorię efektów, np. pewne poprawki do ruchów planet, promieniowanie grawitacyjne itd., są tak małe, że nie dadzą się wykryć przy zastosowaniu znanych dotychczas detektorów. Inne przewidywane przez tę teorię efekty, jak np. czarne dziury, mogą występować jedynie w dużych, kosmicznych obiektach i też nie zostały jeszcze definitywnie wykryte.



Elipsa Merkurego wykonuje pełny obrót w 3 mln lat

od doświadczenia do teorii

Historia różnych dziedzin fizyki toczyła się różnie. Przeważnie rozpoczynała się od doświadczalnego odkrycia jakiegoś nowego zjawiska (obiektu, procesu, własności). W takiej sytuacji w pierwszym okresie badawczym dominował eksperyment, a więc konstrukcja odpowiednich detektorów, pomiary różnych charakterystycznych wielkości, szukanie fenomenologicznych reguł, korelacji, struktur i związków. Oczywiście i na tym etapie — jak już wspominaliśmy — dużą rolę odgrywa tworzenie nowych pojęć i konstrukcji myślowych, bez których eksperyment byłby ślepy i przypadkowy. Koncepcje teoretyczne wprowadzane wówczas są więc nieodzowne dla właściwego ukierunkowania doświadczeń. Nie stanowią one jeszcze teorii, ale są przygotowaniem teorii, której celem jest wykrycie istotnych dla danego typu obiektów czy zjawisk, strukturalnych i dynamicznych związków przyczynowych oraz zapisanie ich w odpowiedniej matematycznej formie.

od teorii do doświadczenia

Z chwilą skonstruowania takiej teorii zaczyna się badanie i sprawdzanie jej konsekwencji. Od metody indukcyjnej, stosowanej na poprzednim konstrukcyjnym etapie, przechodzi się do metody dedukcyjnej, polegającej na wyciąganiu logicznych wniosków z założeń i równań teorii. Dobra teoria nie powinna zawierać logicznych i matematycznych sprzeczności, powinna tłumaczyć przyczynowo znane już fakty i przewidywać nowe. Za szczególnie cenne osiągnięcie teorii uważa się potwierdzenie przez doświadczenie wszystkich przewidywań nieznanych uprzednio faktów. Oczywiście zgodność z faktami doświadczalnymi musi być ilościowa, a nie tylko jakościowa. Ważną zaletą każdej teorii fizycznej jest więc możliwie dokładna zgodność z doświadczeniem.

Zdarzyło się niejednokrotnie w historii fizyki, że pewne nieoczekiwane przewidywania teorii dawały asumpt do wykrycia zupełnie nowych zjawisk i do powstania nowych działów fizyki. Czasem więc i doświadczenie wyprzedza teorię, czasem — na odwrót — teoria wyprzedza doświadczenie. Maxwell np. przewidział na podstawie skonstruowanej przez siebie teorii elektryczności i magnetyzmu istnienie fal

elektromagnetycznych i zbadał teoretycznie ich własności. Fale te wykrył doświadczalnie H. Hertz wiele lat później. Szczególna teoria względności Einsteina przewidywała szereg zaskakujących efektów, jak np. wzrost masy lub wydłużanie się czasu życia nietrwałych cząstek będących w ruchu względem obserwatora, co zostało później potwierdzone doświadczalnie.

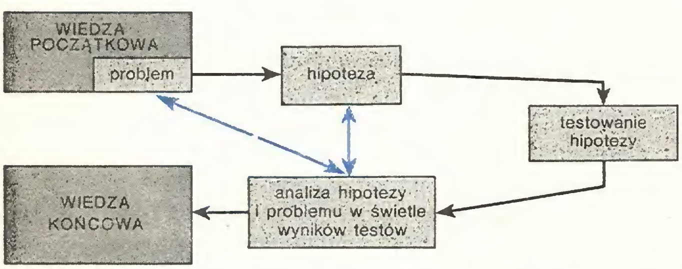
Można podać bardzo wiele takich przykładów swoistego wyprzedzania doświadczenia przez teorię. Z drugiej jednak strony — wszystkie teorie fizyczne powstały w wyniku analizy odpowiednich danych doświadczalnych. Nie ma w fizyce sprawdzonych teorii, które byłyby „wymyślone" na drodze czysto racjonalnych rozważań, nie opierających się na żadnych faktach doświadczalnych. Jest to możliwe w matematyce, która jako nauka czysto racjonalna nie musi się opierać na żadnych realnych faktach. Ale fizyka jest nauką empiryczną, nie może być zatem czysto racjonalna, lecz musi być racjonalno-empiryczna.

Sprzężenie teorii z doświadczeniem

Czasami po pewnym okresie badań eksperymentalnych i przygotowawczej działalności koncepcyjnej dobra teoria powstawała stosunkowo szybko, tj. bez wielu nieudanych prób. Niejednokrotnie powstanie dobrej teorii poprzedzał szereg próbnych hipotez, które po konfrontacji z doświadczeniem trzeba było odrzucać lub modyfikować, usuwając stopniowo zauważone braki i luki.

cykl badawczy

Racjonalno-empiryczna metoda naukowa, stosowana w fizyce i innych naukach empirycznych, opiera się na bardzo charakterystycznych cyklach badawczych. Typowy cykl badawczy składa się z następujących etapów (rys. 1): stan początkowy (wyjściowy) wiedzy, postawienie problemu, wysunięcie hipotezy rozwiązującej ten problem (przynajmniej częściowo),



Rys. 1. Struktura typowego cyklu badawczego. Strzałki niebieskie przedstawiają sprzężenia zwrotne

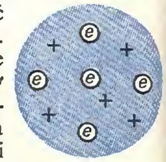
testowanie hipotezy (logiczne, matematyczne i empiryczne), analiza hipotezy i problemu w świetle wyników testów (odrzucenie, korekta lub przyjęcie hipotezy), nowy stan wiedzy.

Jeśli testy nie potwierdzają w pełni hipotezy, należy ją poprawić lub zastąpić nową hipotezą, albo też skorygować sam problem, rozpoczynając w ten sposób nowy cykl badawczy. Jeśli po pewnej liczbie takich cykli dojdziemy do zgodnej z wszystkimi faktami hipotezy rozwiązującej nasz problem, możemy ją uznać za dobrą teorię i włączyć do zasobu sprawdzonej wiedzy naukowej. Wynikiem takiego postępowania jest więc osiągnięcie nowego, wyższego poziomu wiedzy.

Punktem wyjściowym każdego cyklu badawczego jest zastany przez badacza stan wiedzy, z którego wyłania się problem (pytanie). Czasem rozwiązanie problemu wynika z już znanych faktów oraz już istniejących teorii i nie wymaga nowych hipotez ani stosowania podanego cyklu badawczego. Innymi słowy, odpowiedź na tego typu pytanie tkwi *implicite* w zastanej wiedzy, ale wymaga zebrania i przetworzenia pewnych informacji. Jeśli np. zapytamy astronoma, jakie będzie położenie planety Wenus jutro o godzinie 23, to z pewnością nie odpowie nam z miejsca, lecz najpierw zajrzy do odpowiednich tablic astronomicznych, następnie wykona konieczne obliczenia i dopiero potem będzie mógł podać współrzędne planety w żądanej chwili. Tego rodzaju problemy występują z reguły w nauczaniu oraz w zastosowaniach wiedzy naukowej, gdzie odgrywają bardzo istotną rolę. Do poznania nie wnoszą jednak nic istotnie nowego, ponieważ właściwie nie doskonalą i nie rozszerzają zasobu naszej wiedzy, wydobywają tylko z ogromu tej wiedzy aktualnie interesującą nas informację.

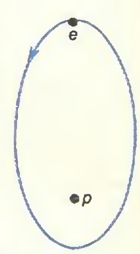
Z poznawczego punktu widzenia znacznie ważniejszy jest inny typ problemów, które są implikowane przez aktualną wiedzę naukową, ale których rozwiązanie nie wynika z tej wiedzy, lecz wymaga jej rozszerzenia. Na przykład — po odkryciu możliwości wzbudzania atomów oraz jonizacji, tj. wyrywania elektronów z atomów lub molekuł, powstały następujące problemy: Jakie siły trzymają elektrony w atomie? Jakie ruchy może wykonywać elektron w atomie? Jakie jest pochodzenie dodatniego ładunku, który neutralizuje ujemny ładunek elektronów? Jak rozłożone są ładunki dodatnie i ujemne w atomie?

przykład: hipotezy budowy atomu



Pierwotny model atomu

Pierwsze hipotezy opierały się na założeniu, że elektrony trzymane są w atomie siłami elastycznymi (niewiadomego pochodzenia) i mogą wykonywać drgania o stałej częstości dokoła położenia równowagi. Dodatni ładunek i masa atomu miały być rozmazane po całym jego wnętrzu. Taki prosty model teoretyczny tłumaczył sporo zjawisk, ale nie potrafił dać odpowiedzi na dwa ostatnie z podanych wyżej pytań. Poza tym nie potrafił wyjaśnić wielkiej liczby częstości drgań występujących nawet w najprostszym jednoelektronowym atomie wodoru. Dopiero z przeprowadzonej przez Rutherforda analizy rozproszeń cząstek α przez atomy wynikło, że cały dodatni ładunek atomu i prawie cała jego masa skupione są w jądrze o rozmiarach ok. 10^5 razy mniejszych od rozmiarów atomu i że elektrony krążą dokoła jądra trzymane na uwięzi przez siły kulombowskiego przyciągania pochodzące od dodatniego ładunku jądra. Dwa lata później (1913 r.) Bohr zaproponował swój model atomu wodoru, w którym uwzględnił wyniki Rutherforda, dodał przy tym do mechaniki Newtona trzy nowe kwantowe postulaty. Również model Bohra nie był w pełni zadowalający, gdyż nie dawał możliwości obliczenia natężeń linii widmowych, nie nadawał się do obliczeń poziomów cięższych atomów itd. Wreszcie w 1925 r. powstała mechanika kwantowa (falowa) atomu, którą po pewnych udoskonaleniach, związanych z wprowadzeniem spinu i statystyki, uwzględnieniem efektów relatywistycznych itp., można było uznać za dobrą teorię atomów, molekuł i wszelkich ciał z nich złożonych.



Model atomu wodoru Bohra

Widać na tym przykładzie, że w celu rozwiązania drugiego rodzaju problemów należy wysunąć odpowiednią hipotezę teoretyczną, której zasadność i poprawność trzeba wszechstronnie sprawdzić. Zgodne z hipotezą wyniki testów uzasadniają ją, tzn. umacniają przekonanie o jej słuszności, ale z logicznego punktu widzenia jedynie jej nie wykluczają i nie stanowią nigdy niezbitego dowodu jej prawdziwości. Dotyczy to nawet bardzo dobrze ugruntowanych, wszechstronnie sprawdzonych i empirycznie potwierdzonych teorii. Ani więc próbne, robocze hipotezy, ani też dobrze sprawdzone teorie fizyczne nie są niepodważalne i absolutnie prawdziwe. Nawet najlepsze teorie mają ograniczony zasięg stosowalności, tzn. tłumaczą i opisują w zgodzie z doświadczeniem tylko pewien ograniczony zakres realnych faktów; ponadto opis ten ma zawsze ograniczoną dokładność. Nawet najlepiej sprawdzone teorie nie mają więc charakteru ostatecznej, dokładnej i absolutnej prawdy, przedstawiają jedynie prawdy cząstkowe, oparte na pewnych przybliżeniach. I tak np. wspomniana mechanika kwantowa atomu też nie jest

uniwersalną i bezwzględnie dokładną teorią, choć jej zasięg stosowalności jest naprawdę ogromny. Zasięg ten nie obejmuje jednak ani cząstek elementarnych, ani wnętrza jądra, ani licznych procesów promieniowania, kreacji i anihilacji cząstek itd.

Scharakteryzowana wyżej metoda badawcza, oparta na powtarzaniu stopniowo doskonalonych cykli badawczych, ma tę zaletę, że świetnie pasuje do przybliżonego i stopniowego charakteru naszego poznania. Pozwala ona bowiem na systematyczne doskonalenie naszej wiedzy, która polega nie tylko na odkrywaniu nowych faktów, lecz także na tworzeniu nowych coraz lepszych teorii, opisujących przyczynowo coraz większy zakres faktów z coraz większą dokładnością. Metoda naukowa fizyki nie jest więc metodą zdobywania prawd absolutnych, lecz metodą doskonalenia prawd cząstkowych, metodą kolejnych przybliżeń.

Rola teorii w poznaniu przyrody

Doskonalenie wiedzy naukowej jest możliwe dzięki zawartym w cyklu badawczym sprzężeniom zwrotnym między hipotezą i realnymi faktami, czyli między teorią i doświadczeniem. Struktura tego cyklu uwzględnia również ogromną rolę teorii w poznaniu świata. Może się wydawać, że teoria jest niepotrzebnym, bo jak stwierdziliśmy wyżej — nie całkiem pewnym elementem poznania. Realne fakty są niewątpliwie pewniejsze i ich znajomość stanowi oczywiście niezwykle cenną część wiedzy naukowej. Wiedza uznająca tylko realne fakty, a odrzucająca wszelkie teorie wykraczające poza porządkowanie faktów, byłaby jednak wiedzą często opisową, nie wchodzącą w przyczyny zjawisk, pozbawioną możliwości tłumaczenia i przewidywania. Byłaby to zatem wiedza zdecydowanie uboższa — zarówno w sensie poznawczym, jak i praktycznym — od wiedzy naukowej, która również opiera się na realnych faktach, ale w celu ich zrozumienia, wyjaśnienia, przewidywania i praktycznego wykorzystania posługuje się teoriami.

Okazuje się zresztą, że nawet samo uporządkowanie i klasyfikacja faktów oparta na odpowiednich teoriach jest z reguły znacznie lepsza, prostsza, lepiej uzasadniona i głębsza. Koncepcje, prawa i zasady teoretyczne tworzą poza tym konieczny dla głębszego poznania język pojęć i wytyczają sposób myślenia oraz sposób patrzenia na badane zjawiska. Koncepcje teoretyczne sugerują często, co należy badać, czym i jak. Innymi słowy, teorie bardzo często zarówno sugerują kierunki i obiekty dalszych badań empirycznych, jak też wskazują możliwe narzędzia i metody badawcze. Warto sobie np. uświadomić ogromną stymulującą rolę hipotezy atomowej w XIX w., kiedy to przez długi czas opierała się ona jedynie na prostej interpretacji kilku makroskopowych praw zaobserwowanych głównie w procesach chemicznych. Hipoteza atomowa dała jednak liczne konkretne wskazówki dotyczące badań doświadczalnych, sugerując, jak szukać innych, coraz wszechstronniejszych i coraz bardziej bezpośrednich eksperymentalnych metod badania atomów, molekuł i jonów.

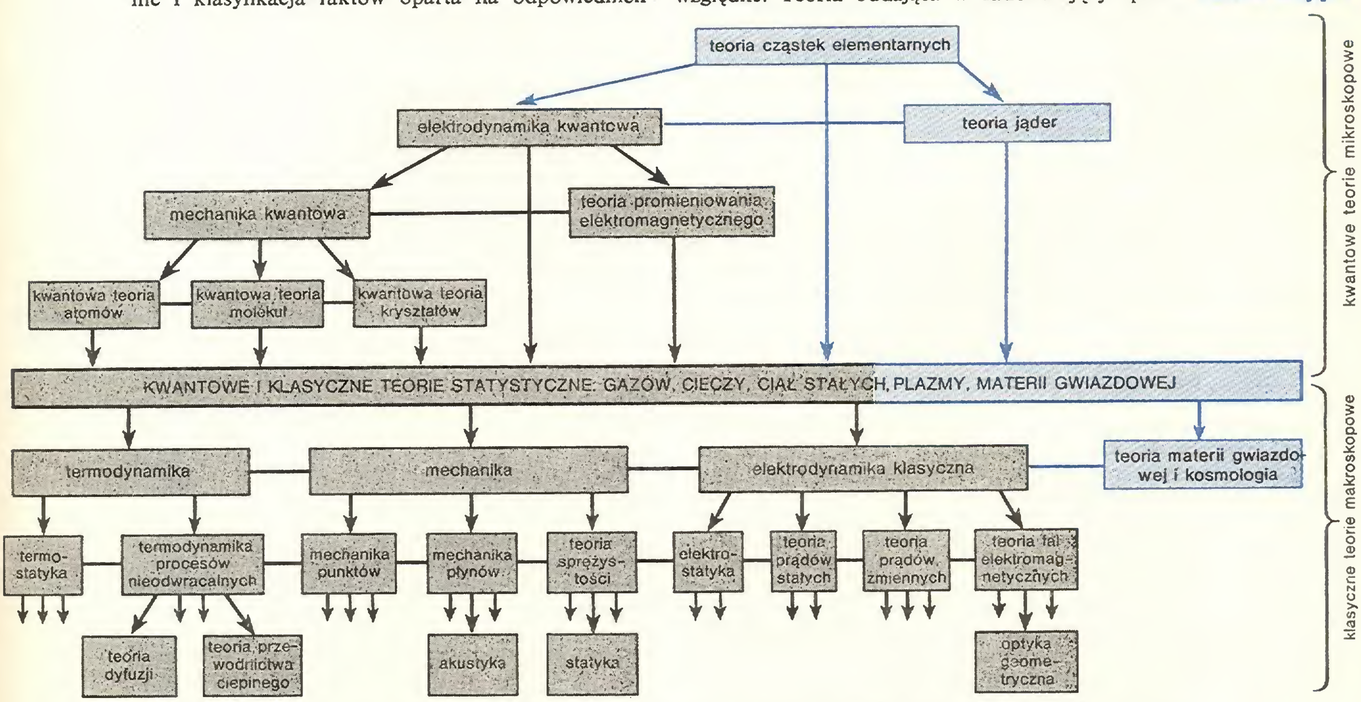
stymulująca rola hipotezy atomowej

Teorie naukowe są więc podstawową formą, sposobem i narzędziem poznania. Niemożliwe byłoby głębsze, przyczynowe poznanie i szersze praktyczne zastosowanie zjawisk przyrody, gdybyśmy się ograniczyli wyłącznie do spisów i opisów obserwowanych faktów. Naukowe poznanie i wykorzystanie praw przyrody odbywa się za pomocą ustawicznie doskonalonych i sprawdzanych przyczynowych teorii, a nie za pomocą najbardziej nawet skrupulatnych katalogów faktów. Niech ktoś spróbuje skonstruować działający samochód bez żadnej teorii, nie znając w ogóle praw fizyki, mając tylko do dyspozycji dokładny katalog wszystkich części samochodu!

Zbudowanie możliwie dokładnej, przyczynowej, potwierdzonej przez realne fakty teorii naukowej jest więc w pewnym sensie głównym celem i uwieńczeniem określonego etapu badań. Z reguły etap taki zawiera nie tylko jeden cykl badawczy, lecz całą spiralę, a nawet wiele spiral składających się z szeregu cykli badawczych, zbliżających się poprzez różne nieudane i częściowo udane, bezpłodne i płodne hipotezy robocze do zadowalającej teorii badanego zakresu zjawisk.

Oczywiście określenie „zadowalająca teoria" jest względne. Teoria oddająca w zadowalający sposób

teoria zadowalająca



Rys. 2. Hierarchiczna struktura współczesnych teorii fizycznych. Zaznaczono tylko najważniejsze teorie fizyczne. Na szarym tle podano teorie w zasadzie zadowalające, dobrze sprawdzone, na niebieskim — teorie w budowie, czyli jeszcze niezadowalające. Relacje implikacji przedstawione są strzałkami; strzałki czarne oznaczają implikacje dobrze sprawdzone, niebieskie — implikacje spodziewane. Linie poziome oznaczają związki (często obustronne) między teoriami tego samego poziomu

pewien zakres zjawisk może się stać niezadowalająca, gdy zwiększymy wymagania dokładności lub powiększymy zakres faktów, np. przez odkrycie nowych zjawisk nie objętych przez dotychczasową teorię. Trzeba wtedy szukać nowej, ogólniejszej i dokładniejszej teorii. W fizyce zdarzało się już wielokrotnie, że ta nowa, doskonalsza teoria nie była sprzeczna z poprzednimi (starymi) teoriami i wcale nie obalała ich, lecz tworzyła nowy, ogólniejszy, szerszy i głębszy punkt widzenia. Stare teorie okazywały się przypadkami szczególnymi lub granicznymi, tj. pewnymi przybliżeniami nowej. Teorie fizyczne są dziś ściśle powiązane w bardzo ciekawą hierarchiczną strukturę, która przypomina trochę drzewo genealogiczne. Wiele gałązek i konarów tego drzewa już znamy, ale do centralnego pnia, z którego wszystko wyrasta, jeszcześmy się nie dostali (rys. 2).

Rys. 2 przedstawia schemat powiązań między dobrze sprawdzonymi teoriami fizycznymi i teoriami obecnie tworzonymi.

Przyczynowość i determinizm

Fizyka interesuje się szczególnie prawami, które mają charakter związków przyczynowych, tzn. określają pewną sytuację zwaną skutkiem, jeśli tylko spełnione są warunki zwane przyczyną. Rozumienie charakteru związku przyczynowego nie jest w nauce ani w filozofii zupełnie jasne i jednoznaczne. Niektórzy uważają, że związek przyczynowy dotyczy tylko zmian w czasie, a więc zdarzeń i procesów. W tym ujęciu przyczyna p jest zdarzeniem (lub procesem) poprzedzającym w czasie inne zdarzenie s, zw. skutkiem. Przy tym zawsze, jeśli w chwili t zajdzie p, to w późniejszej chwili $t'>t$ zajdzie również s. Liczne równania ruchu (np. równania Newtona, Maxwella, Schrödingera, równania przewodnictwa cieplnego, dyfuzji) opisują takie czasowe związki przyczynowe. Podanie stanu początkowego układu w chwili t i wartości wszystkich oddziaływań (sił) w przedziale czasu od t do $t'>t$ określa jednoznacznie stan układu w dowolnej chwili późniejszej t'. Skutek można uznać w tym przypadku za zdarzenie polegające na tym, że układ znajduje się w określonym stanie w chwili t'. Natomiast za przyczynę trzeba uznać cały zespół warunków, a mianowicie zdarzenie polegające na wystąpieniu określonego stanu początkowego w chwili t oraz zadziałanie określonych sił i praw ruchu w przedziale od t do t'. Stan końcowy układu jest zdeterminowany przez stan początkowy i równania ruchu w przedziale od t do t'. Tak rozumiana przyczynowość wiąże się więc ściśle z determinizmem.

związki przyczynowe

determinizm

Zamiast pojęciami przyczyny i skutku determinizm operuje pojęciami stanu początkowego i końcowego oraz pojęciem równań ruchu. Przy tych samych oddziaływaniach stan początkowy wyznacza jednoznacznie stan końcowy. Natomiast przy różnych oddziaływaniach różne stany początkowe mogą prowadzić do tego samego stanu końcowego i na odwrót — z tego samego stanu początkowego możemy otrzymać różne stany końcowe.

zdarzenia elementarne

Głębsze podejście do zagadnienia przyczynowości oparte jest na dokładniejszej analizie oddziaływań między parami zdarzeń elementarnych $\varphi(A)$ i $\varphi(A')$ zachodzących w punktach $A = (t, x, y, z)$ i $A' = (t', x', y', z')$ czasoprzestrzeni. (Zdarzeniem elementarnym może być np. określona wartość jakiegoś pola $\varphi(A)$ w punkcie A. Mając teorię opisującą rozchodzenie się pola φ, możemy powiedzieć, czy i jak wartość pola $\varphi(A)$ w punkcie A wpływa na wartość pola $\varphi(A')$ w punkcie A'.) Stawiamy pytanie jaki jest zbiór zdarzeń elementarnych $\varphi(A)$, które stanowią łącznie przyczynę zdarzenia $\varphi(A')$.

Dokładna odpowiedź zależy od rodzaju interesujących nas zdarzeń elementarnych $\varphi(A)$. Najistotniejsze ogólne ograniczenia daje teoria względności, która podaje następujące warunki konieczne istnienia przyczynowego związku między zdarzeniami zachodzącymi w A i zdarzeniem zachodzącym w A':

$$t' \geqslant t, \quad c^2(t'-t)^2 \geqslant (r'-r)^2.$$

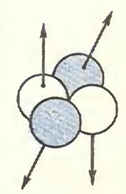
Jeśli któryś z tych warunków nie jest spełniony, zdarzenia zachodzące w A nie mogą mieć wpływu na zdarzenia zachodzące w A'. Mówimy wtedy, że punkty A i A' są przyczynowo nie powiązane. Znaczenie pierwszej nierówności jest oczywiste. Natomiast druga nierówność oznacza, że prędkość rozchodzenia się wszelkich realnych oddziaływań fizycznych nie może być większa od prędkości światła c w próżni. W wypadku oddziaływań elektromagnetycznych prędkość ta wynosi dokładnie c, dla innych oddziaływań — może być mniejsza lub równa c.

zdarzenia przyczynowo nie powiązane

W nieco luźniejszym sformułowaniu związek przyczynowy łączy się zawsze z jakimś pytaniem typu: dlaczego? W zależności od rodzaju dopuszczalnych pytań otrzymamy różne „rodzaje" przyczyn. W odniesieniu do ruchu punktu materialnego możemy np. zadać pytanie: Dlaczego pęd cząstki zmienia się w taki, a nie inny sposób? Gdy znamy równania dynamiki Newtona, odpowiedź na takie pytanie jest łatwa: Pęd zmienia się w określony sposób, ponieważ na cząstkę działa odpowiednia siła. Gdyby na cząstkę nie działały żadne siły, pęd jej nie zmieniałby się w ogóle. Bezpośrednią przyczyną sprawczą zmiany pędu jest więc siła. Jeśli jednak nieco inaczej zapytamy: Dlaczego przyspieszenie wynosi w chwili t tyle a tyle? — to odpowiedź będzie inna. A mianowicie o wartości przyspieszenia w danej chwili decyduje nie tylko siła, lecz także masa cząstki i chwilowa wartość jej prędkości. Za przyczynę określonej zmiany prędkości (czyli określonej wartości przyspieszenia) wypada więc uznać nie tylko siłę, lecz także masę i prędkość w interesującej nas chwili.

dlaczego...?

Widać z tego przykładu, że określenie przyczyny i skutku zależy od sposobu stawiania pytań. Pytania typu „dlaczego" mogą w ogóle nie dotyczyć zmian zachodzących w czasie ani żadnych zmian. Możemy np. zadać pytanie: Dlaczego jądro helu ma podwójny ładunek dodatni i spin równy zeru? Odpowiedź będzie brzmiała: Ponieważ składa się ono z dwóch dodatnio naładowanych protonów i dwóch neutronów, przy tym wszystkie one są w stanie podstawowym z momentem orbitalnym równym zeru, a spiny obu protonów i obu neutronów muszą być wtedy — zgodnie z zakazem Pauliego — skierowane przeciwnie. Podane przyczyny mają teraz charakter nie dynamiczny, lecz czysto strukturalny. Pytanie nie dotyczy zmian w czasie, więc nie było potrzeby odwoływania się do sił jądrowych i równań ruchu. Wystarczyło podanie prostych informacji o strukturze jądra i prostego prawa w postaci zakazu Pauliego.



Model jądra helu

Fizycy są skłonni traktować determinizm i przyczynowość raczej szeroko, stosując te pojęcia do sytuacji, w których własności obiektów wynikają teoretycznie z ich struktury lub odpowiednich dynamicznych praw. Związek przyczynowy w tym szerszym ujęciu przypomina więc logiczną relację wynikania $p \rightarrow s$, a więc: jeśli zachodzi p, to musi także zachodzić s. Szczególnie poglądowe i ciekawe jest wynikanie z odpowiednich teorii własności obiektów złożonych z własności ich części składowych.

przyczynowość jako relacja wynikania

Bardzo wygodny i poglądowy jest podział obiektów fizycznych według stopnia złożoności, co odpowiada z grubsza ich wzrastającym rozmiarom. Mamy więc poziom: cząstek elementarnych, jąder, atomów, molekuł i jonów, zwykłych ciał makroskopowych, Ziemi i planet, gwiazd, galaktyk i Wszechświata. Do chemii należy poziom atomów i drobin; fizyka klasyczna zajmuje się głównie ciałami makroskopowymi o niezbyt wielkich rozmiarach; geologia i geofizyka zajmują się Ziemią, jej składem i budową; astronomia zajmuje się ostatnimi czterema poziomami; biologia sięga od poziomu cząsteczek do poziomu ciał makroskopowych.

Zasada, że prawa i własności obiektów należących

do wyższego poziomu wynikają z odpowiednich praw i własności ich części składowych należących do niższego poziomu, jest w fizyce w wielu wypadkach bardzo dobrze i dokładnie potwierdzona; niekiedy dokładność pozostawia jeszcze wiele do życzenia. Na przykład wynikanie własności jąder z własności nukleonów i innych cząstek elementarnych nie budzi wątpliwości, ale wobec braku zadowalającej ogólnej teorii cząstek elementarnych przewidywania są jeszcze mało dokładne. Natomiast bardzo dokładnie potrafimy wytłumaczyć i obliczyć wszystkie fizyczne i chemiczne własności atomów, ponieważ mechanika kwantowa jest bardzo dobrą teorią atomu. Okazuje się, że skomplikowane własności atomów są ściśle określone przez stosunkowo proste, czysto elektromagnetyczne i mechaniczne własności jąder i otaczających je elektronów (jak masy i ładunki elektryczne oraz momenty magnetyczne jądra i elektronu). Oczywiście wynikanie to możemy wykazać jedynie za pomocą teorii fizycznych — w tym wypadku za pomocą mechaniki kwantowej.

Z kolei struktura geometryczna i inne fizyczne i chemiczne własności molekuł wynikają z budowy zewnętrznych powłok elektronowych atomów wchodzących w skład tych molekuł. Dalej — fizyka statystyczna, teoria ciała stałego i inne teorie ukazują, jak własności ciał makroskopowych (gazów, cieczy, ciał stałych) wynikają z ich struktury atomowej, molekularnej lub jonowej. Od czasów Newtona wiemy, że ruchy planet, komet i gwiazd wynikają z tych samych praw mechaniki, które obowiązują makroskopowe ciała na Ziemi. Wiele własności Słońca i innych gwiazd jest wynikiem występujących w nich reakcji między jądrami i cząstkami elementarnymi. Wyjaśniono m.in., że źródłem ogromnej energii promieniowania wysyłanego przez Słońce i inne gwiazdy są reakcje jądrowe.

Liczne i różnorodne relacje wynikania, które już znamy, dają bardzo głębokie i ciekawe powiązania między różnymi gałęziami właściwej fizyki a innymi naukami przyrodniczymi. Związki te sprawiają, że mimo powstawania nowych specjalności naukowych, mimo odkrywania nowych zjawisk, z teoretycznego punktu widzenia nauki przyrodnicze tworzą coraz bardziej jednolitą całość. Ten niezwykle ważny i frapujący proces jednoczenia wszystkich nauk przyrodniczych na bazie fizyki nie został zakończony. Spodziewamy się odkrycia jeszcze wielu zupełnie nowych, fundamentalnych związków po powstaniu zadowalającej teorii cząstek elementarnych. Również organizmy żywe nadal kryją wiele tajemnic. Nie ma jednak powodów do niecierpliwości. Trzeba bowiem pamiętać, że fizyka cząstek elementarnych powstała dopiero ok. 30 lat temu i że ciągle odkrywamy w tej dziedzinie nowe zjawiska i nowe prawa. Niemal równie młoda jest biologia molekularna, badająca fizykochemiczne podstawy życia.

jedność nauk przyrodniczych

Matematyka jako język fizyki

Matematyka jest nauką czysto racjonalną, która nie zajmuje się odkrywaniem i badaniem empirycznych faktów, lecz tworzeniem i doskonaleniem specjalnych języków formalnych. Języki te są potrzebne do ścisłego, jasnego i komunikatywnego przedstawienia abstrakcyjnych pojęć i relacji, których opisanie i zbadanie za pomocą języka potocznego byłoby zbyt kłopotliwe, niedokładne, niejasne, czy wręcz niemożliwe.

Dla uniknięcia ewentualnych nieporozumień należy podkreślić, że twórcza praca matematyków nie ma nic wspólnego z propozycjami sztucznych języków w rodzaju esperanto. W tym ostatnim wypadku chodzi tylko o wierne tłumaczenie zdań występujących w językach naturalnych (narodowych) na nowy język, który dzięki swej prostocie miałby szanse stania się językiem światowym. W twórczych badaniach matematyków nie chodzi o tłumaczenie znanych już zdań i relacji na nowy ystem słów czy innych znaków, lecz o poszukiwanie i badanie nowych, nieznanych uprzednio relacji, do których opisu trzeba tworzyć specjalne języki symboli (znaków) i reguł postępowania.

Teorie matematyczne nie muszą się odnosić do realnej rzeczywistości ani wynikać z badania realnych faktów; nie podlegają więc empirycznej weryfikacji, lecz jedynie sprawdzeniu logicznej poprawności. W tej sytuacji jest oczywiste, że nie wszystkie możliwe, logicznie dopuszczalne teorie matematyczne i relacje nadają się do opisu rzeczywistości. W fizyce np. zwykle spośród wielu matematycznie dopuszczalnych rozwiązań równań opisujących dane zjawisko trzeba wybrać te, które spełniają jeszcze pewne dodatkowe warunki czysto fizyczne. Rozwiązania równań nie spełniające tych warunków odrzuca się jako nieprzydatne dla fizyka. Tak więc dopiero nauki empiryczne oraz praktyka (technika) stwierdzają, które z utworzonych przez matematykę języków, relacji, twierdzeń, zdań i pojęć znajdują zastosowanie do opisu rzeczywistości.

Ze względu na abstrahowanie od realnych faktów symboliczny język matematyki jest prawdziwie uniwersalny, tzn. nadaje się do opisu bardzo wielu formalnie identycznych relacji między jakościowo zupełnie odmiennymi zjawiskami. Identyczne pod względem formalnym (tj. czysto matematycznym) równania falowe mogą opisywać tak różne zjawiska jak rozchodzenie się głosu, światła lub fal radiowych. Odmienna treść fizyczna wynika dopiero z wyboru rozwiązań oraz interpretacji wielkości spełniających te równania i odpowiednie dodatkowe warunki fizyczne.

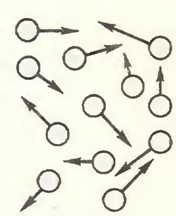
$$\Delta f = \frac{1}{c^2}\frac{\partial^2 f}{\partial t^2}$$

Równanie falowe

Jako uniwersalny, a przy tym bardzo bogaty i ścisły język, matematyka ma ogromne znaczenie dla całej nauki. Proces matematyzacji nauki zaczął się historycznie od astronomii i fizyki, gdzie ilościowy opis zjawisk przyrody i związane z tym szerokie stosowanie matematyki są najbardziej naturalne. Sformułowanie większości praw i wszystkich teorii fizycznych byłoby wręcz niemożliwe bez wszechstronnej pomocy matematyki. Potrzeby fizyki wielokrotnie inspirowały powstanie nowych działów matematyki. Tak było w czasach nowożytnych z geometrią analityczną, teorią równań różniczkowych i całkowych, rachunkiem wariacyjnym czy teorią dystrybucji, które powstały niejako na zamówienie fizyki i przy współudziale fizyków.

Modele mechaniczne i matematyczne

Hipotezy robocze, a także dobrze sprawdzone teorie fizyczne opierały się dawniej na poglądowych modelach geometryczno-mechanicznych. Modelem molekuły np. był układ odpowiednio przestrzennie rozmieszczonych punktów masowych odpowiadających atomom. Odległości między atomami miały być stałe, tak że pod wieloma względami molekuła zachowywała się jak bryła sztywna, mogąca wykonywać tylko ruchy postępowe i obrotowe dokoła środka masy. Modelem gazu był w teorii kinetycznej układ bardzo wielu sztywnych kulek o określonych promieniach i masach, poruszających się bezładnie (ruchem termicznym), mogących się zderzać oraz oddziaływać na odległość siłami van der Waalsa.



Model gazu wg teorii kinetycznej

Stosowanie takich mechanicznych modeli było wynikiem podstawowej dla całej fizyki roli pojęć i praw mechaniki, co jest z pewnością także związane z faktem, że używane w fizyce detektory i przyrządy pomiarowe dokonują z reguły konwersji badanych zjawisk na zjawiska mechaniczno-wizualne, które potrafimy bezpośrednio obserwować i mierzyć. Można wyróżnić kilka podstawowych elementów, z których złożone są modele mechaniczne. Są to: punkty masowe (korpuskuła, cząstka), bryły sztywne, ciągłe

modele mechaniczne

ośrodki materialne i pola (np. pola sił działających na cząstkę lub pola różnych innych wielkości fizycznych, fale). Z elementów tych można złożyć bardzo wiele klasycznych, mechanicznych modeli. Teoria oparta na takim mechanicznym modelu jest zwykle dość poglądowa, a wiele własności można wydedukować, zrozumieć i opisać prawie bez użycia matematyki. Dopiero gdy pytamy o konkretne wartości wielkości fizycznych lub bardziej złożone prawa, musimy skorzystać z pomocy matematyki. Koncepcji podstawowych dostarcza więc w takich modelach mechanika, a matematyka odgrywa tylko rolę pomocniczą; służy mianowicie do ścisłego sformułowania modelu mechanicznego i dokonywania obliczeń.

W XX w. odkryto zjawiska kwantowe, powstały teorie kwantowe, w których związek stosowanych symboli matematycznych (np. operatorów, wektorów stanu, pól kwantowych) z pomiarem jest pośredni i bardzo złożony. Okazało się m.in., że elektron nie jest ani czystą korpuskułą, ani falą, lecz w pewnych sytuacjach przejawia własności korpuskularne, w innych — pozornie sprzeczne z tamtymi — własności falowe. Korpuskularno-falowy, dualny chacharakter elektronu, protonu i innych „cząstek" nie da się wytłumaczyć za pomocą żadnego klasycznego modelu mechanicznego. Jeszcze gorzej przedstawia się sytuacja w zjawiskach, w których obserwujemy kreację i anihilację elektronów i innych cząstek kwantowych.

modele matematyczne

Modele mechaniczne okazały się więc w XX w. nieprzydatne do opisu przynajmniej niektórych zjawisk. Zmusiło to fizyków do coraz częstszego stosowania tzw. modeli matematycznych. Model matematyczny to układ równań wraz z określeniem fizycznego charakteru występujących w nim symboli i ich związku z pomiarami. Sprawa interpretacji tych rownań za pomocą odpowiednich mechanicznych wyobrażeń staje się sprawą drugorzędną. Najważniejszą sprawą jest, by założone równania modelu matematycznego dawały wyniki ilościowo zgodne z doświadczeniem. Model jest tym lepszy, im większy jest zakres jego stosowalności i dokładności oraz im więcej jego przewidywań potwierdziło doświadczenie. We współczesnej fizyce stosuje się — z jednej strony (w charakterze hipotez roboczych) — modele matematyczne o bardzo wąskim zakresie i małej dokładności, z drugiej — modele bardzo dokładnie oddające rzeczywistość w szerokim zakresie zjawisk. W tym ostatnim przypadku mówimy raczej o teorii, a nie o modelu, gdyż tę nazwę rezerwuje się zwykle dla mocno przybliżonych, fragmentarycznych, próbnych schematów teoretycznych, głównie takich, które dotyczą jakiegoś szczególnego, konkretnego obiektu lub zjawiska.

Modele matematyczne są zazwyczaj znacznie mniej poglądowe od mechanicznych. Poglądowość nie jest jednak najważniejszą zaletą teorii. Nie możemy żądać od wszystkich praw natury, by były poglądowe w sensie mechanicznym. Zresztą uznanie teorii za poglądową zależy trochę od przyzwyczajenia, rozwoju nauki, przyswojenia odpowiednich pojęć itp. W każdym razie rozwój fizyki zmusił badaczy do zarzucenia w pewnych sytuacjach klasycznych modeli mechanicznych, ponieważ okazały się one sprzeczne z rzeczywistością.

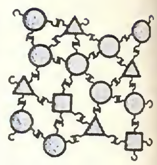
Zarys rozwoju fizyki

Od niepamiętnych, prehistorycznych czasów człowiek obserwował zjawiska przyrody, powodowany zarówno naturalną ciekawością, jak i chęcią praktycznego wykorzystania zauważonych regularności. Od biernej, przypadkowej obserwacji do celowego, czynnego, racjonalno-empirycznego działania naukowego droga jest jednak bardzo daleka. Nie da się ustalić określonej daty, czy nawet stulecia, które można by uznać za początek fizyki, ponieważ działalność badawcza człowieka nabierała bardzo powoli cech zwanych dziś naukowymi. Niewątpliwie można jednak wyróżnić w historii pewne przełomowe okresy, w których zaszły istotne i raczej trwałe zmiany warunkujące postęp nauki.

grecka filozofia przyrody VI–III w. p.n.e.

Przełomowym dla fizyki i wielu innych nauk był z pewnością okres między VI a III w. p.n.e., kiedy to w starożytnej Grecji powstała filozofia rozumiana jako niezależne od religii i magii, racjonalne poszukiwanie prawdy o przyrodzie i o człowieku. Niezmiernie ciekawa jest historia myśli greckiej tego okresu, która od stosunkowo prymitywnych rozważań Talesa nad prapoczątkiem świata przechodziła do coraz bardziej dociekliwych pytań i coraz wnikliwszych i wszechstronniejszych odpowiedzi. W okresie tym powstały liczne racjonalne modele, mające wyjaśnić różne zjawiska przyrody. Choć często propozycje greckich filozofów wydają się dziś bardzo naiwne, to jednak są i takie, które zadziwiają swoją trafnością. Zdecydowanie wyróżnia się wśród nich atomowy model materii, zapoczątkowany przez Leukippa, a rozwinięty przez Demokryta w IV w. p.n.e. Model ten jest stosunkowo bliski atomistycznym ideom współczesnej fizyki.

Wszystkie modele przyrody filozofów greckich opierały się jednak na powierzchownych, jakościowych obserwacjach i spekulatywnych, nie sprawdzonych przesłankach. Były to więc modele czysto racjonalne i jakościowe, nastawione na wyjaśnienie, a nie na ilościowe przewidywanie. Brak odpowiednich kryteriów doświadczalnych nie pozwalał odróżnić propozycji trafnych od nietrafnych, płodnych od bezpłodnych. Wprawdzie występowali w bogatej kulturze greckiej również zwolennicy empiryzmu, był to jednak empiryzm raczej czysto obserwacyjny, bierny i jakościowy, który też nie zdołał się przekształcić w ścisłą racjonalno-empiryczną naukę.



Struktura atomowa materii wg Demokryta

Wydaje się, że filozofowie greccy byli zbyt niecierpliwi i ambitni, że chcieli od razu znaleźć uniwersalne i fundamentalne prawa przyrody, struktury i przyczyny; nie zdawali sobie sprawy, że trzeba przedtem lepiej i dokładniej poznać zjawiska przyrody. Najpierw trzeba wiedzieć, „co i jak", a potem można próbować głębszych odpowiedzi na pytanie typu „dlaczego".

Wielką zasługą i trwałą zdobyczą filozofii greckiej było jednak samo postawienie problemu racjonalnej teorii przyrody, która by tłumaczyła nawet bardzo złożone zjawiska za pomocą niewielkiej liczby stosunkowo prostych elementów składowych i praw. W filozofii i nauce greckiej można znaleźć także podstawowe idee dotyczące jedności materii oraz istnienia uniwersalnych praw przyrody. Znajdziemy tam także niemal wszystkie elementy dzisiejszej metody naukowej, jak indukcja i dedukcja, analiza i konstrukcja pojęć, hipoteza, przyczynowość, racjonalizm, empiryzm, pomiar, eksperyment, obserwacja, matematyczny zapis. Jednak znajdziemy tam również wiele koncepcji i tendencji niepotrzebnych lub wręcz błędnych.

ἄτομος

ἐμπειρία

ἰδέα

μετρον

Dysponując w zasadzie wszystkimi potrzebnymi elementami składowymi procesu badawczego, filozofowie i uczeni greccy nie doszli do odpowiedniej ich syntezy, która jest potrzebna do skutecznego funkcjonowania metody naukowej. Nie opracowali w dostatecznym stopniu koniecznych dla rozwoju fizyki technik eksperymentalnych, nie zwrócili należytej uwagi na rolę pomiaru i eksperymentu, na ścisły, matematyczny opis zjawisk ani ilościową zgodność teorii z doświadczeniem. Toteż nie udało im się uruchomić tak istotnego sprzężenia zwrotnego między teoretyczną hipotezą a faktami doświadczalnymi, które jest podstawą samodoskonalącego mechanizmu empirycznej kontroli, sprawdzania i korygowania wysuwanych hipotez.

I tak np. idealista Platon wysoko cenił matematykę,

Platon i Arystoteles

ale lekceważył zmysłowe poznanie, a empiryk Arystoteles doceniał wprawdzie wagę zmysłowego poznania i logicznego rozumowania, ale posługiwał się głównie metodami jakościowymi i nie doceniał ścisłego matematycznego opisu zjawisk. Żaden z nich nie zdawał sobie sprawy z roli pomiaru i eksperymentu.

W nauce i filozofii starożytności, a także później, aż do końca średniowiecza, brakowało więc podstawowych uzgodnień, co się objawiło w skrajnie przeciwnych postawach filozofów (uczonych) w sprawach zasadniczych dla rozwoju metody naukowej i samych badań. Jeszcze dziwniejsze jest dziś dla nas pomieszanie w działalności tych samych ludzi elementów racjonalnych z irracjonalnymi, materializmu z idealizmem, wyrafinowania logicznego z naiwnością i prymitywizmem przesłanek, grubego zmysłowego empiryzmu z metafizyczną, oderwaną od rzeczywistości spekulacją.

Archimedes



Archimedes (III w. p.n.e.)

Z fizyków starożytności niewątpliwie największym i najbliższym nam ideowo był Archimedes. Jego wielkość polegała nie tylko na odkryciu kilku podstawowych praw mechaniki, lecz także na umiejętności łączenia teorii z praktyką i matematyki z eksperymentem. Wydaje się, że Archimedes trafił już na trop właściwej, skutecznej metody naukowej, ale nie zdążył jej należycie opracować i rozpowszechnić. Nie znalazł też godnych następców i po jego śmierci rozwój fizyki właściwie zatrzymał się na kilkanaście stuleci.

Jedynym działem fizyki, w którym starożytność przekazała wiekom późniejszym wiele cennych, konkretnych wiadomości, była statyka. Natomiast w zagadnieniach ruchu i dynamiki panowały przez niemal 2000 lat zupełnie błędne koncepcje Arystotelesa. To samo można powiedzieć o błędnym fizycznym modelu świata Eudoksosa–Arystotelesa, który był sprzeczny z zasadami jedności materii, powszechności praw fizyki i przyczynowości, a ponadto miał raczej jakościowy i zamknięty charakter. Ogromny autorytet Arystotelesa jako filozofa i twórcy wielu nauk sprawił niestety, iż fizyczna astronomia zatrzymała się w rozwoju.

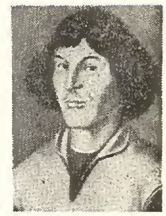
okres odrodzenia

Dopiero w okresie późnego średniowiecza i potem w okresie odrodzenia coraz więcej ludzi zaczęło sobie zdawać sprawę ze słabości ówczesnej, czysto racjonalnej, spekulatywnej nauki. Niektórzy, np. Roger Bacon oraz — dużo później — Francis Bacon, napisali nawet specjalne filozoficzne traktaty, zawierające propozycje nowej empirycznej metody naukowej i wizje nowego społeczeństwa zorganizowanego na naukowych podstawach. Sami jednak nie zastosowali proponowanych metod do systematycznych badań naukowych, nie zdołali więc porwać współczesnych wielkimi odkryciami lub teoriami naukowymi. Inni, jak np. Leonardo da Vinci, dokonali wielu szczegółowych, ale niesystematycznych odkryć i wynalazków, które również nie zostały podjęte, kontynuowane i rozpowszechnione. Ani jedni, ani drudzy nie odegrali więc większej roli w powstaniu nowożytnego przyrodoznawstwa. Decydujące dla jego powstania i rozwoju okazały się dopiero fascynujące przykłady skuteczności nowego stylu myślenia w postaci wielkiej i stymulującej teorii lub w postaci dłuższej i systematycznej serii ważnych i ciekawych odkryć doświadczalnych. Tak więc nie R. Bacon ani F. Bacon i nie Leonardo da Vinci, lecz Kopernik, Galileusz i Newton są ojcami nowożytnej fizyki i całego przyrodoznawstwa.

przełom w XVI i XVII w.

Największy przełom w historii nauk fizycznych dokonywał się w XVI i XVII w. dzięki rozwinięciu w tym okresie skutecznej i niemal całkowicie dojrzałej metody naukowej. Proces rozwijania nowej, racjonalno-empirycznej metody badawczej rozpoczął Mikołaj Kopernik. Jego dzieło okazało się wielką, niezwykle ważną i płodną syntezą renesansowego empiryzmu ze średniowieczną metafizyką. Od humanistów odrodzenia przejął Kopernik antydogmatyczną i empiryczną, dociekliwą i poszukującą, krytyczną postawę, która była wolna od wszelkiego mistycyzmu i walczyła otwarcie z więzami tradycji, wątpliwymi dogmatami i autorytetami. Od swoich filozofujących poprzedników przejął racjonalny i ścisły, a przy tym śmiały teoretyczny sposób myślenia, umożliwiający głębsze poznanie i zrozumienie świata zamiast poprzestawania na jakościowym, czysto zmysłowym poznaniu i powierzchownym opisie obserwowanych zjawisk. Kopernikańska płodna synteza tych uprzednio rozdzielonych i przeciwstawnych lub źle wymieszanych nurtów, postaw i sposobów myślenia stanowiła narodziny nowożytnego przyrodoznawstwa.

Kopernik



Mikołaj Kopernik (1473–1543)

Znajdujemy więc w dziele Kopernika niemal wszystkie podstawowe zasady dzisiejszej metody naukowej, bez tak częstych u jego poprzedników domieszek religijnych, magicznych, mistycznych itp. Kopernik wierzył w sposób oczywisty w realność świata i jego poznawalność za pomocą zmysłowych obserwacji i rozumowania. Jego metoda jest ścisła, oparta na pomiarach i matematyce. W przytaczanych argumentach fizycznych na rzecz układu heliocentrycznego korzystał w sposób widoczny z założenia jednolitości materii i powszechności praw fizyki. Unikał wszelkich irracjonalnych i idealistycznych interpretacji obserwowanych zjawisk.

Kopernik był przekonany o wyższości swojej metody naukowej i krytykował dowolność i aprioryczność, jakościowy charakter i brak uzasadnienia założeń u wielu swych poprzedników. Był przekonany, że dobra teoria powinna być nie tylko ilościowo zgodna z doświadczeniem, ale powinna być również zgodna z „naturą rzeczy". Choć nie jest całkowicie jasne, jak rozumiał to wymaganie, to z jego własnych rozważań można sądzić, że dobra teoria powinna uwypuklać proste symetrie i prawa oraz powinna mieć jakieś głębsze, np. dynamiczne znaczenie. W sposób widoczny Kopernik szukał fizycznego uzasadnienia swojej heliocentrycznej teorii i zaproponował sam ciekawą wielograwitacyjną hipotezę, która była pierwszym krokiem w kierunku prawa powszechnego ciążenia.

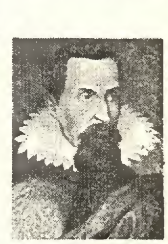
Zasługi Kopernika dla rozwoju fizyki polegają jednak nie tyle na wadze jego własnych odkryć i hipotez, co na wadze problemów fizycznych, jakie dzieło jego otworzyło dla następnych pokoleń astronomów i fizyków. W przeciwieństwie do zamkniętych od strony fizycznej systemów Eudoksosa–Arystotelesa oraz Ptolemeusza teoria Kopernika stanowiła system otwarty. Stała się więc przebogatym źródłem fundamentalnych i płodnych dla rozwoju fizyki pytań dotyczących: kinematycznych i dynamicznych praw i zasad mechaniki, fizycznej struktury układu planetarnego i Wszechświata oraz podstawowych zasad metody naukowej przyrodoznawstwa.

Galileusz



Galileusz, Galileo Galilei (1564–1642)

Tak więc na samym progu ery nowożytnej astronomia, jako najbardziej ilościowa z ówczesnych nauk fizycznych, stała się dzięki Kopernikowi inspiratorką rozwoju całej fizyki. Dzieło Kopernika kontynuował świadomie Galileusz. Uznając i popierając system heliocentryczny Kopernika — ugruntowany na początku XVII w. przez odkrycia Keplera — Galileusz skoncentrował swe wysiłki głównie na ziemskiej fizyce. Przeprowadził systematyczne i wnikliwe pomiary i obserwacje różnych zjawisk. On pierwszy zrozumiał, że wbrew pozorom siła wiąże się z przyspieszeniem, a nie z prędkością. Skonstruował szereg nowych detektorów (termoskop, luneta, waga hydrostatyczna). Odkrył prawo ruchu wahadła, księżyce Jowisza, fazy Wenus itp. Przede wszystkim jednak pogłębił metodę naukową w zakresie detekcji i pomiarów zjawisk oraz wprowadził szereg ważnych dla dalszego rozwoju dynamiki pojęć.



Johannes Kepler (1571–1630)

W XVII w. zostały położone doświadczalne i teoretyczne podwaliny wielu działów fizyki. Zaczęto systematyczne badania gazów i cieczy oraz ciał sprężystych, zjawisk optycznych i falowych (W. Snell, R. Boyle, B. Pascal, Ch. Huygens, O. Römer, P. de

Fermat, Torricelli, Otto von Guericke i in.). Następcy Galileusza skonstruowali w XVII w. wiele nowych detektorów, np. baroskop do badania zmian ciśnienia, spektroskop pryzmatyczny do badania składu światła, zegary astronomiczne.

Nowe, ilościowe metody fizyki, rozwijane przez Galileusza i jego następców, wymagały adekwatnych metod matematycznych. Niejako na zamówienie fizyki stworzył R. Descartes geometrię analityczną, a I. Newton i G.W. Leibniz — rachunek różniczkowy i całkowy.

Newton

Słynne dzieło I. Newtona *Philosophiae naturalis principia mathematica* (1687) zawiera już kompletną koncepcję nowożytnej fizyki jako nauki ścisłej, zarazem eksperymentalnej i teoretycznej, opartej na pomiarach i szerokim stosowaniu matematyki. W pracy tej Newton sprecyzował pojęcia przestrzeni, czasu, układu odniesienia, masy, siły, pędu itd. Podał też trzy podstawowe prawa dynamiki, zwane często ze względu na ich wagę zasadami mechaniki. Pierwsza zasada postuluje istnienie inercjalnych układów odniesienia; druga podaje równanie wiążące siłę z pochodną pędu; trzecia dotyczy pewnych ogólnych własności symetrii sił wzajemnego oddziaływania między ciałami fizycznymi. Newton odkrył również prawo powszechnego ciążenia jako uniwersalne prawo natury, dotyczące wszelkich ciał fizycznych. Korzystając z podanych przez siebie praw ruchu, wyrażonych w ścisłej matematycznej postaci, rozwiązał wiele konkretnych problemów właściwej fizyki oraz astronomii. Szczególnie ciekawe rezultaty dały prawa Newtona zastosowane do ruchu planet. Okazało się, że ruch punktu masowego w centralnym polu sił grawitacyjnych odwrotnie proporcjonalnych do kwadratu odległości odbywa się po krzywych stożkowych (koło, elipsa, parabola, hiperbola) w zależności od warunków początkowych. Newton nie tylko fizycznie uzasadnił system Kopernika i trzy prawa Keplera, ale podał jednolitą teorię ruchów ciał na Ziemi oraz planet, komet i innych ciał niebieskich. W ten sposób astronomia stała się po prostu częścią fizyki, a jednocześnie prawa ziemskiej mechaniki objęły co najmniej cały układ planetarny. Zasady jedności materii i powszechności praw fizyki znalazły w mechanice układu planetarnego swoje pierwsze ścisłe potwierdzenie.

$$F = G\frac{mM}{R^2}$$

$$G = 6{,}672 \times 10^{-11}\,\mathrm{Nm^2/kg^2}$$

Stała grawitacyjna



Sir Isaac Newton (1642–1727)

Newton przyczynił się również niezwykle do rozwoju optyki, nauki o cieple, mechaniki ośrodków ciągłych itd. Choć jego zasługi dotyczą głównie teorii, to jednak dokonał on także wielu odkryć eksperymentalnych (np. odkrycie dyspersji światła, rozszczepienia światła białego, dyfrakcji i interferencji w cienkich płytkach, odkrycie prawa stygnięcia ciał, praw lepkości).

Newton podał w jasnej i ścisłej postaci zasady nowoczesnej, empiryczno-racjonalnej metody naukowej, która jest stosowana bez zasadniczych zmian do dziś. Dopiero wiek XX przyniósł pewne pogłębienie jego metody naukowej. W ciągu minionych prawie 300 lat metodę tę przejęły bez większych zmian inne nauki przyrodnicze i technika, a ostatnio adaptują ją również nauki społeczne i humanistyczne.

XVIII w. — rozwój mechaniki

Mechanika Newtona stanowiła przez niemal 200 lat wzór teorii fizycznej, źródło inspiracji dla filozofów i przedmiot powszechnego zainteresowania ludzi wykształconych. Następcy Newtona rozwijali przez cały XVIII w. mechanikę teoretyczną, stosując prawa Newtona do różnych typów sił i różnych układów, np. układów wielu punktów i ośrodków ciągłych (gazów, cieczy, ciał sprężystych). W związku z potrzebami doskonalonej mechaniki teoretycznej powstały w tym okresie nowe działy matematyki, np. teoria równań różniczkowych zwyczajnych i cząstkowych, teoria równań całkowych, rachunek wariacyjny itp. Wielu spośród następców Newtona (np. L. Euler, J.L. Lagrange, J. d'Alembert, J.D. Bernoulli, P.S. Laplace, C.G.J. Jacobi, W.R. Hamilton) przyczyniło się tak jak on do rozwoju zarówno fizyki, jak i matematyki. Jeśli chodzi o ścisłą i niezwykle owocną współpracę fizyki z matematyką, był to z pewnością złoty wiek nauki, który trwał do połowy XIX stulecia.

rozwój badań eksperymentalnych

Wiek XVIII cechuje nie tylko wspaniały rozwój mechaniki teoretycznej, znamionują go także ważne osiągnięcia eksperymentalne w innych działach fizyki. Skonstruowano lub w istotny sposób udoskonalono wiele nowych detektorów i przyrządów pomiarowych (np. termometr, barometr, kalorymetr, teleskopy różnych typów, waga analityczna, waga skręceń). Poznano podstawowe prawo elektrostatyki (Ch.A. Coulomb), odkryto prawo zachowania masy w reakcjach chemicznych (A.L. Lavoisier, M.W. Łomonosow), rozwinęła się szybko kalorymetria. Rewolucja przemysłowa w Anglii przyspieszyła rozwój badań eksperymentalnych.

Pierwszą połowę XIX w. charakteryzuje znaczna liczba przełomowych odkryć doświadczalnych z zakresu nauk o cieple, elektryczności, magnetyzmie oraz optyki i chemii — przy jednoczesnym braku dobrych teorii badanych zjawisk. W dziedzinie zjawisk cieplnych odkryto równoważność pracy i ciepła (B. Rumford, H. Davy, J.P. Joule). Stwierdzono też, że praca może się przemieniać w ciepło bez żadnych ograniczeń, natomiast ciepło można przemienić w pracę tylko z pewnymi ograniczeniami (S. Carnot). Odkrycia te zadały pierwszy poważny cios błędnej hipotezie cieplika, przypisującej ciepłu substancjalny, niezniszczalny charakter. W chemii odkryto prawa stosunków wielokrotnych, stosunków równoważnikowych i stosunków stałych (J.L. Proust, J.B. Richter, L.J. Gay-Lussac, J. Dalton, A. Avogadro). W celu wytłumaczenia tych praw Dalton przypomniał starą hipotezę atomową. W optyce dokładne zbadanie zjawisk interferencji i dyfrakcji światła (T. Young, A.J. Fresnel) obaliło korpuskularny model światła, którego zwolennikiem był sam Newton. Konstrukcja siatki dyfrakcyjnej (J. Fraunhofer) i różnych typów spektrometrów oraz odkrycie linii absorpcyjnych w widmie słonecznym stało się początkiem spektrometrii optycznej. Astronomia otrzymała nowe, potężne narzędzia do badania chemicznego składu zewnętrznych warstw Słońca, gwiazd i planet oraz procesów zachodzących w tych warstwach, jak również procesów zachodzących w przestrzeni międzygwiazdowej. Spektrometria znalazła też rychło liczne zastosowania do badania składu chemicznego i zanieczyszczeń różnych substancji, dzięki możliwości bardzo dokładnego wykrywania charakterystycznych dla poszczególnych pierwiastków linii widmowych. Innym ważnym dla optyki osiągnięciem było odkrycie polaryzacji światła odbitego (E.L. Malus). Po skonstruowaniu efektywnych polaryzatorów (W. Nicol) rozpoczął się okres intensywnych badań zjawiska polaryzacji światła (A.J. Fresnel, T. Young, E. Malus, D.F.J. Arago, D. Brewster). J.B.L. Foucault zmierzył prędkość światła w wodzie i wykazał, że jest ona zgodna z modelem falowym, natomiast zdecydowanie sprzeczna z modelem korpuskularnym światła. Zjawisko polaryzacji świadczyło, że światło jest falą poprzeczną.

rozwój nauki o elektryczności

W dziedzinie elektryczności skonstruowano wiele nowych urządzeń służących do wytwarzania pól i prądów elektrycznych oraz pól magnetycznych (np. stosy, ogniwa, akumulatory, maszyny elektrostatyczne, kondensatory, cewki) oraz do pomiaru wielkości elektrycznych i magnetycznych (galwanometr, woltomierz, amperomierz, busola stycznych). Za pomocą ogniw odkryto i zbadano zjawisko elektrolizy (H. Davy, M. Faraday), a następnie za pomocą elektrolizy wykryto ważne pierwiastki nie występujące na Ziemi w stanie wolnym (np. potas i sód) oraz stwierdzono złożony charakter niektórych substancji (np. wody). W 1820 r. H.Ch. Oersted, badając własności prądów stałych, odkrył, że prąd elektryczny ma coś wspólnego z magnetyzmem, ponieważ odchyla igłę busoli. Wkrótce A.M. Ampère odkrył wzajemne mag-



Michael Faraday (1791–1867)

netyczne oddziaływanie dwóch prądów. G.S. Ohm wykrył prawo przewodnictwa metali, zwane dziś jego imieniem.

Największe zasługi dla rozwoju nauki o elektryczności i magnetyzmie położył jednak w 1 połowie XIX w. M. Faraday. Odkrył on w 1831 r. bardzo ważne prawo indukcji, wg którego zmienne w czasie pole magnetyczne powoduje wystąpienie zamkniętych linii sił pola elektrycznego. W tym samym czasie Faraday odkrył zjawisko indukcji prądów przez zmienne prądy, co było zresztą naturalną konsekwencją zjawiska Oersteda i pierwotnego prawa indukcji. Odkrycia te wykazywały istnienie głębokich związków między elektrycznością i magnetyzmem, wprowadzały nowe obiekty badań: zmienne w czasie prądy i pola elektryczne oraz magnetyczne, inspirowały konstrukcje nowych ważnych urządzeń w postaci cewki, elektromagnesu, galwanoskopu, transformatora, które wkrótce zrewolucjonizowały technikę, a z nią całą naszą cywilizację. Faraday dokonał jeszcze wielu innych wiekopomnych odkryć, zdobywając sławę najlepszego eksperymentatora, jeśli nie w całej historii fizyki, to przynajmniej XIX w. Wystarczy wymienić jeszcze prawa elektrolizy (stała Faradaya), skroplenie gazów, odkrycie diamagnetyków i paramagnetyków, zbadanie dielektryków, odkrycie skręcenia polaryzacji światła przy przechodzeniu przez namagnesowane ośrodki.

$F = 9{,}648 \times 10^4$ C/mol

Stała Faradaya

Faraday był nie tylko genialnym eksperymentatorem, był równie genialnym teoretykiem. Wprawdzie nie sformułował matematycznie teorii zjawisk elektromagnetycznych, ale wprowadził podstawowe koncepcje teoretyczne. Wyniki jego doświadczeń skłoniły go do odrzucenia panującego wówczas mechanistycznego obrazu zjawisk elektromagnetycznych, w którym podstawową rolę odgrywały ładunki elektryczne i magnetyczne działające na odległość siłami centralnymi odwrotnie proporcjonalnymi do kwadratu odległości. Faraday wprowadził inny opis, oparty na pojęciu lokalnego działania pola elektrycznego i magnetycznego na znajdujący się w nim ładunek elektryczny lub prąd. Taki model stał się bardzo płodny w przyszłości, kiedy się okazało, że szybko zmienne pola elektromagnetyczne mogą się oderwać od wytwarzających je ładunków i rozchodzić niezależnie od chwilowych położeń tych ładunków.

Mimo tak wielu, tak fundamentalnych odkryć eksperymentalnych, konstrukcji wielu nowych detektorów, urządzeń i aparatów pomiarowych, we wszystkich działach fizyki, oprócz właściwej mechaniki, brak było zadowalających teorii i panował w nich dość duży nieład. Dopiero w połowie XIX w. nastąpił przełom pod tym względem. Zaczęło się od sformułowania w latach czterdziestych I zasady termodynamiki, a wkrótce, w latach 1850–51 — II zasady (S. Carnot, R. Mayer, J.P. Joule, H. Helmholtz, R. Clausius, W. Thomson-Kelvin). I zasada jest uogólnieniem prawa zmiany energii, znanego z mechaniki, na procesy, w których oprócz pracy występuje wymiana ciepła. II zasada określa kierunek zachodzących w przyrodzie procesów termodynamicznych. Z zasad tych wynika istnienie temperatury bezwzględnej i entropii. Obie zasady termodynamiki implikują ścisły związek między zjawiskami mechanicznymi i cieplnymi, które przez długie wieki były uważane za niezależne od siebie. Choć zasady termodynamiki są sformułowane zupełnie ogólnie (abstrahują od szczególnych własności układów) oraz czysto fenomenologicznie, to jednak zastosowanie ich do konkretnych ciał implikuje wiele ciekawych związków między bardzo różnorodnymi zjawiskami i procesami.

zasady termodynamiki

hipoteza atomowo-molekularna

W połowie XIX w. została ostatecznie przyjęta przez chemików hipoteza atomowo-molekularna (St. Cannizzaro), a — w związku z powstaniem termodynamiki i ze wzrastającym zainteresowaniem zjawiskami cieplnymi — stała się ona domeną teoretycznych badań fizyków. Zarówno bowiem termodynamika, jak i powstała kilkanaście lat później elektrodynymika były teoriami czysto fenomenologicznymi i makroskopowymi. Występowały w nich liczne empiryczne parametry i funkcje stanu (np. temperatura, ciśnienie, równanie stanu, entropia, energia wewnętrzna, stała dielektryczna, przewodnictwo elektryczne), których teorie te nie potrafiły ani zinterpretować, ani obliczyć.

Problem wyjaśnienia zjawisk makroświata jako wypadkowych odpowiednich własności atomów i molekuł (czyli mikroświata) podjął w 1845 r. J.J. Waterston. Szybki rozwój tego kierunku badań datuje się jednak dopiero od prac A.K. Kröniga i R.E. Clausiusa rozpoczętych w 1856 r. Badacze ci zbudowali bardzo poglądowy model atomowo-molekularny gazów, zwany teorią kinetyczną gazów. Począwszy od 1860 r. do szybkiego rozwoju tej teorii przyczynił się J.C. Maxwell. Na gruncie tej pierwszej ilościowej teorii mikroświata zinterpretowano temperaturę, ciśnienie, entropię i inne makroskopowe funkcje stanu, wyrażając je przez wartości średnie odpowiednich wielkości mikroskopowych. W ten sposób zjawiska cieplne zostały sprowadzone do odpowiednich zjawisk mechanicznych zachodzących w mikroświecie. Wychodząc z założenia, że między molekułami gazu występuje proste oddziaływanie, udało się też wytłumaczyć postać równania stanu gazów (van der Waals). Zastosowanie znalezionego przez Maxwella rozkładu prędkości w jednorodnym gazie o stałej temperaturze pozwoliło na uściślenie obliczeń wielu parametrów i uczyniło teorię kinetyczną znacznie bardziej realistyczną. Maxwell podał też prostą mikroskopową teorię takich zjawisk jak przewodnictwo cieplne, lepkość i dyfuzja gazów.

kinetyczna teoria gazów

Stworzenie zadowalającej matematycznej teorii zjawisk elektromagnetycznych i magnetycznych jest również zasługą Maxwella. Dokonał on tego w serii prac rozpoczętych w 1861 r., a uwieńczonych ogłoszeniem w 1873 r. wielkiego dzieła *Treatise on Electricity and Magnetism.* W pracach tych Maxwell przełożył najpierw prawa Gaussa, Ampère'a–Oersteda i Faradaya na ścisły język matematycznych równań różniczkowych, posługując się przy tym polowymi koncepcjami Faradaya. Równanie odpowiadające pierwotnemu prawu Oersteda okazało się w przypadku gęstości ładunku zależnej od czasu sprzeczne z prawem zachowania ładunku. Maxwell zmienił to równanie i doszedł do niesprzecznego układu, zwanego dziś równaniami elektrodynamiki klasycznej albo po prostu równaniami Maxwella. Z równań elektrodynamiki wynikało istnienie nowego, nie znanego wówczas zjawiska fal elektromagnetycznych. Maxwell wykazał, że powinny to być fale poprzeczne o ściśle określonych właściwościach. Zwrócił też uwagę na to, że fale elektromagnetyczne mają wszystkie znane wówczas własności światła. Założył więc w zgodzie ze wspomnianymi doświadczeniami Faradaya, że światło jest po prostu falą elektromagnetyczną o stosunkowo małej długości fali. Istnienie fal elektromagnetycznych o dużych długościach zostało potwierdzone przez H.R. Hertza wiele lat później, dopiero w 1888 r.

elektrodynamika Maxwella



James Clerk Maxwell (1831–1879)

Po licznych i bardzo różnorodnych weryfikacjach teorii Maxwella, w końcu XIX w. nikt już nie wątpił w jej poprawność. Podobnie jak w końcu XVII w. mechanika Newtona, tak w 2 połowie XIX w. elektrodynamika Maxwella oznaczała niezwykle ważny przewrót w strukturze fizyki. W ramach jednej wspólnej teorii zostały połączone trzy wielkie klasy zjawisk (elektryczne, magnetyczne i optyczne), które przez długi okres były uważane za niezależne. Pominąwszy elektrostatykę i magnetostatykę, pola elektryczne i magnetyczne są zawsze nierozłączne, a związki między nimi są opisane przez równanie Maxwella. Dlatego mówi się obecnie poprawniej o zjawiskach elektromagnetycznych. Optyka przestała istnieć jako niezależny dział fizyki, a stała się działem elektrodynamiki (→ Elektrodynamika).

$$\operatorname{div}\vec{D} = \varrho$$

$$\operatorname{rot}\vec{E} = -\frac{\partial \vec{B}}{\partial t}$$

$$\operatorname{div}\vec{B} = 0$$

$$\operatorname{rot}\vec{H} = \frac{\partial \vec{D}}{\partial t} + \vec{j}$$

Równania Maxwella

ziarnistość elektryczności

W ostatniej ćwierci XIX w. zaczęły się mnożyć doświadczenia wykazujące coraz bardziej przekonywająco, że elektryczność występuje w postaci ziarnistej, a nie jest ciągłym fluidem jak sądzono dawniej. Wskazywały na to już prawa elektrolizy. Znacznie większą rolę odegrały tu jednak badania wyładowań elektrycznych w rozrzedzonych gazach (J. Plücker, H. Geissler, J. Hittorf, W. Crookes, E. Goldstein). Wykazano m.in., że promienie katodowe składają się z ujemnie naładowanych cząstek. Dokonane po 1897 r. pomiary stosunku masy do ładunku (J.J. Thomson, W. Kaufmann, E. Wiechert) doprowadziły do wniosku, że cząstki te, nazwane elektronami, są przeszło 1800 razy lżejsze od najlżejszego atomu wodoru. Przeprowadzone niezależnie pomiary ładunków elektrycznych (R.A. Millikan) wykazały, że są one zawsze całkowitymi wielokrotnościami ładunku elektronu.

pierwsze badania mikroświata

Wyładowania w rozrzedzonych gazach stały się pierwszą bardziej bezpośrednią metodą eksperymentalnego badania mikroświata. Zastosowanie niskich ciśnień, a tym samym małych gęstości, oznaczało poważne zmniejszenie wzajemnych oddziaływań, czyli stworzenie warunków coraz lepszej wzajemnej izolacji mikrocząstek. Poza tym w eksperymentach tych udało się po raz pierwszy wytworzyć wiązki jednakowych naładowanych mikrocząstek (protonów, elektronów, jonów) o bardzo zbliżonych pędach. Wprawdzie historia fizyki niskich ciśnień zaczęła się już w XVII w., od Torricellego i Guerickego, ale dopiero w 2 połowie XIX w. nastąpił jej gwałtowny rozwój — głównie dzięki temu, że stała się ona podwaliną niemal wszystkich metod badania mikroświata.

elektronowa teoria Lorentza

Jeszcze przed definitywnym doświadczalnym odkryciem elektronu H.A. Lorentz stworzył podwaliny pierwszej mikroskopowej, statystycznej teorii zjawisk elektromagnetycznych, nazwanej później elektronową teorią Lorentza. Lorentz przyjął, zgodnie z późniejszymi odkryciami doświadczalnymi, że atomy i molekuły zbudowane są z elektrycznie naładowanych cząstek. Założył dalej, że w mikroświecie słuszne są równania Maxwella w postaci znanej dla próżni — z tym zastrzeżeniem, że jedynym prądem elektrycznym jest prąd konwekcyjny, wywołany ruchem ładunków. Wprowadzając jeszcze kilka innych poglądowych założeń i utożsamiając makroskopowe wielkości z wartościami średnimi odpowiednich wielkości mikroskopowych, z modelu Lorentza można było otrzymać wiele wniosków dotyczących elektromagnetycznych własności ciał.

atomistyka podstawą unifikacji fizyki

W tym samym okresie do rozwoju kinetycznej teorii gazów w kierunku coraz głębszej i coraz wszechstronniejszej statystycznej teorii materii ogromnie się przyczynili przede wszystkim L. Boltzmann, a potem J.W. Gibbs. Po odkryciu elektrycznej struktury materii oraz powstaniu teorii Lorentza i statystycznej fizyki zaczął się na przełomie XIX i XX w. nowy etap unifikacji całej makroskopowej fizyki na bazie atomistyki, zakładającej, że wszystkie ciała makroskopowe składają się z określonych mikrocząstek (atomów, molekuł, jonów, elektronów). Własności tych mikrocząstek i prawa nimi rządzące miały determinować wszystkie własności ciał makroskopowych i prawa makroświata.

Program ten był w owym czasie bardzo trudny do zrealizowania. Nie znano jeszcze struktury atomów, molekuł i jonów, nie znano najważniejszych praw mikroświata. Wnioskowanie o własnościach mikroświata na podstawie ich hipotetycznego wpływu na zjawiska makroświata było bardzo trudne, ponieważ średniowanie nie jest procedurą jednoznacznie określoną ani odwracalną. W tej sytuacji na przełomie XIX i XX stulecia powstała silna opozycja przeciwko całej atomistyce, wskazująca na brak bezpośrednich metod doświadczalnych, które by potwierdziły realne istnienie i pozwoliły na zbadanie struktury atomów i innych mikrocząsteczek. Wielkim, ale krótkotrwałym triumfem przeciwników atomistyki było wykrycie w tym okresie wielu fundamentalnych sprzeczności między przewidywaniami teorii statystycznych a faktami doświadczalnymi. W rezultacie sprzeczności te doprowadziły do powstania ok. 1925 r. nowej, poprawnej teorii atomu i molekuły — mechaniki kwantowej.

obserwacja pojedynczych mikrocząstek

Również żądanie bardziej bezpośrednich, doświadczalnych dowodów istnienia mikrocząstek zostało wkrótce spełnione dzięki konstrukcji nowych detektorów i innych narzędzi badawczych, jak np. komory Wilsona, pęcherzykowe, iskrowe, liczniki różnego rodzaju naładowanych i neutralnych cząstek, czułe emulsje fotograficzne, mikroskopy elektronowe, dalej — liczne akceleratory przygotowujące strumienie pożądanych mikrocząstek o określonym pędzie itp. (→ Detekcja cząstek wielkiej energii, Akceleratory). Urządzenia te pozwoliły nie tylko na mierzenie nowych, statystycznych wielkości odnoszących się bezpośrednio do określonych mikrocząstek (jak przekroje czynne, rozkłady kątowe i energetyczne oraz inne rozkłady prawdopodobieństwa), lecz także na wizualne oglądanie torów pojedynczych naładowanych mikrocząstek oraz ich zderzeń, procesów tworzenia się i rozpadów, a nawet obserwację większych drobin (w mikroskopie elektronowym).

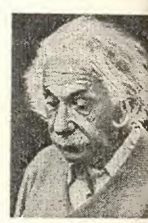
W ten sposób odpowiedź na trzy fundamentalne pytania eksperymentatora: Co, jak i czym mierzyć w fizyce mikroświata — stawała się stopniowo coraz bogatsza i głębsza.

sprzeczność między mechaniką i elektrodynamiką

Inny ważny kierunek badań z końca XIX i początku XX w. wiązał się z narastającą sprzecznością między dwoma największymi teoriami makroświata: mechaniką Newtona a elektrodynamiką Maxwella. Według elektrodynamiki prędkość rozchodzenia się fal elektromagnetycznych w próżni c jest stała i izotropowa. Newtonowska koncepcja przestrzeni i czasu prowadzi do wniosku, że jest to możliwe tylko w jednym wyróżnionym układzie inercjalnym. W innych poruszających się względem niego układach prędkość c powinna zależeć od kierunku rozchodzenia się fali.

pomiar prędkości światła

W końcu XIX w. A.A. Michelson i E.W. Morley wykonali serię doświadczeń, które miały wykazać, jak prędkość Ziemi dodaje się do prędkości światła. Okazało się — wbrew fizyce Newtona — że prędkość światła w dwóch poruszających się wzajemnie układach jest taka sama i że jest izotropowa w obu układach. Usiłowania pogodzenia obu teorii bez zmieniania żadnej z nich nie udały się. W 1905 r. A. Einstein zrobił zgodne z wynikiem doświadczeń Michelsona–Morleya założenie, że prędkość światła jest stała i jednakowa we wszystkich układach inercjalnych, oraz, że wszystkie układy inercjalne są fizycznie równoprawne i nie ma układu wyróżnionego (eteru). Einstein wykazał, że współrzędne przestrzenne i czasowe tego samego zdarzenia, mierzone za pomocą sygnałów świetlnych w dwóch poruszających się wzajemnie ruchem jednostajnym układach inercjalnych, są związane tzw. transformacją Lorentza, odmienną od dotychczas stosowanej transformacji Galileusza. Z transformacji Lorentza wynikały zaskakujące konsekwencje, np. względność takich pojęć jak przedział czasowy, kolejność czasowa, odległość, masa, które w fizyce Newtona były uważane za bezwzględne, tj. niezależne od wyboru układu inercjalnego. Einstein pogodził elektrodynamikę Maxwella z mechaniką Newtona zmieniając tę ostatnią, gdyż jej równania nie były niezmiennicze względem transformacji Lorentza. Najważniejszą zmianą było zastąpienie w prawie dynamiki Newtona stałej masy masą zależną od prędkości. Ze zmienionej w ten sposób mechaniki wyniknął słynny wzór wyrażający równoważność masy i energii. Einstein założył, że wzór ten jest słuszny w przypadku układów zawierających nie tylko cząstki, lecz również pola (→ O niektórych podstawowych pojęciach fizycznych).



Albert Einstein (1879–1955)

Transformacja Lorentza implikuje skomplikowane prawo dodawania prędkości, które jednakże dla prędkości małych w porównaniu z nieprzekraczalną

szczególna teoria względności

według Einsteina prędkością światła c przechodzi w dawne proste prawo dodawania wektorów. To samo dotyczy innych równań, tak że można powiedzieć, że teoria Einsteina, zwana też szczególną teorią względności, nie jest sprzeczna z dawną mechaniką, tylko obejmuje szerszy zakres zjawisk. Dawna mechanika pozostaje słuszna w zakresie prędkości dużo mniejszych od c.

Zaskakujące konsekwencje szczególnej teorii względności zostały potwierdzone w tysiącach eksperymentów i w praktycznym zastosowaniu (np. w akceleratorach szybkich cząstek, w reakcjach między cząstkami elementarnymi i jądrami, w reaktorach i bombach atomowych). Do tych sukcesów doprowadziła Einsteina systematyczna rewizja podstawowych pojęć dawnej fizyki i wprowadzenie — zamiast starych, apriorycznych definicji — określeń operacyjnych, opartych na użyciu fal elektromagnetycznych jako optymalnego środka przesyłania sygnałów i wszelkiej informacji. Z przykładu tego wynika, że po pierwsze — trzeba być bardzo ostrożnym z pojęciami, które nie są wprowadzone operacyjnie, lecz apriorycznie, po drugie — nie należy zakładać pochopnie stosowalności teorii w zakresie jeszcze nie sprawdzonym.

Kilkanaście lat później Einstein zaproponował nową, wysoce zgeometryzowaną teorię grawitacji, która nosi też nazwę ogólnej teorii względności

ogólna teoria względności

(opiera się ona na lokalnej równoważności zjawisk występujących w przypieszonym układzie i w polu grawitacyjnym). W teorii tej postuluje się, że obecność mas grawitacyjnych zakrzywia przestrzeń, która już nie jest prostą przestrzenią Euklidesa, lecz przestrzenią typu Riemanna. Teoria ta daje w pierwszym przybliżeniu te same wyniki co mechanika Newtona. Otrzymane w następnym przybliżeniu odstępstwa od teorii Newtona są bardzo małe, tak że do dziś zostały potwierdzone jedynie w paru przypadkach. Teoria grawitacji Einsteina stała się jednak podstawą nowoczesnej kosmologii i prowadzi do wielu ciekawych wyników (→ Kosmologia).

Jak już wspomniano wyżej, po długim stosunkowo okresie rozwoju i coraz większych sukcesów, na przełomie XIX i XX stulecia atomistyka znalazła się

kryzys atomistyki na początku XX w.

nagle w stanie głębokiego kryzysu. Przede wszystkim okazało się, że wbrew pierwotnej koncepcji atomu — odkryte przez chemików i fizyków realne atomy nie są ani niepodzielne, ani niezniszczalne, ani nie są najmniejszymi cząstkami materii. Wskazywało na to odkrycie elektronu jako składnika atomu oraz odkrycie naturalnej promieniotwórczości szeregu pierwiastków (H. Becquerel, M. Skłodowska-Curie i P. Curie). Rychło wykazano, że promienie alfa i beta występujące w tych zjawiskach są korpuskularne i że emitujący je pierwiastek przekształca się w inny pierwiastek (E. Rutherford, W. Crookes, W. Ramsay, K. Fajans, F. Soddy).

Z drugiej strony zwrócono uwagę na szereg faktów doświadczalnych (np. kształt widma promieniowania ciała doskonale czarnego, ciepło właściwe wieloatomowych gazów i ciał stałych, mnogość linii widmowych nawet najprostszych atomów, zjawisko fotoelektryczne), które się nie dały pogodzić z ówczesnymi prawami fizyki makroskopowej postulowanymi w fizyce statystycznej również w odniesieniu do obiektów mikroświata. Zaczęto więc szukać dodatkowych ograniczeń, które by z jednej strony — usunęły powstałe rozbieżności, a z drugiej — nie były rażąco sprzeczne ze znanymi prawami, choć może obce i niepotrzebne w fizyce makroświata.

Ten nowy etap rozwoju atomistyki zapoczątkował w 1900 r. M. Planck, proponując rozwiązanie problemu promieniowania ciała doskonale czarnego za

kwant energii

pomocą dodatkowego założenia kwantów energii (tj. określonych porcji), które może wypromieniować lub zaabsorbować oscylator harmoniczny. Następny krok zrobił w 1905 r. Einstein wprowadzając pojęcie fotonu (cząstki lub kwantu światła) dla wytłumaczenia

foton

efektu fotoelektrycznego.



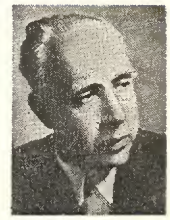
Ernest Rutherford (1871–1937)

Tymczasem badano pilnie widma promieniowania i inne własności atomu, odkrywając szereg empirycznych reguł (np. reguła kombinacyjna Ritza). Przełomowe znaczenie miało jednak dopiero odkrycie w 1911 r. dodatnio naładowanego jądra atomowego przez Rutherforda. Opierając się na tym odkryciu, N. Bohr już w 1913 r. zaproponował nowy model atomu. Model Bohra był półklasyczny, tzn. opierał się na równaniach mechaniki Newtona uzupełnionych trzema dodatkowymi kwantowymi postulatami. Model Bohra wyjaśniał częściowo podstawowe własności atomów z jednym elektronem, ale nie odpowiadał na wiele istotnych pytań, nie pozwalał na obliczenie wielu ważnych własności atomów, nie nadawał się do obliczeń bardziej złożonych układów.

model atomu Bohra

powstanie mechaniki kwantowej

W 1924 r. L. de Broglie wysunął śmiałą hipotezę, że podobnie jak światłu (fotonom) przypisano własności korpuskularne w oddziaływaniu z materią, tak i typowym cząstkom (np. elektronom) należy przypisać własności falowe. De Broglie podał wzór wiążący długość fali z pędem elektronu, ale nie podał równania falowego. Zrobił to wkrótce E. Schrödinger, proponując swoje słynne równanie, które się stało filarem nowej mechaniki mikroświata, zwanej mechaniką kwantową (lub falową). Inne, nieco ogólniejsze, bardziej abstrakcyjne (operatorowe), ale w zasadzie równoważne sformułowanie mechaniki kwantowej podał W. Heisenberg. W następnych paru latach ewolucja mechaniki kwantowej polegała na wprowadzeniu probabilistycznej interpretacji funkcji falowej, na uwzględnieniu świeżo odkrytego spinu elektronu, zakazu Pauliego lub statystyki Bosego–Einsteina w przypadku układów wielocząstkowych oraz efektów relatywistycznych (P.A.M. Dirac) i na wprowadzeniu innych formalnych i fizycznych uogólnień i udoskonaleń.

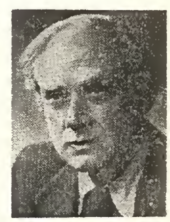


Niels Bohr (1885–1962)



Louis de Broglie (ur. 1892)

Nowa teoria świata atomów, molekuł, jonów i elektronów dawała znakomite rezultaty. Pozwoliła na dokładne obliczenie energii wiązania, poziomów wzbudzonych, prawdopodobieństw przejścia, czasów życia i wielu innych efektów, wobec których dawne teorie i modele były bezradne. Gdy mamy do czynienia z prostszymi układami, rachunki nie są nawet specjalnie trudne, a ich zgodność z danymi doświadczalnymi znakomita. W przypadku wielocząstkowych układów równania mechaniki kwantowej są wprawdzie bardzo skomplikowane, ale numeryczne rachunki wszędzie dają dobrą zgodność teorii z doświadczeniem.



Paul Adrien Maurice Dirac (ur. 1902)

Mechanika kwantowa dokonała poważnych zmian w podstawowych pojęciach fizyki. W fizyce przedkwantowej, zwanej dziś klasyczną, przez symbole oznaczające w równaniach teorii pęd, położenie, moment pędu itp. należało rozumieć dostatecznie dokładne wyniki pomiarów tych wielkości, czyli liczby. W mechanice kwantowej wielkości fizyczne są reprezentowane w równaniach przez operatory, których związek z pomiarem jest dość złożony, a mianowicie dopiero wartości własne operatora reprezentującego daną wielkość są liczbami, które się pokrywają z możliwymi wynikami pomiaru. Nieprawdziwe okazują się w związku z tym dwa milczące

rewizja podstawowych pojęć

założenia teorii pomiaru robione w fizyce klasycznej: a) zakłócenie stanu dowolnego obserwowanego ciała spowodowane przez przyrząd pomiarowy można uczynić dowolnie małym; b) można przeprowadzić jednoczesne pomiary dwóch dowolnych, algebraicznie niezależnych wielkości fizycznych z dowolną dokładnością i tak, by pomiar jednej nie zakłócał pomiaru drugiej. O nieprawdziwości tych założeń świadczą m.in. znane relacje nieokreśloności Heisenberga. Metodologiczne i filozoficzne konsekwencje mechaniki kwantowej są nawet ciekawsze i dalej idące niż teorii względności.

Z biegiem czasu okazało się, że mechanika kwantowa nie opisuje całkiem poprawnie oddziaływania elektronów, atomów itp. z polem promieniowania

elektrodynamika kwantowa — połowa XX w.

elektromagnetycznego (np. nie daje możliwości obliczenia efektu Comptona czy anihilacji pary negaton–pozyton). W celu opisania takich i wielu innych efektów trzeba było wprowadzić tzw. drugie kwantowanie, w wyniku którego powstała w połowie XX w. kwantowa elektrodynamika (S. Tomonaga, F.J. Dyson, R.P. Feynman, J. Schwinger). Teoria ta opisuje wszelkie układy złożone z elektronów i pola elektromagnetycznego w doskonałej zgodzie z doświadczeniem.

Jest rzeczą niezwykle interesującą, że równania mechaniki kwantowej Schrödingera można wyprowadzić z równań elektrodynamiki kwantowej jako pewne przybliżenie. Inne przybliżenie, zwane klasycznym, prowadzi od elektrodynamiki kwantowej do klasycznych równań Maxwella i równań Newtona. Elektrodynamikę kwantową można więc uznać za najbardziej fundamentalną z istniejących sprawdzonych teorii fizycznych. Jest ona teorią, która opisuje w sposób jednolity ogromny zakres zjawisk. Dla większości zjawisk zachodzących w świecie atomów, molekuł, jonów oraz złożonych z nich ciał makroskopowych wystarczająco dokładną teorią jest zresztą już mechanika kwantowa. Opisuje ona nie tylko wszystkie fizyczne i chemiczne własności wymienionych mikrocząstek, ale za pośrednictwem fizyki statystycznej opisuje doskonale również własności ciał makroskopowych. Tak to już w XX w. dokonał się następny bardzo ważny krok w kierunku zjednoczenia coraz większej liczby działów fizyki i innych nauk przyrodniczych w ramach wspólnej, jednolitej teorii. Objęcie niemal całej makroskopowej fizyki oraz chemii, a przypuszczalnie także biologii, przez jedną fizyczną teorię, którą jest elektrodynamika kwantowa lub w nieco skromniejszym wymiarze mechanika kwantowa, ma bardzo proste i poglądowe uzasadnienie. Jest nim atomowa struktura wszystkich normalnych ciał oraz czysto elektromagnetyczny charakter wszystkich istotnych w tym zakresie oddziaływań elementarnych.

uniwersalność elektrodynamiki kwantowej

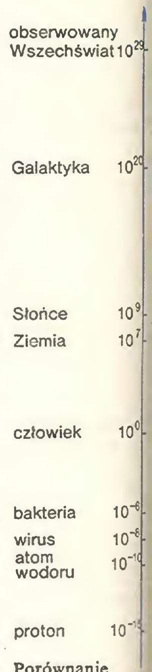
Poza zasięgiem elektrodynamiki kwantowej znajdują się tylko fizyka jądrowa i fizyka cząstek elementarnych oraz niektóre zjawiska zachodzące w skali kosmicznej, np. we wnętrzu gwiazd, gdzie zresztą oprócz grawitacji istotną rolę odgrywają własności jąder i cząstek elementarnych.

fizyka jądrowa

Jądro atomowe zostało odkryte już w 1911 r. przez Rutherforda przy okazji jego badań wnętrza atomu. Jednakże rozwój fizyki jądrowej jako odrębnej dyscypliny rozpoczął się właściwie dopiero w latach trzydziestych. Jak zwykle w fizyce, nowe obiekty badań wymagają najpierw stworzenia nowych narzędzi i metod pomiarowych. Podstawową metodą badania jąder stało się ich bombardowanie różnymi cząstkami i śledzenie zachodzących przy tym rozproszeń oraz reakcji. Rozwój fizyki jądrowej, a następnie fizyki cząstek elementarnych, wymagał konstrukcji coraz potężniejszych akceleratorów oraz coraz czulszych liczników i detektorów cząstek. Pierwsze systematyczne badania sztucznie wywołanych reakcji jądrowych rozpoczęto ok. 1930 r. (I. i F. Joliot-Curie, J. Chadwick i in.). W r. 1932 odkryto drugi obok protonu składnik jądra — neutron. Od tego czasu nasza wiedza empiryczna o stanach podstawowych i wzbudzonych jąder oraz o reakcjach jądrowych wzrosła ogromnie, a jednak nie ma jeszcze zadowalającej ogólnej teorii jądra. Zgodnie z hipotezą H. Yukawy przyjmuje się, że oddziaływania między nukleonami przenoszone są głównie przez tzw. pole mezonowe (mezony π). Jednakże oddziaływań tych nie udało się dotychczas opisać teoretycznie w zadowalający sposób (→ Oddziaływania silne, Siły jądrowe). Zamiast tego stosuje się w fizyce jądrowej liczne fragmentaryczne modele, które oddają dość dobre usługi, pozwalając zrozumieć podstawowe własności jąder (→ Modele jądrowe). Ponieważ nukleony oddziałują silnie nie tylko między sobą i z mezonami π, lecz także z innymi cząstkami elementarnymi, trudno uwierzyć, by zadowalająca teoria jądra mogła powstać przed powstaniem teorii cząstek elementarnych.

fizyka cząstek elementarnych

Fizyka cząstek elementarnych jest właściwie najmłodszym z wielkich działów fizyki; w odrębną dyscyplinę wydzieliła się z fizyki jądrowej dopiero w latach pięćdziesiątych, po odkryciu, oprócz cząstek μ i mezonów π, szybko rosnącej liczby nowych, z reguły nietrwałych cząstek (hiperony: Λ, Σ, Ξ, Ω, mezony: k, η, ω, ϱ i cała plejada tzw. rezonansów). Obiekty te traktuje się na równi z elektronami, protonami i neutronami, ponieważ mają one wiele własności cząstek elementarnych (podobne elementarne oddziaływania i symetrie oraz niepodzielność, małe rozmiary itp.). Znamy już bardzo wiele takich obiektów i ciągle odkrywamy nowe. Znamy też sporo ciekawych empirycznych praw i reguł oraz ogromną liczbę reakcji między cząstkami elementarnymi. Nie ma jednak jeszcze zadowalającej teorii cząstek elementarnych (→ Cząstki elementarne i ich oddziaływania).



Porównanie rozmiarów

Ponieważ cząstki elementarne są najmniejszymi (ok. 10^{-13} cm) znanymi cegiełkami, z których zbudowane są wszystkie znane nam formy materii, fizycy przypuszczają, że ich własności powinny wyznaczać własności wszystkich obiektów zarówno mikroświata, jak i makroświata, a więc zarówno jąder, jak i materii gwiezdnej. O własnościach atomów istotnych dla chemii i fizyki normalnych ciał makroskopowych decydują — na szczęście — tylko bardzo nieliczne globalne własności jąder, jak ładunek, masa i spin. Struktura samego jądra, osłoniętego bardzo grubą otoczką elektronową, jest nieistotna dla wspomnianego zakresu zjawisk; ważna jest tu tylko struktura zewnętrznych powłok elektronowych atomów i ich masy. Natomiast, aby obliczyć własności samych jąder (którymi możemy się interesować niezależnie) oraz badać konsekwencje pewnych ekstremalnych stanów w makroświecie, konieczna jest dobra teoria cząstek elementarnych.

Przyszła teoria cząstek elementarnych ma więc wszelkie szanse stania się najbardziej ogólną i fundamentalną teorią fizyczną, zawierającą kwantową elektrodynamikę i teorię jądra jako przypadki szczególne lub graniczne. Według obecnego stanu fizyki możemy powiedzieć, że cała materia w jej wszystkich znanych nam stanach jest zbudowana z tych samych cząstek elementarnych, których oddziaływania podlegają tym samym prawom. Jest to niezwykle poglądowe i przekonywające fizyczne uzasadnienie zasad jedności materii i powszechności praw przyrody.

przyszła teoria cząstek elementarnych

Powstaje frapujące pytanie, czy teoria cząstek elementarnych obejmuje swoim zasięgiem całą materię, czy też będzie ona tylko następnym krokiem po nieskończonej drabinie poznania.

W niniejszym, z natury rzeczy bardzo pobieżnym, szkicu uwzględnione zostały tylko te działy fizyki, które się znalazły w głównym nurcie poznawczym fizyki, wyznaczonym przez powstawanie coraz ogólniejszych i coraz głębszych przyczynowych teorii. Jest wiele działów fizyki, które mają ogromne znaczenie praktyczne i są również ciekawe pod względem poznawczym, ale nikt nie spodziewa się po nich odkryć o fundamentalnym dla rozwoju całej fizyki znaczeniu poznawczym czy też konstrukcji zupełnie nowej ogólnej teorii. Takimi działami są: fizyka ciała stałego, fizyka cieczy, fizyka plazmy, optyka, termodynamika procesów nieodwracalnych itd. Nikt nie ma wątpliwości, że istniejące teorie — mechanika kwantowa i elektrodynamika kwantowa oraz fizyka statystyczna — opisują dobrze wymienione działy fizyki, ale wobec skomplikowanego charakteru występujących w nich układów ścisła teoria jest przydatna tylko jako punkt wyjścia do licznych uproszczonych modeli. Ze względu na bogactwo występujących w takich układach różnorodnych zjawisk praca badawcza jest tu nie tylko bardzo ciekawa, ale ma przede wszystkim kapitalne znaczenie praktyczne.

Zarys niniejszy ogranicza się do właściwej fizyki.

O chemii czy astronomii wspomniano tylko wówczas, gdy nauki te przyczyniły się w sposób istotny do rozwoju głównego nurtu fizyki. Trzeba jednak podkreślić, że w astrofizyce odkryto w XX w. wiele nowych, fascynujących zjawisk, które mogą mieć ogromne znaczenie dla poznania struktury i ewolucji Wszechświata. Ogromne znaczenie zarówno poznawcze, jak i praktyczne ma też fizyka życia, a w szczególności biologia molekularna.

Praktyczne i kulturotwórcze znaczenie fizyki

technika a nauka

Technika jest z pewnością dużo starsza od nauki; stniała i rozwijała się, gdy nauki jeszcze nie było. Wobec braku naukowej, teoretycznej i empirycznej podbudowy techniki, jej rozwój był w dawnych czasach niezmiernie powolny, zależał najczęściej od szczęśliwego przypadkowego odkrycia, niekiedy — od genialnego pomysłu. Również utrwalanie i przekazywanie nabytych praktycznych umiejętności było w tych warunkach bardzo utrudnione. Zresztą nawet przez pierwsze paręset lat historii nowożytnej fizyki jej szybki rozwój w niewielkim zaledwie stopniu oddziaływał na technikę. Z wyjątkiem wynalazku maszyny parowej technika tego okresu opierała się niemal wyłącznie na osiągnięciach mechaniki starożytnej, a więc głównie na statyce i teorii mechanizmów.

Przełom nastąpił dopiero w 2 połowie XIX w., po powstaniu elektrodynamiki, a także termodynamiki. Praktyczne wykorzystanie termodynamiki do budowy silników cieplnych i wielu zagadnień przemysłowej chemii, ciepłownictwa itp. postępowało stosunkowo wolno. Natomiast zastosowanie elektrodynamiki potoczyło się od razu lawinowo. Rozpoczął się „wiek elektryczności", który trwa nadal. Szerokie wykorzystanie zjawisk elektromagnetycznych spowodowało przede wszystkim zupełną rewolucję sił wytwórczych i podniesienie poziomu materialnego i kulturalnego społeczeństw. Rewolucja ta opierała się głównie na wprowadzeniu wydajnych, niezawodnych i ogólnie dostępnych elektromagnetycznych metod przetwarzania, przesyłania, akumulowania i coraz bardziej zróżnicowanego wykorzystania taniej energii. Z drugiej strony — nie mniejszą rewolucję społeczno-kulturalną spowodowały elektromagnetyczne metody zbierania, zapisu, przesyłania, przetwarzania, odtwarzania i powielania wszelkiej informacji, w tym także informacji czysto kulturalnej Trzeci zakres zastosowania elektrodynamiki dotyczy automatyzacji procesów produkcyjnych i nieprodukcyjnych, zdalnego kierowania itp.

wiek elektryczności

Aż do drugiej wojny światowej technika wykorzystywała tylko makroskopowe zjawiska i prawa klasycznej fizyki. W ostatnich 30 latach coraz częściej wykorzystuje się zjawiska kwantowe z zakresu fizyki atomu, molekuły i ciała stałego, kwantowej optyki oraz fizyki plazmy, które się zajmują zjawiskami elektromagnetycznymi na poziomie fizyki mikroświata.

W tym samym wprawdzie czasie wzrosła liczba nowych zastosowań mechaniki i termodynamiki, ale ich wagi nie da się porównać z wagą zastosowań zjawisk elektromagnetycznych. Prymatu pod tym względem nie zdołała też odebrać elektrodynamice fizyka jądrowa — mimo swego wielkiego wpływu na politykę (bomby jądrowe). Zastosowanie izotopów, reaktorów jądrowych, różnego typu promieniowania, choć bardzo ważne, nie jest jeszcze tak powszechne i tak istotne ani dla przemysłu, ani dla praktyki pozaprzemysłowej jak zastosowanie elektrodynamiki. Warto zresztą podkreślić, że fizyka jądrowa nie występuje bynajmniej w roli konkurenta elektrodynamiki, tzn. nie ma mowy o wypieraniu metod elektromagnetycznych przez jądrowe. Chodzi z reguły o dodanie do możliwości elektrodynamiki nowych możliwości, wynikających z wykorzystania fizyki jądrowej.

zastosowanie fizyki jądrowej

Warto wspomnieć tu także o drugiej, równoczesnej rewolucji, którą spowodowało w ostatnim stuleciu zastosowanie chemii. Chemia nie tylko udoskonaliła technologię wytwarzania wielu „starych" materiałów, ale wprowadziła do produkcji ogromną liczbę nowych tworzyw sztucznych, paliw, leków, nawozów, detergentów, środków spożywczych itp.

zastosowanie chemii

Dzięki coraz szerszemu wykorzystaniu zdobyczy fizyki i chemii oraz w pewnym stopniu biologii nastąpił w ostatnich 100 latach niebywale szybki rozwój przemysłu. Zrodzona przez te nauki nowa technika zmienia w zawrotnym tempie oblicze świata, decyduje o bogactwie materialnym i o poziomie kulturalnym narodów, zmienia siły wytwórcze, rodzaj i skalę wytwarzanych dóbr, zmienia stosunki społeczne i polityczne we wszystkich częściach naszego globu.

W pierwszym okresie unaukowienia techniki, który trwał z grubsza do połowy XX w., technika korzystała niemal wyłącznie z końcowych rezultatów badań naukowych, czyli z gotowej wiedzy naukowej w postaci odkrytych zjawisk, skonstruowanych aparatów i teorii. Był to okres dominacji indywidualnych wynalazców, którzy eksploatowali fizykę w sposób „chałupniczy", polegający na wielokrotnych próbach albo opartych na zasadzie „chybił trafił", albo na tzw. genialnych pomysłach wynalazczych, albo też na dość nieuchwytnej i niepowtarzalnej „technicznej smykałce".

Rozpoczętą w połowie XX w. rewolucję naukowo-techniczną charakteryzuje zerwanie z poprzednimi, „chałupniczymi" i nienaukowymi metodami wykorzystania fizyki do celów praktycznych. Współczesna fuzja techniki z fizyką polega więc na powszechnym stosowaniu metod naukowych w intensywnie i systematycznie prowadzonych pracach zespołowych, zmierzających do optymalnego wykorzystania poznanych przez nauki fizyczne zjawisk i praw natury.

rewolucja naukowo-techniczna

Dopiero bardzo niedawno zdano sobie sprawę, że do praktycznego wykorzystania nadaje się nie tylko produkt końcowy pracy fizyków, którym jest gotowa wiedza naukowa, lecz także stosowana przez fizyków metoda naukowa, styl myślenia i działania. Fizyka bowiem nie tylko zaspokaja nasz głód poznania i naturalną ciekawość świata, jest nie tylko przebogatą skarbnicą aparatów, materiałów, zjawisk, praw i teorii tworzących szeroką bazę techniki, fizyka stworzyła również nowy styl myślenia i wszelkiego działania opartego na racjonalno-empirycznej, ścisłej metodzie naukowej. Okazuje się, że metody opracowane pierwotnie w badaniach naukowych do celów czysto poznawczych można zaadaptować w coraz szerszym zakresie do praktycznego działania. Adaptacja ta zasadza się m.in. na podobieństwie struktury omówionego wyżej cyklu badawczego i typowego cyklu planowego działania praktycznego, który się składa z następujących elementów: stan początkowy, zadanie, plan, realizacja planu, analiza planu i jego realizacji (ewentualność zmian i korekt), stan końcowy. Uruchomienie skutecznie działających sprzężeń zwrotnych między zadaniem a planem i jego realizacją, na wzór sprzężenia między problemem, hipotezą i testami empirycznymi, jest dziś jednym z podstawowych problemów rewolucji naukowo-technicznej.

metoda naukowa w działalności praktycznej

Podstawą i warunkiem skuteczności wszelkiej działalności praktycznej jest umiejętność trafnego przewidywania przebiegu ważnych dla praktyki procesów oraz rezultatów ludzkiego działania. Dobrze uzasadnione podstawy maksymalnie trafnego przewidywania daje dziś tylko nauka. W skomplikowanym i szybko zmieniającym się świecie dawne źródła umiejętności przewidywania, a mianowicie doświadczenie życiowe, praktyka, tradycja i historia, są dziś wysoce niewystarczające.

Ogromna rola fizyki w tworzeniu kultury materialnej jest — mimo wszelkie nieporozumienia i wypacze-

wkład fizyki do kultury umysłowej

nia — znacznie lepiej znana i utrwalona w świadomości społecznej niż znaczenie fizyki dla kultury duchowej (zwanej też często kulturą właściwą), a ściślej — kultury umysłowej.

Fizyka odgrywa ogromną rolę kulturotwórczą; niestety, na ogół nie docenianą. Fizyka stworzyła przede wszystkim nowy, niezależny od religii i magii, naukowy obraz całej przyrody nieożywionej. Wprawdzie fizyczny, materialistyczny obraz świata nie jest ani absolutnie dokładny, ani ostateczny, ale jest z pewnością dokładniejszy i bliższy rzeczywistości od wszelkich obrazów nienaukowych. Obraz ten jest nie tylko sprawdzalny, ale także ustawicznie rozszerzany, pogłębiany i korygowany. Ten niepełny, ale ciągle doskonalony obraz świata urzeka już dziś swoją prostotą, jednolitością oraz jasnością i komunikatywnością.

Wiele elementów fizycznego obrazu świata weszło już do kultury ogólnej szerokich rzesz społeczeństwa. Pojęcia cząstki elementarnej, atomu, molekuły, jądra, masy, prędkości, przyspieszenia, siły, prądu elektrycznego, pól elektrycznych i magnetycznych, fal, praw ruchu, prawa powszechnego ciążenia, układu heliocentrycznego itp. są powszechnie używane i rozumiane (choć nie zawsze całkiem poprawnie) przez większość ludzi nie tylko ze średnim lub wyższym wykształceniem, lecz nawet przez ludzi o wykształceniu podstawowym. Język pojęć fizyki i ilościowy, racjonalno-empiryczny styl myślenia fizyków stał się w znacznym stopniu dobrem powszechnym, wszedł do języka ogólnego, jest używany potocznie, w prasie, w radio i telewizji, a nawet w ... poetyckich porównaniach.

Dobrze sprawdzony i sprawdzalny, obiektywny obraz całego świata materialnego oraz naukowy styl myślenia są niezwykle cennymi dobrami kulturalnymi, które fizyka oferuje społeczeństwu. Już w XVIII w. powstanie mechaniki Newtona, tłumaczącej w jednolity sposób ruchy planet i ciał na Ziemi, stało się przedmiotem powszechnego zainteresowania ludzi wykształconych w kulturalnie zaawansowanych krajach. Fizyka Newtona stała się źródłem inspiracji dla filozofów i działaczy, przyczyniła się walnie do powstania ważnego prądu umysłowego, zwanego oświeceniem. Dzisiaj bezpośrednie oddziaływanie fizyki na kulturę umysłową jest jeszcze większe, choć brak perspektywy czasu utrudnia nam często wykrycie głównych nurtów i kierunków.

Podobnie jak oddziaływanie fizyki na produkcję i konsumpcję dóbr materialnych odbywa się za pośrednictwem techniki, tak też część oddziaływania fizyki na kulturę właściwą odbywa się za pośrednictwem literatury i sztuki. Do literatury zaczęła wchodzić fizyka i inne nauki przyrodnicze już w 2 połowie XIX w. Odkrycia naukowe i oparte na nich wynalazki i pomysły techniczne pobudzały fantazję pisarzy.

fizyka a literatura

Rozpoczął się trwający do dziś bujny rozwój literatury fantastyczno-naukowej oraz powieści przygodowej i detektywistycznej, zawierającej dużo informacji i pomysłów z zakresu fizyki tudzież innych nauk przyrodniczych oraz techniki. Wiek XX przyniósł znaczne pogłębienie tego kierunku, który znajduje wyraz także w rozkwicie tzw. literatury faktu. W społeczeństwach XX w. wystąpił bowiem stymulowany przez naukę głód szeroko pojętej wiedzy, opartej nie na dyletanckich obserwacjach oraz osobistych, subiektywnych przeżyciach pisarzy z bożej łaski, lecz na rzetelnych, naukowo stwierdzonych faktach. Coraz częściej bestsellerem staje się w naszym wieku nie powieść, lecz pamiętnik interesującej osoby, zbiór reportaży, esejów, rozpraw czy książka popularnonaukowa z zakresu archeologii, historii, psychologii, biologii, socjologii, medycyny, chemii, fizyki, astronomii, a nawet matematyki. Ba, niejednokrotnie bestsellerami stają się ładnie napisane książki o wyraźnie naukowym charakterze.

Tworzenie dobrych dzieł literatury faktu jest trudniejsze niż tworzenie dobrych fikcji powieściowych. Oprócz piękna języka, polotu i fantazji wymaga się dobrego przygotowania fachowego, znajomości i dobrego zrozumienia przedstawionych zjawisk, koncepcji i teorii, zdolności syntetycznego widzenia oraz daru prostego wyjaśniania spraw nader skomplikowanych. Połączenie piękna słowa i formy pisarskiej z głęboką, konkretną, a przecież fascynującą treścią, opartą na faktach dostarczanych przez naukę i praktykę, wydaje się realistycznym i pociągającym programem rozwoju literatury.

nowe dziedziny kultury

Jeszcze większy i łatwiejszy do prześledzenia jest wpływ fizyki na kulturę artystyczną. Na bazie fizyki i chemii powstały i nadal powstają nowe dziedziny twórczości artystycznej, których w ogóle nie było 200, 100 czy nawet 50 lat temu. Jako najważniejsze przykłady można tu podać fotografię, kino, radio, telewizję i rodzącą się dopiero sztukę holografii. Są to uznane dziś, nowe dziedziny twórczości, które pozornie wypierają stare — w rodzaju malarstwa, muzyki, śpiewu, tańca i teatru — ale w gruncie rzeczy przyczyniają się do niebywałego ich upowszechniania.

Również muzyka i sztuki plastyczne zmieniają się bardzo dzięki zastosowaniu opartych na fizyce i chemii nowych technik i nowych materiałów. Wystarczy przypomnieć zastosowanie elektroniki i akustyki w muzyce albo rolę nowych materiałów w architekturze.

Należy też podkreślić, że fizyka i chemia stworzyły liczne możliwości wiernego zapisu, utrwalenia, przesyłania i dowolnie częstej, wiernej reprodukcji nietrwałych dóbr kultury. To, co dawniej ginęło bezpowrotnie, możemy teraz utrwalić i powtórzyć w dowolnym czasie. Na fotografii, filmie, płycie gramofonowej lub taśmie magnetycznej możemy utrwalić dowolny występ muzyczny, taneczny, teatralny lub wokalny i odtworzyć kiedy mamy tylko czas i ochotę. Możemy w krótkim czasie porównać wykonanie tego samego utworu przez różne, może nie istniejące już zespoły, możemy oglądać i słuchać zmarłych już wykonawców, brać niemal bezpośredni udział w minionych lub geograficznie odległych, a ważnych i ciekawych zdarzeniach kulturalnych i politycznych. Czy nie jest to wspaniałe rozszerzenie i wzbogacenie możliwości kulturalnych każdego człowieka? Czy nasi dziadowie mieli takie możliwości?

fizyka a muzyka i sztuki plastyczne

Nauka jest dziś nie tylko ważną częścią kultury ogólnej, nie tylko oddziałuje coraz silniej na wszystkie niemal formy kultury materialnej i duchowej, ale staje się w naszych oczach główną, dominującą siłą kulturotwórczą, wyznaczającą kierunki rozwoju kulturalnego oraz inspirującą powstawanie nowych prądów umysłowych i nowych form działalności kulturalnej. Przodującą rolę odgrywa w tym procesie fizyka, jako najbardziej podstawowa nauka przyrodnicza i baza nowoczesnej techniki.

M. Bunge *Foundations of Physics*, Berlin 1967; L.N. Cooper *Istota i struktura fizyki*, Warszawa 1975; A. Einstein i L. Infeld *Ewolucja fizyki*, Warszawa 1959; R.P. Feynman i in. *Feynmana wykłady z fizyki* t. 1–3, Warszawa 1972–74; J. Werle *Rozwój i perspektywy fizyki*, Warszawa 1969.

PODSTAWOWE POJĘCIA I TEORIE

O niektórych podstawowych pojęciach fizycznych · Zasady zachowania · Termodynamika fenomenologiczna · Termodynamika statystyczna · Przejścia fazowe i zjawiska krytyczne · Elektrodynamika · Teoria pola ·

O niektórych podstawowych pojęciach fizycznych

Andrzej Staruszkiewicz

Fizyka klasyczna

Terminem „fizyka klasyczna" określa się zespół pojęć i praw ruchu uważanych za prawdziwe od połowy XVII do końca XIX w. Za początek fizyki klasycznej należy uważać rozpoznanie przez fizyków XVII w., zwłaszcza I. Newtona, prawdziwości prawa ruchu nazwanego później pierwszym prawem Newtona: ciało, na które nie działa żadna siła, pozostaje w spoczynku lub porusza się ruchem jednostajnym i prostoliniowym. Twierdzenie to nie jest ani oczywiste, ani nawet zrozumiałe bez dalszych wyjaśnień.

I prawo Newtona

Człowiek płynący statkiem porusza się względem przedmiotów na brzegu, lecz pozostaje w spoczynku względem statku. O jakim więc ruchu i jakim spoczynku mówi pierwsze prawo? Jak można stwierdzić, że na ciało nie działa żadna siła? Jeszcze w drugiej połowie XIX w. Ernst Mach uznał, że pytania te podważają cały gmach mechaniki klasycznej; krytyka Macha wywarła pewien wpływ na Alberta Einsteina — twórcę szczególnej i ogólnej teorii względności.

Newton rozumiał pierwsze prawo w sposób, który być może — jak utrzymywali krytycy od G.W. Leibniza do Macha — jest filozoficznie niedorzeczny, lecz za to trafnie ujmuje fizyczną istotę zagadnienia: gdyby usunąć wszystkie rzeczy wypełniające przestrzeń, tak jak usuwa się płyn z naczynia, to pozostałby pusty pojemnik przedmiotów materialnych — przestrzeń Euklidesowa; gdyby następnie do przestrzeni tej wprowadzić małą cząstkę i nadać jej prędkość, to kolejne położenia cząstki ułożyłyby się wzdłuż linii prostej; gdyby wreszcie na prostej tej zaznaczyć ciąg punktów tak, by między dwoma sąsiednimi była zawsze ta sama odległość, to chwile przebywania cząstki w tych punktach byłyby oddzielone jednakowymi odstępami czasu.

Krytyka newtonowskiego rozumienia pierwszego prawa jest łatwa, lecz jałowa; chwalebne skądinąd dążenie do ścisłości logicznej musi być poprzedzone wiedzą o faktach; w mechanice klasycznej podstawowym faktem jest bezwładność ciał, z której zdaje sprawę pierwsze prawo ruchu.

Współcześnie w celu jasnego ujęcia zasad mechaniki wprowadzamy pojęcie czterowymiarowej czasoprzestrzeni Galileusza. Jest to zbiór wszystkich zdarzeń, tj. procesów fizycznych zachodzących w ściśle określonym miejscu i ściśle określonym czasie, jak np. zderzenie dwu nieskończenie małych cząstek. Za zdarzenie uważamy przy tym nie sam proces fizyczny, lecz wyłącznie jego miejsce w czasie i przestrzeni. G.W. Leibniz uważał taką abstrakcję za niedopuszczalną i na tej zasadzie odrzucał mechanikę Newtona. Czasoprzestrzeń Galileusza ma określoną strukturę afiniczną i metryczną. Struktura afiniczna polega na tym, że można wyróżnić pewne linie jako linie proste i pewne pary prostych jako proste równoległe. Struktura metryczna polega na tym, że istnieje wyróżniona rodzina równoległych hiperpłaszczyzn, zwanych hiperpłaszczyznami równoczesności. Każda z nich jest trójwymiarową przestrzenią Euklidesową. Nadto z każdą parą hiperpłaszczyzn równoczesności, np. z hiperpłaszczyznami *1* i *2*, związana jest liczba t_{12} — upływ czasu newtonowskiego między *1* i *2;* dla trzech hiperpłaszczyzn *1, 2, 3,* $t_{13} = t_{12}+t_{23}$.

czasoprzestrzeń Galileusza

hiperpłaszczyzny równoczesności

Mówimy, że współrzędne czasoprzestrzenne t, x_1, x_2, x_3 są dostosowane do struktury czasoprzestrzeni Galileusza, jeżeli hiperpłaszczyzny równoczesności dane są równaniem $t = \text{const}$, jeżeli upływ czasu newtonowskiego między dwiema takimi hiperpłaszczyznami jest równy różnicy współrzędnej t i jeżeli kwadrat odległości dwu równoczesnych i nieskończenie bliskich zdarzeń wynosi $(dx_1)^2+(dx_2)^2+(dx_3)^2$. O współrzędnych czasoprzestrzennych dostosowanych mówimy, że tworzą układ inercjalny.

układ inercjalny

Terminu „układ inercjalny" używa się w dwu różnych znaczeniach. Oznacza on albo współrzędne czasoprzestrzenne dostosowane do struktury metrycznej czasoprzestrzeni — rozumie się wtedy, że współrzędne te pokrywają całą czasoprzestrzeń — albo niewielkie ciało, np. laboratorium, wolne w dostatecznym stopniu od wpływu innych ciał i nie wykonujące ruchu obrotowego. Gdy mówimy: równania Newtona przyjmują najprostszą postać w układzie inercjalnym, to mamy na myśli pierwsze znaczenie; gdy mówimy: dla czasów rzędu kilku sekund laboratorium spoczywające na Ziemi może być uważane za układ inercjalny, to mamy na myśli drugie znaczenie. W dalszym ciągu artykułu używamy terminu „układ inercjalny" w obu znaczeniach.

Pierwsze prawo ruchu mówi, że w układzie inercjalnym ruch cząstki, na którą nie działają żadne siły, spełnia równanie:

$$\frac{d^2x_i}{dt^2} = 0, \quad i = 1, 2, 3.$$

Jeżeli współrzędne t, x_1, x_2, x_3 tworzą układ inercjalny, to współrzędne:

$$t' = t, \quad x'_i = x_i + v_i t$$

(gdzie v_i są stałymi) też tworzą układ inercjalny; jest to matematyczny wyraz zasady względności Galileusza, która mówi, że przy pomocy żadnego doświadczenia mechanicznego nie można odróżnić spoczynku od ruchu po prostej ze stałą prędkością.

Jak można wyznaczyć hiperpłaszczyzny równoczesności? Newton nie powiedział tego jasno. Wyznaczenie hiperpłaszczyzn równoczesności nabrało jednak wielkiej wagi w XVIII w. w związku z rozwojem żeglugi; znając bowiem hiperpłaszczyzny równoczesności można z położenia gwiazd oznaczyć długość geograficzną. Parlament angielski wyznaczył nagrodę za opracowanie metody oznaczania długości; połowę nagrody dostał zegarmistrz londyński Harrison, którego zegar przewieziony z Londynu do Singapuru i z powrotem wskazywał czas różniący się tylko o kilka sekund od czasu wskazywanego przez podobny zegar pozostawiony w Londynie. Przyznając nagrodę Harrisonowi parlament zgodził się milcząco na następujące określenie równoczesności: niech w układzie inercjalnym spoczywają dwa jednakowe zegary sprężynowe wskazujące ten sam czas; jeżeli jeden z nich przenieść w inne miejsce, to wskazanie czasu t przez zegar przeniesiony pozostanie zdarzeniem równoczesnym ze wskazaniem czasu t przez zegar pozostawiony w spoczynku.

klasyczne określenie równoczesności

Określenie to nie ogranicza ani drogi przeniesienia, ani prędkości, a więc zakłada, że upływ czasu newtonowskiego dt wskazywany przez zegar sprężynowy jest różniczką zupełną; rzecz jasna, zmiany prędkości przy przenoszeniu powinny być tak łagodne, by nie powstały uszkodzenia mechaniczne zegara. Powyższym określeniem równoczesności posługujemy się do dziś w życiu codziennym.

II prawo Newtona

Drugie prawo ruchu Newtona rozwija myśl zawartą w pierwszym: skoro ciało oddzielone zupełnie od innych porusza się ze stałą prędkością, to każda zmiana prędkości wywołana jest obecnością innych ciał. Każde ciało ma wg Newtona masę bezwładną, która jest tym większa, im trudniej jest zmienić jednostajny i prostoliniowy ruch ciała. Drugie prawo we współczesnym ujęciu mówi, że iloczyn masy bezwładnej i przyspieszenia równy jest w każdej chwili sile przyłożonej do ciała:

$$m\frac{d^2x_i}{dt^2} = f_i, \quad i = 1, 2, 3.$$

Sądzi się czasem, że drugie prawo jest tylko określeniem pojęcia siły za pomocą czysto kinematycznego pojęcia przyspieszenia. Newton i tu rozumiał rzecz głębiej; w dziele *Optyka* napisał: „całe zadanie filozofii przyrody polega na tym, by z obserwowanych zjawisk odczytać siły, a następnie z sił wyprowadzić dalsze zjawiska".

teoria grawitacji

Pięknym rozwiązaniem tak rozumianego podstawowego zadania fizyki jest newtonowska teoria grawitacji. Newton znał prawa ruchu planet Keplera. Udowodnił, że prawa te wynikają z założenia, że między Słońcem a planetą działa siła przyciągająca odwrotnie proporcjonalna do kwadratu odległości planety od Słońca i skierowana wzdłuż prostej łączącej oba ciała. Na tej zasadzie przyjął, że między każdymi dwoma ciałami we Wszechświecie działa siła przyciągająca proporcjonalna do iloczynu mas, odwrotnie proporcjonalna do kwadratu odległości i skierowana wzdłuż prostej łączącej oba ciała.

masa bezwładna i masa ciężka

Masa, o której mówi prawo grawitacji, jest pojęciem różnym od pojęcia masy bezwładnej, o której mówi drugie prawo ruchu. Jest jednak faktem doświadczalnym, że obie te masy są zawsze równe. Najprostszym dowodem tego jest ruch wahadła matematycznego, którego okres nie zależy od materiału, z którego wykonano wahadło. Równość masy ciężkiej i masy bezwładnej, będąca jedynie faktem w teorii grawitacji Newtona, stała się dla Einsteina punktem wyjścia do sformułowania ogólnej teorii względności. Odległość występująca w prawie grawitacji jest odległością wzdłuż hiperpłaszczyzn równoczesności; stąd widać, że sformułowanie prawa grawitacji zakłada opisaną wcześniej strukturę metryczną czasoprzestrzeni Galileusza.

Prawo grawitacji pozwala przewidzieć na drodze czysto dedukcyjnej zjawiska nie objęte prawami Keplera, np. perturbacje, które do ruchu planety wprowadza sąsiedztwo innej planety.

III prawo Newtona

Trzecie prawo ruchu mówi, że jeżeli ciało A działa na ciało B siłą $\vec{f}_{AB}$, to ciało B działa na ciało A siłą $\vec{f}_{BA} = -\vec{f}_{AB}$. Status trzeciego prawa jest wyraźnie niższy niż status pierwszego i drugiego prawa. Po pierwsze, np. dla układu ciał oddziaływających grawitacyjnie, trzecie prawo jest zawarte w prawie grawitacji. Po drugie, w szczególnej teorii względności, która jest uogólnieniem fizyki klasycznej, pierwsze prawo pozostaje bez zmian, drugie wymaga tylko niewielkiej zmiany, podczas gdy trzecie może się ostać tylko po znacznych zmianach w sposobie rozumienia oddziaływań między ciałami.

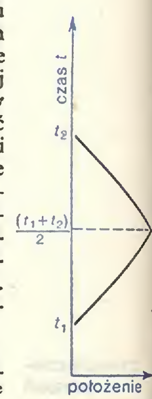
Szczególna teoria względności

„Szczególna teoria względności" jest to wyjątkowo niefortunny termin, którym obejmujemy zespół pojęć i praw ruchu uważanych od początku XX w. za dokładniejsze od praw fizyki klasycznej. Zasady szczególnej teorii względności zostały podane przez Einsteina w 1905 r. Einstein stwierdził, że o ile równoczesność zdarzeń nieskończenie bliskich jest koincydencją przestrzenno-czasową i może być stwierdzona przez prostą obserwację, o tyle równoczesność zdarzeń odległych nie jest czymś oczywistym. Einstein określił równoczesność zdarzeń odległych (rys. 1) następująco: jeśli w układzie inercjalnym spoczywa zegar i jeśli sygnał świetlny wysłany w chwili t_1 z miejsca spoczynku zegara i odbity od jakiejś przeszkody powraca w chwili t_2, to odbicie sygnału jest zdarzeniem równoczesnym ze wskazaniem przez zegar czasu $\frac{1}{2}(t_1+t_2)$. Z określenia tego wynika, że zdarzenia równoczesne w jednym układzie inercjalnym nie są na ogół równoczesne w innym układzie inercjalnym. Istotnie, niech prosta $x = 0$ na rys. 2 przedstawia zegar spoczywający, a prosta $x = vt$ — zegar poruszający się z prędkością v; jeśli nadto łamana ABC przedstawia bieg światła, to w punkcie B następuje odbicie. W układzie zegara spoczywającego zdarzenie B jest równoczesne ze zdarzeniem D określonym równością $AD = DC$, natomiast w układzie zegara poruszającego się zdarzenie B jest równoczesne ze zdarzeniem D' określonym równością: $AD' = D'C'$, gdzie zdarzenie C' jest odebraniem przez zegar poruszający się sygnału odbitego.

równoczesność zdarzeń

Względność równoczesności wynikająca z określenia Einsteina, tj. zależność równoczesności od układu inercjalnego, została odebrana przez współczesnych Einsteina jako coś całkowicie zmieniającego pojęcie czasu; wrażenie to zostało utrwalone w niefortunnej nazwie „teoria względności". Obecnie, z perspektywy przeszło siedemdziesięciu lat, uderza nas ciągłość przejścia od fizyki klasycznej do szczególnej teorii względności. W porównaniu z teoriami rzeczywiście i głęboko zmieniającymi fizyczny obraz świata – ogólną teorią względności i mechaniką kwantową — zmiany, które wprowadza do fizyki klasycznej szczególna teoria względności, są nieznacznymi uzupełnieniami tego samego w zasadzie obrazu rzeczywistości.

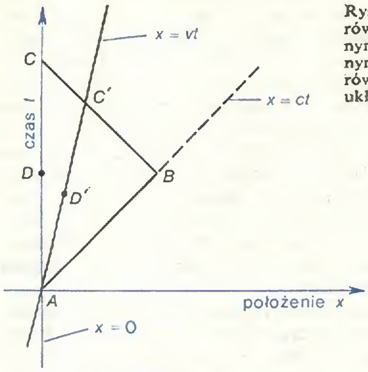
W trzy lata po ukazaniu się pracy Einsteina H. Min-



Rys. 1. Odbicie sygnału jest zdarzeniem równoczesnym ze wskazaniem czasu $\frac{1}{2}(t_1+t_2)$ przez zegar spoczywający w układzie inercjalnym

kowski podał prostą interpretację geometryczną zasad szczególnej teorii względności. Interpretacja Minkowskiego jest tak przekonywająco trafna, że współcześnie za szczególną teorię względności niektórzy uważają właśnie interpretację Minkowskiego.

Wróćmy do rys. 2; zdarzenia leżące na prostej $D'B$ są równoczesne w układzie zegara poruszającego



Rys. 2. Zdarzenia równoczesne w jednym układzie inercjalnym nie są na ogół równoczesne w innym układzie inercjalnym

się, podobnie jak zdarzenia na każdej prostej równoległej do $D'B$. Prosta równoległa do $D'B$ i przechodząca przez punkt A jest więc w układzie zegara poruszającego się odpowiednikiem osi x; nazwijmy ją osią x'. Podobnie prosta $x = vt$ jest odpowiednikiem osi t; nazwijmy ją osią t'. Łatwo stwierdzić, że zdarzenie o współrzędnych t, x ma w układzie poruszającym się współrzędne:

przekształcenie Lorentza

$$t' = \frac{t-(v/c^2)x}{\sqrt{1-v^2/c^2}}, \qquad x' = \frac{-vt+x}{\sqrt{1-v^2/c^2}},$$

gdzie c jest prędkością światła; związki te nazywamy przekształceniem Lorentza. Otóż:

$$(ct')^2-(x')^2 = (ct)^2-(x)^2.$$

czasoprzestrzeń Minkowskiego

Minkowski interpretuje formę kwadratową $(ct)^2-(x)^2$ jako kwadrat odległości zdarzenia o współrzędnych t, x od początku układu, przekształcenie Lorentza zaś jako odpowiednik obrotu. Interpretacja Minkowskiego polega więc na przypisaniu czasoprzestrzeni struktury metrycznej różnej od struktury metrycznej czasoprzestrzeni Galileusza. Strukturę tę można opisać następująco: z każdymi dwoma zdarzeniami związana jest liczba będąca czasoprzestrzennym odpowiednikiem kwadratu odległości między dwoma punktami w geometrii Euklidesowej. Dla dwu nieskończenie bliskich zdarzeń liczba ta ma postać:

$$ds^2 = c^2dt^2-dx^2-dy^2-dz^2,$$

gdzie t jest czasem mierzonym przez zegar spoczywający w wybranym układzie inercjalnym, c — prędkością światła, zaś współrzędne x, y, z mają to samo znaczenie, co w fizyce klasycznej (są to współrzędne kartezjańskie prostokątne na każdej płaszczyźnie równoczesności, $ct = \text{const}$); płaszczyzny równoczesności są zbiorami zdarzeń równoczesnych w wybranym układzie inercjalnym. O ile jednak płaszczyzny równoczesności zmieniają się wraz ze zmianą układu inercjalnego, o tyle liczba $(ds)^2$ dla dwu ustalonych zdarzeń ma tę samą wartość we wszystkich układach inercjalnych.

Treść fizyczną szczególnej teorii względności można ująć w następującą zasadę: wszystkie prawa fizyki dotyczące czasu i przestrzeni powinny być twierdzeniami geometrii metrycznej Minkowskiego.

Zbadamy dokładniej strukturę metryczną czasoprzestrzeni Minkowskiego. O współrzędnych t, x, y, z takich, że kwadrat odległości dwu bliskich zdarzeń ma postać:

$$c^2dt^2-dx^2-dy^2-dz^2$$

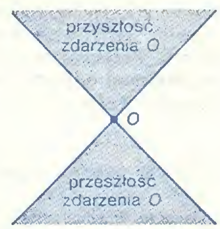
mówimy, że są dostosowane do struktury metrycznej czasoprzestrzeni; mówimy też, że tworzą układ inercjalny. Omawiając czasoprzestrzeń Galileusza zwróciliśmy uwagę na dwa różne znaczenia terminu „układ inercjalny". Pojęcie układu inercjalnego jako niewielkiego ciała, wolnego w dostatecznym stopniu od wpływu innych ciał i nie wykonującego ruchu obrotowego, przenosi się bez zmiany do szczególnej teorii względności; ulega natomiast zmianie pojęcie układu inercjalnego jako układu współrzędnych dostosowanych do struktury metrycznej czasoprzestrzeni, bo zmienia się właśnie sama struktura metryczna.

Obierzmy dowolne zdarzenie O jako początek układu współrzędnych czasoprzestrzennych. Kwadrat odległości zdarzenia X o współrzędnych t, x, y, z od początku układu:

$$s^2 = c^2t^2-x^2-y^2-z^2$$

może być dodatni, ujemny lub równy zeru. Mówimy, że zdarzenia O, X tworzą wektor czasowy — gdy $s^2 > 0$, przestrzenny — gdy $s^2 < 0$ i zerowy — gdy $s^2 = 0$. Zbiór wektorów zerowych o początku O tworzy stożek świetlny zdarzenia O. Stożek ten dzieli czasoprzestrzeń na trzy rozłączne podzbiory (rys. 3):

stożek świetlny



Rys. 3. Stożek świetlny zdarzenia O

wnętrze górnej części stożka zwane przyszłością zdarzenia O, wnętrze dolnej części stożka zwane przeszłością zdarzenia O i zewnętrze stożka. Stożki światła każdego zdarzenia stanowią absolutny element czasoprzestrzeni Minkowskiego, podobnie jak hiperpłaszczyzny równoczesności stanowią absolutny element czasoprzestrzeni Galileusza.

Stożek świetlny określa przeszłość i przyszłość jednego zdarzenia, a nie przeszłość i przyszłość w ogóle; tych ostatnich pojęć nie da się określić w czasoprzestrzeni Minkowskiego. K. Gödel, sławny matematyk, zwraca uwagę, że kryje się tu ciekawa filozoficznie różnica między fizyką klasyczną a szczególną teorią względności. Intuicja nasza zdaje się dostrzegać różnicę w sposobie istnienia rzeczy przeszłych i przyszłych. Ma to dobre oparcie w strukturze metrycznej czasoprzestrzeni Galileusza, która jest pocięta płaszczyznami równoczesności na rozłączne zbiory zdarzeń; o zdarzeniach leżących na jednej płaszczyźnie równoczesności możemy powiedzieć, że współistnieją, o innych — że były lub będą. W czasoprzestrzeni Minkowskiego nie da się wykonać żadnej takiej geometrycznie uzasadnionej konstrukcji. Gödel wyciąga stąd wniosek, że wszystkie zdarzenia istnieją tak samo, jedynie nasza intuicja, przestając spostrzegać pewne z nich, uznaje je za minione. Ciekawe, czy przekonanie, że świat „staje się", a nie „istnieje w czasie i przestrzeni", jest wrodzone umysłowi ludzkiemu, czy też ukształtowało się wyraźnie pod wpływem fizyki Newtona.

przeszłość i przyszłość

Rozważymy trzy stereotypowe poglądy łączone niekiedy ze szczególną teorią względności.

„Żaden sygnał w przyrodzie nie może mieć prędkości większej niż prędkość światła". Nie wykluczone, że zdanie to jest prawdziwe; nie jest ono jednak ani potrzebne dla uzasadnienia szczególnej teorii względności ani z niej nie wynika. Przeciwnie, w ramach szczególnej teorii względności można łatwo zbudować modele teoretyczne, w których występują sygnały o prędkości większej niż prędkość światła.

„Dwa zdarzenia tworzące wektor przestrzenny nie mogą być powiązane przyczynowo". Pojęcie więzi przyczynowej jest antropomorfizmem, któremu bar-

dzo trudno nadać znaczenie fizyczne. Jeżeli za więź przyczynową uznać istnienie zależności funkcyjnej, to pogląd ten nie jest na ogół prawdziwy. Twierdzenie Gaussa–Ostrogradskiego mówi np., że strumień pola elektrycznego przez zamkniętą powierzchnię jest proporcjonalny do ładunku elektrycznego wewnątrz powierzchni, a więc ustala zależność funkcyjną między zjawiskami oddzielonymi przestrzennie dowolnie dużą odległością.

„Ciała nie mogą oddziaływać na odległość, lecz za pośrednictwem pola, tj. materii istniejącej w przestrzeni między ciałami i mogącej istnieć także pod nieobecność samych ciał". Nie wykluczone, że twierdzenie to jest prawdziwe, niemniej nie wynika ono ze szczególnej teorii względności. W ramach tej ostatniej można z powodzeniem budować teorie działania na odległość na wzór newtonowskiej teorii grawitacji.

działanie z bliska i na odległość

Różnicę między działaniem z bliska, tj. przez określone środowisko, a działaniem na odległość, tj. przez „próżnię", rozumiano dobrze już w XVII wieku. Ch. Huygens i G.W. Leibniz odrzucali newtonowską teorię grawitacji jako opartą na wykluczonym a priori działaniu na odległość. Newton, zgadzając się w zasadzie z krytyką, rozumiał jednak, że dokładny opis zjawisk jest równie ważny jak znajomość ich mechanizmu. Przykład ten powinien być ostrzeżeniem przed pochopnym odrzucaniem działania na odległość jako sposobu opisu.

Ogólna teoria względności

„Ogólna teoria względności" jest nazwą równie niefortunną jak „szczególna teoria względności". Nazywamy w ten sposób współczesną teorię grawitacji odkrytą przez Einsteina w latach 1911–16. Newtonowskie prawo grawitacji zakłada strukturę metryczną czasoprzestrzeni Galileusza, a więc jest niezgodne ze strukturą metryczną czasoprzestrzeni Minkowskiego. Jeżeli tę ostatnią uważamy za bliższą rzeczywistości, to musimy zmienić prawo grawitacji. Klucz do nowej teorii grawitacji Einstein widział w równości masy bezwładnej i ciężkiej — fakcie przypadkowym w teorii Newtona. Z faktu tego wynika m.in., że masa niewielkiego ciała poruszającego się w polu grawitacyjnym nie występuje w równaniach Newtona, tj. że ruch wszystkich ciał w polu grawitacyjnym ma to samo przyspieszenie. Daleko idący wniosek Einsteina polega na przyjęciu, że skoro ruch wszystkich ciał jest taki sam, to ruch ten jest własnością czasoprzestrzeni, a nie samych ciał. Jest to hipoteza o strukturze metrycznej czasoprzestrzeni, którą można opisać następująco.

rzeczywista czasoprzestrzeń

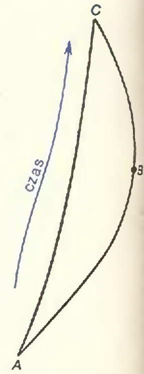
Rzeczywista czasoprzestrzeń ma się tak do czasoprzestrzeni Minkowskiego, jak zakrzywiona powierzchnia w trójwymiarowej przestrzeni Euklidesa do płaszczyzny: małe kawałki każdej gładkiej powierzchni można uważać za kawałki płaszczyzny; podobnie kawałki czasoprzestrzeni dostępne w doświadczeniu są małe w skali kosmosu i mogą być uważane za część czasoprzestrzeni Minkowskiego; większe części czasoprzestrzeni wykazują jednak własności metryczne różne od odpowiednio dużych części czasoprzestrzeni Minkowskiego.

Jak można znaleźć rzeczywisty kształt czasoprzestrzeni? Jest to zadanie technicznie trudne, lecz bardzo proste pojęciowo, niczym nie różniące się od zadania znalezienia np. kształtu powierzchni Ziemi. Należy jedynie wprowadzić określenia i hipotezy wiążące nieznaną geometrię ze zjawiskami dostępnymi obserwacji.

Kształt Ziemi można znaleźć umawiając się, że łańcuch napięty między dwoma bliskimi punktami powierzchni Ziemi tworzy odcinek prostej o długości równej liczbie ogniw. Mówimy tu, rzecz jasna, o pojęciowej zasadzie mierzenia, a nie o rzeczywistych pomiarach, które znacznie wygodniej robić przy pomocy teodolitu. Wykonując triangulację, tj. mierząc szereg stykających się bokami trójkątów, znajdziemy łatwo odstępstwa od hipotezy, że badana powierzchnia jest płaszczyzną.

triangulacja czasoprzestrzenna

Kształt czasoprzestrzeni można znaleźć umawiając się, że zegar spadający swobodnie w polu grawitacyjnym wskazuje upływ czasu równy długości linii czasoprzestrzennej, którą tworzy jego ruch. Niech linie *AB*, *AC* i *BC* na rys. 4 przedstawiają swobodne spadanie trzech jednakowych zegarów. W czasoprzestrzeni Minkowskiego, znając czas *AB*, czas *AC* i prędkość, z jaką zegary mijają się w punkcie *A*, można obliczyć czas *BC*, podobnie jak w geometrii Euklidesa oblicza się bok z dwu pozostałych boków i kąta między nimi. Jeżeli czas *BC* okaże się różny od obliczonego w ramach geometrii Minkowskiego, to będzie to znaczyć, że obszar czasoprzestrzeni zawierający trójkąt *ABC* nie może być uważany za część czasoprzestrzeni Minkowskiego. Przy pomocy takiej triangulacji czasoprzestrzennej można wyznaczyć dokładny kształt każdej części czasoprzestrzeni.



Rys. 4. Linie *AB*, *AC* i *BC* przedstawiają swobodne spadanie trzech jednakowych zegarów

Czy jest uzasadnione przyjąć, że swobodnie spadający zegar odmierza czas równy długości linii, którą tworzy jego ruch? Można by odpowiedzieć, że jest to definicja długości krzywej czasoprzestrzennej, którą przyjmujemy nie tłumacząc jej racjonalności; jest to jednak odpowiedź niesłuszna. Prawidłowa odpowiedź brzmi: intuicja oparta na wiedzy fizycznej i matematycznej każe nam przypuszczać, że czas mierzony przez swobodnie spadający zegar jest istotnie bardzo dobrą miarą długości krzywej czasoprzestrzennej. Samo pojęcie długości jest w ramach ogólnej teorii względności pojęciem podstawowym; jest nam ono znane przed każdym rzeczywistym pomiarem. Wykonując rzeczywiste pomiary nie tylko nie uzależniamy mierzonych długości od ich roboczych definicji, lecz przeciwnie — przeprowadzamy analizę i krytykę wyników opierając się na znajomości ogólnej teorii względności i innych działów fizyki jako hipotez podstawowych.

równanie Einsteina

Krzywizna czasoprzestrzeni nie jest dowolna, lecz ograniczona równaniami Einsteina. Równania te dopuszczają szeroki zbiór rozwiązań, a każde rozwiązanie jest a priori dopuszczalną formą czasoprzestrzeni. Problemy, które nastręcza interpretacja rozwiązań równań Einsteina zobrazujemy na dwu przykładach.

Niech współrzędne t, x, y, z tworzą układ inercjalny w czasoprzestrzeni Minkowskiego. Utożsamiając punkty hiperpłaszczyzny $z = 0$ z odpowiednimi punktami hiperpłaszczyzny $z = a > 0$ otrzymamy czasoprzestrzeń różną od czasoprzestrzeni Minkowskiego, lecz spełniającą równania Einsteina tak jak sama czasoprzestrzeń Minkowskiego. Można też utożsamić punkty hiperpłaszczyzny $t = 0$ z odpowiednimi punktami hiperpłaszczyzny $t = b > 0$; otrzymamy czasoprzestrzeń o cyklicznym czasie, tj. czasoprzestrzeń, w której wszystkie zdarzenia okresowo się powtarzają. Istnienie zamkniętych linii czasowych tak dalece komplikuje i tak niełatwą analizę rozwiązań równań Einsteina, że wielu fizyków zajmujących się tymi zagadnieniami jest skłonnych uznać, iż rozwiązania z zamkniętymi liniami czasowymi są a priori niedopuszczalne.

Mechanika kwantowa

Mechanika kwantowa powstała w latach 1925–2 w wyniku długotrwałych prób znalezienia opisu zjawisk atomowych, np. liniowego widma emisyjnego gazów. Próby te utrwaliły pogląd, że opis taki jest niemożliwy w ramach tradycyjnych pojęć. Na przykład teoria Bohra atomu wodoru pozwala w zasadzie prawidłowo opisać widmo atomu wodoru zakładając, że w atomie tym elektron krąży wokół protonu tak jak planeta wokół Słońca, lecz po niektórych tylko orbitach Keplera, co jest niezrozumiałą innowacją.

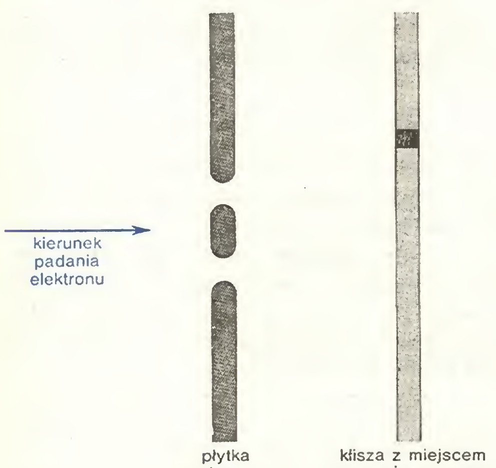
Tytuł pierwszej z prac E. Schrödingera: *Kwantowanie jako problem wartości własnych*, dobrze oddaje kierunek myśli autora. Od dawna wiedziano, że równania różniczkowe opisujące ruch drgający, np. równanie struny, mają przy pewnych warunkach brzegowych rozwiązania tylko dla niektórych częstości, zw. częstościami własnymi. Skoro atom wodoru wysyła promieniowanie o ściśle określonych częstościach, to widocznie atom ten wykonuje ruch drgający, a celem fizyki jest znalezienie równania opisującego ten ruch.

równanie Schrödingera

Schrödinger podał równanie, zwane od tego czasu równaniem Schrödingera, które pozwala odtworzyć teorię Bohra atomu wodoru w sposób matematycznie zrozumiały, tj. znaleźć charakterystyczne dla tego atomu częstości mniej więcej tak, jak znajduje się częstości drgającej struny. Równanie Schrödingera dla atomu wodoru jest równaniem różniczkowym funkcji falowej elektronu $\psi(t, x, y, z)$, gdzie współrzędne t, x, y, z tworzą układ inercjalny w czasoprzestrzeni Galileusza. Równanie Schrödingera dla atomu helu, który zawiera dwa elektrony, jest równaniem określającym funkcję falową dwu elektronów $\psi(t, x_1, y_1, z_1, x_2, y_2, z_2)$ itd. Jądra atomowe, które są znacznie cięższe od elektronów, uważamy za nieruchome centra sił, można jednak zbudować równanie Shrödingera uważając jądro za ruchomą część atomu. Jakie jest znaczenie fizyczne funkcji falowej? Odpowiedź na to pytanie dał M. Born; ażeby jednak właściwie ocenić charakter tej odpowiedzi, rozpatrzymy proste doświadczenie, które w różnych wersjach było wielokrotnie wykonywane.

interpretacja funkcji falowej

Elektron pada na płytkę z dwoma małymi otworkami, przez które może on przejść na drugą stronę (rys. 5); za otworkami jest klisza fotograficzna, która rejestruje każdy przechodzący elektron. Doświadczenie daje następujące wyniki: miejsca zaczernione przez elektron w kolejnych doświadczeniach są różne, a różnicy nie da się powiązać z jakąkolwiek niejednolitością stanów elektronu poprzedzających przejście przez otworki; przy wielokrotnym powtarzaniu doświadczenia liczba śladów na kliszy układa się w obraz o kilku maksimach i minimach, jak przy interferencji światła na dwu otworkach.



Rys. 5. Przejście pojedynczego elektronu przez płytkę z dwoma otworami

Wydaje się, że przed teorią można by postawić dwa zadania: przewidzieć, w którym miejscu na kliszy powstanie zaczernienie w pojedynczym doświadczeniu; wyjaśnić, skąd biorą się maksima i minima przy wielokrotnym powtarzaniu doświadczenia.

Interpretacja Borna funkcji falowej elektronu daje wyczerpującą odpowiedź na drugie pytanie, lecz jednocześnie wyklucza danie odpowiedzi na pierwsze pytanie. Mianowicie liczba $|\psi(t, x, y, z)|^2\, dxdydz$ jest wg Borna proporcjonalna do prawdopodobieństwa znalezienia elektronu w objętości $dxdydz$ w chwili t, jeżeli poszukiwanie takie zostanie przedsięwzięte. Postępowanie teoretyczne, które wraz z interpretacją Borna doprowadziło do odkrycia ogromnej liczby nowych faktów, zobrazujemy na przykładzie opisanego wyżej doświadczenia. Zakładamy, że ruch elektronu jest opisany — cokolwiek by to znaczyło — funkcją falową $\psi(t, x, y, z)$. Równanie Schrödingera dla tej funkcji rozwiązujemy przy warunkach brzegowych odpowiadających założonej nieprzepuszczalności płytki, np. przyjmując, że funkcja falowa znika na brzegu płytki. Jest to postępowanie od dawna znane w fizyce teoretycznej, identyczne w istocie z postępowaniem zastosowanym przez A. J. Fresnela w początkach XIX w. w celu objaśnienia ugięcia światła. Wreszcie przyjmujemy, że proces oddziaływania elektronu z kliszą można traktować jako lokalizację elektronu w niewielkiej objętości i stosujemy interpretację Borna, by przewidzieć natężenie lokalizacji przy wielokrotnym powtarzaniu doświadczenia.

Mechanika kwantowa a szczególna teoria względności

interpretacja mechaniki kwantowej

Bohr i Heisenberg rozwinęli i zaostrzyli interpretację probabilistyczną Borna do postaci zwanej kopenhaską interpretacją mechaniki kwantowej. Wolimy nie wchodzić w szczegóły tej interpretacji, wydaje się nam bowiem, że interpretacja kopenhaska jest przykładem i wynikiem filozoficznej emfazy, której w fizyce lepiej unikać, o ile tylko jest to możliwe. Nie zaostrzona filozoficznie interpretacja Borna wystarcza w zupełności do praktycznego stosowania mechaniki kwantowej.

Einstein — będący jednym z twórców mechaniki kwantowej — nigdy nie przyjął interpretacji kopenhaskiej, ponieważ przypuszczał, że interpretacja kopenhaska może prowadzić do istnienia sygnałów o prędkości większej niż prędkość światła. Jak wyjaśniliśmy wcześniej, istnienie takich sygnałów nie uważamy dziś za coś a priori niemożliwego lub sprzecznego z teorią względności; ta ostatnia jest tylko hipotezą o strukturze metrycznej czasoprzestrzeni i niewiele mówi o własnościach materii wypełniającej czasoprzestrzeń.

Nie wypowiadając się w sprawach niejasnych, chcielibyśmy podkreślić jednak to, co jest jasne (przynajmniej dla nas).

Spekulacje filozoficzne związane z mechaniką kwantową są wątpliwej wartości przede wszystkim ze względu na prowizoryczny charakter samej mechaniki kwantowej. Mechanikę kwantową należy bowiem uzgodnić ze szczególną teorią względności. Jedyną współczesną teorią, która jest syntezą mechaniki kwantowej i szczególnej teorii względności jest tzw. elektrodynamika kwantowa. Są trzy (co najmniej) powody, by uważać elektrodynamikę kwantową za teorię wymagającą dopracowania. Po pierwsze, w rachunkach związanych z tą teorią pojawiają się wyrażenia matematycznie nieokreślone (całki rozbieżne); istnieje wprawdzie sposób jednoznacznego obliczania tych wyrażeń, lecz sam fakt jest sygnałem trudności pojęciowych teorii. Po drugie, zasady mechaniki kwantowej w zastosowaniu do pola elektromagnetycznego prowadzą do sprzeczności z równaniami Maxwella, które są równaniami ruchu tego pola. Większość badaczy nie przywiązuje wagi do tego, lecz np. G. Källén, jeden z najlepszych znawców elektrodynamiki kwantowej, napisał wyraźnie: „jest to w najlepszym wypadku brak elegancji w teorii". Po trzecie, w elektrodynamice kwantowej występuje stała bezwymiarowa $\hbar c/e^2 = 137{,}036\ldots$, gdzie $\hbar$ jest stałą Plancka, c — prędkością światła, a e — ładunkiem elektronu. Stała ta wg jednomyślnej opinii wielu fizyków powinna być obliczona teoretycznie,

braki elektrodynamiki kwantowej

lecz zdaje się to być niemożliwe w ramach elektrodynamiki kwantowej.

Jasne jest także, że w ostatnich kilkunastu latach zmieniła się wyraźnie atmosfera intelektualna wokół zagadnienia podstaw mechaniki kwantowej. Atmosferę z końca lat trzydziestych opisuje L. Infeld, któremu Einstein powiedział „tu w Princeton uważają mnie za starego durnia"; chodziło oczywiście o opozycję Einsteina wobec interpretacji kopenhaskiej. Dziś nikt zapewne nie pomyślałby czegoś takiego.

E. J. Dijksterhuis *The Mechanization of the World Picture*, Oxford 1969; L. Infeld *Szkice z przeszłości*, Warszawa 1966; A. Trautman *Teoria względności*, Wrocław 1971; E.T. Whittaker *A History of the Theories of Aether and Electricity*, New York 1960.

Zasady zachowania

Wojciech Kopczyński

Zasady zachowania dotyczą tych wielkości fizycznych, o których można powiedzieć, że nie dają się zniszczyć, tzn., które podczas procesów jakim podlegają układy fizyczne izolowane od otoczenia (zwane też układami odosobnionymi) pozostają niezmienione. Zachowanie wielkości fizycznych może przy tym być słuszne bezwzględnie, lub też obowiązywać jedynie dla niektórych procesów. Zasady zachowania określają warunki, których spełnienie zapewnia stałość tych wielkości fizycznych. Ich przegląd rozpocznijmy od zasad poznanych najwcześniej.

Zasady zachowania w mechanice Newtona

Pęd pojedynczej cząstki określamy jako iloczyn jej masy przez prędkość

$$\vec{p} = m\vec{v}.$$

Gdy na cząstkę nie działa żadna siła, to z I zasady dynamiki Newtona wynika, że jej pęd jest zachowany.

zasada zachowania pędu

Pęd układu fizycznego złożonego z n cząstek jest sumą pędów poszczególnych cząstek. Z II zasady Newtona wynika, że jest on stały, gdy suma sił działających na cząstki jest równa zeru:

$$\vec{p}_1+\vec{p}_2+ \dots +\vec{p}_n = \overrightarrow{\text{const}}, \text{ gdy } \vec{F}_1+\vec{F}_2+ \dots +\vec{F}_n = \vec{0}.$$

Załóżmy, że wszystkie siły działające w tym układzie są siłami wewnętrznymi, tzn. $\vec{F}_1 = \vec{F}_{12}+\vec{F}_{13} \dots + +\vec{F}_{1n}$, $\vec{F}_2 = \vec{F}_{21}+\vec{F}_{23}+ \dots +\vec{F}_{2n}$ itd., przy czym $\vec{F}_{ij}$ oznacza siłę działającą na i-tą cząstkę a pochodzącą od cząstki j-tej. Przyjmijmy też, że siły te spełniają III zasadę Newtona, $\vec{F}_{ij} = -\vec{F}_{ji}$. Wtedy suma wszystkich sił działających na cząstki będzie równa zeru, gdyż siła $\vec{F}_{12}$ odejmie się od siły $\vec{F}_{21}$, $\vec{F}_{13}$ od $\vec{F}_{31}$ itd. Zatem, całkowity pęd układu cząstek będzie zachowany.

Z zasady zachowania pędu oraz z zasady zachowania masy wynika, że środek masy układu porusza się ruchem jednostajnym i prostoliniowym.

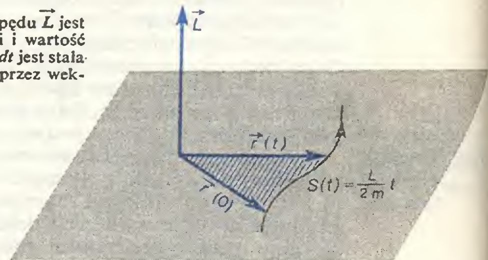
Moment pędu $\vec{L}$ cząstki względem pewnego punktu O określa się jako iloczyn wektorowy promienia wodzącego cząstki $\vec{r}$, poprowadzonego z O, przez pęd tej cząstki $\vec{p}$, $\vec{L} = \vec{r}\times\vec{p}$. Moment pędu spełnia równanie

$$\frac{d\vec{L}}{dt} = \vec{M},$$

gdzie $\vec{M} = \vec{r}\times\vec{F}$ nazywa się momentem siły $\vec{F}$ względem punktu O. Siła $\vec{F}$ jest siłą centralną, gdy istnieje punkt O taki, że siła ta jest w każdym punkcie równoległa do promienia wodzącego $\vec{r}$ poprowadzonego z tego punktu O. Nazywa się on centrum siły $\vec{F}$. Moment siły centralnej $\vec{F}$ względem centrum znika, a zatem moment pędu względem tego punktu jest stały:

$$\vec{L} = \overrightarrow{\text{const}}, \quad \text{gdy } \vec{F} \parallel \vec{r}.$$

Przedstawiona powyżej zasada zachowania momentu pędu dla pojedynczego punktu materialnego niesie dwie informacje. Po pierwsze, wektory $\vec{r}$ i $\vec{v}$, spełniające związek $\vec{r}\times\vec{v} = \vec{L}/m$, muszą stale pozostawać w płaszczyźnie przechodzącej przez centrum i prostopadłej do wektora $\vec{L}$, a zatem ruch cząstki powinien się odbywać w tej płaszczyźnie. Po drugie, prędkość polowa cząstki, wynosząca $\vec{L}/2m$, pozostaje stała (rys. 1). II prawo Keplera oraz informacja, że tor planety jest krzywą płaską oznaczają właśnie, że moment pędu planety w ruchu wokół Słońca spełnia zasadę zachowania. Zasada ta jest słuszna w przybliżeniu, gdy pomijamy wpływ innych ciał niebieskich na ruch planety. W lepszym przybliżeniu zachowany jest moment pędu całego Układu Słonecznego.



Rys. 1. Gdy moment pędu $\vec{L}$ jest stały, ruch jest płaski i wartość prędkości polowej dS/dt jest stała (S — pole zakreślone przez wektor wodzący cząstki)

Gdy mamy układ złożony z dwóch cząstek, oddziaływających między sobą siłami spełniającymi III zasadę Newtona, suma momentów sił działających na te cząstki wynosi

$$\vec{M} = \vec{r}_1\times\vec{F}_1+\vec{r}_2\times\vec{F}_2 = (\vec{r}_1-\vec{r}_2)\times\vec{F}_1.$$

zasada zachowania momentu pędu

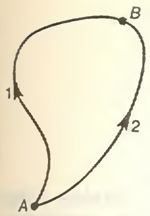
Uogólniając definicję podaną dla pojedynczej cząstki, siłę $\vec{F}_1$ nazwiemy centralną, jeśli jest ona równoległa do prostej łączącej cząstki. Jeśli $\vec{F}_1$ jest centralna, to $\vec{M} = 0$, a więc całkowity moment pędu układu, określony jako suma momentów pędu cząstek, jest zachowany, $\vec{L} = \vec{L}_1+\vec{L}_2 = \overrightarrow{\text{const}}$. Nietrudno uogólnić zasadę zachowania momentu pędu na sytuację, gdy układ składa się z n cząstek:

$$\vec{L} = \vec{L}_1+\vec{L}_2+ \dots +\vec{L}_n = \overrightarrow{\text{const}},$$
$$\text{gdy } \vec{M}_1+\vec{M}_2+ \dots +\vec{M}_n = \vec{0}.$$

Podkreślamy, że ostatnia równość zachodzi wtedy, gdy siły pomiędzy każdą parą cząstek spełniają zasadę równej akcji i reakcji oraz skierowane są wzdłuż prostej łączącej cząstki danej pary.

Rozważmy pojedynczą cząstkę poruszającą się w niezależnym od czasu polu sił $\vec{F}(\vec{r})$. Jednym z wniosków z II zasady Newtona jest stwierdzenie, że przyrost energii kinetycznej cząstki jest równy wykonanej nad nią pracy, tzn.

$$\frac{1}{2}m\vec{v}_B^2-\frac{1}{2}m\vec{v}_A^2 = \int_A^B \vec{F}d\vec{r}.$$



Rys. 2. Gdy pole sił jest potencjalne, to praca wzdłuż drogi *1* równa jest pracy wzdłuż drogi *2*

Pole sił nazywamy potencjalnym, jeżeli praca wykonana nad cząstką przy jej przesunięciu z punktu A do punktu B nie zależy od wyboru toru łączącego te punkty (rys. 2):

$$\int_{\text{wzdłuż } 1} \vec{F} d\vec{r} = \int_{\text{wzdłuż } 2} \vec{F} d\vec{r}.$$

Wtedy możemy wprowadzić pojęcie potencjału V. Potencjał $V(\vec{r}_A)$ pola sił $\vec{F}(\vec{r})$ w punkcie $\vec{r}_A$ jest to praca wykonana przy przesunięciu cząstki od tego punktu do pewnego dowolnie wybranego punktu $\vec{r}_0$, czyli

$$V(\vec{r}_A) = \int_A^0 \vec{F} d\vec{r}.$$

Definicja ta ma sens, dzięki założeniu o potencjalności pola $\vec{F}(\vec{r})$. Potencjał V jest wyznaczony niejednoznacznie; wybierając inny punkt $\vec{r}_0$, dodajemy do potencjału stałą. Stwierdziwszy następnie, że praca wykonana przy przesunięciu cząstki z punktu A do B równa się różnicy potencjałów między tymi punktami,

$$\int_A^B \vec{F} d\vec{r} = V(\vec{r}_A) - V(\vec{r}_B),$$

zasada zachowania energii

dochodzimy do wniosku, że suma energii kinetycznej i potencjalnej cząstki nie zależy od punktu, w którym znajduje się cząstka,

$$\mathcal{E} = \frac{1}{2} m\vec{v}_A^2 + V(\vec{r}_A) = \frac{1}{2} m\vec{v}_B^2 + V(\vec{r}_B),$$

a więc pozostaje stała w czasie.

Ze sformułowanej powyżej zasady zachowania energii dla jednego punktu materialnego wynika, że prędkość punktu materialnego o ustalonej energii jest funkcją jego położenia.

Pojęcie potencjału uogólnia się na dwie lub więcej cząstek. Jest on wówczas funkcją położeń tych cząstek: $\vec{r}_1$, $\vec{r}_2$, itd. Na przykład, potencjał dwóch ładunków Q_1, Q_2 oddziaływających elektrostatycznie zależy od ich wzajemnej odległości $r = |\vec{r}_1 - \vec{r}_2|$ w następujący sposób: $V = -Q_1Q_2/4\pi\varepsilon_0 r$, gdzie ε_0 jest przenikalnością elektryczną próżni. Prawo zachowania energii dla n cząstek wyraża się formułą

$$\mathcal{E} = \sum_{i=1}^{n} \frac{1}{2} m_i \vec{v}_i^2 + V(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_n) = \text{const.}$$

Istnieje wiele sił nie mających potencjału. Zalicza się do nich wszystkie siły zależne od prędkości cząstek. Jednakże nie wszystkie siły tego typu powodują niezachowanie energii. Przykład: siła Lorentza $\vec{F} = Q\vec{v} \times \vec{B}$ działająca na ładunek Q poruszający się w polu indukcji magnetycznej $\vec{B}$, jest prostopadła do kierunku ruchu cząstki, nie wykonuje więc żadnej pracy, a co za tym idzie, nie wpływa na zmianę energii cząstek.

Występują jednak siły niepotencjalne prowadzące do rozpraszania energii mechanicznej. Są to przede wszystkim siły tarcia i oporu ośrodka. Siły te są pochodzenia międzycząsteczkowego, a więc mają naturę elektromagnetyczną. A zatem, w zjawiskach, w których występują te siły, zasada zachowania energii obowiązuje na poziomie mikroskopowym. Opór ośrodka powoduje zamianę energii kinetycznej ciała jako całości na energię kinetyczną cząsteczek lub na inne postacie energii wewnętrznej. Przekonanie o niezniszczalności energii prowadzi do sformułowania I zasady termodynamiki. Zawiera ona najogólniej wyrażoną zasadę zachowania energii: energia układu odosobnionego jest zachowana. Termodynamika pozwala przy tym teoriom szczegółowym na

I zasada termodynamiki

dokładne określenie tego czym jest energia różnorakich typów układów fizycznych.

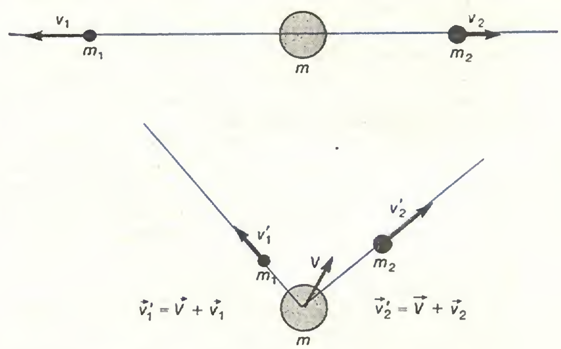
Zasady zachowania są często pomocne w rozwiązywaniu równań ruchu. Pozwalają one zmniejszyć liczbę równań, a także zastąpić niektóre równania różniczkowe drugiego rzędu równaniami pierwszego rzędu. Oto dwa przykłady, w których ta rola zasad zachowania jest szczególnie wyraźna. Zasada zachowania pędu pozwala zmniejszyć liczbę równań ruchu o trzy — dzięki temu rozwiązywanie równań ruchu 2 ciał sprowadza się do badania ich ruchu względnego, a więc do rozwiązywania równań analogicznych do równań Newtona dla 1 ciała. Rozwiązując zagadnienie ruchu cząstki po zadanej krzywej, możemy zastąpić równanie Newtona, będące równaniem różniczkowym drugiego rzędu, równaniem pierwszego rzędu, wyrażającym zasadę zachowania energii.

rola zasad zachowania w mechanice

Zasady zachowania mają szczególne znaczenie, gdy nie znamy sił działających między cząstkami, a także wtedy, gdy siły działają przez krótki okres, a zainteresowani jesteśmy zachowaniem się układu przed i po tym okresie; tak jest np. podczas zderzeń i rozpadów cząstek. Teoria tych zjawisk opiera się nie na równaniach ruchu, lecz wyłącznie na prawach zachowania. Dobitnym przykładem takiego zastosowania zasad zachowania jest zasada działania silnika odrzutowego, gdy nie wnikając w charakter oddziaływania pomiędzy rozpalonymi gazami a ściankami komory spalania, a posługując się jedynie zasadą zachowania pędu, potrafimy zrozumieć zjawisko odrzutu.

Energia, pęd i masa w fizyce newtonowskiej i w szczególnej teorii względności

W fizyce nierelatywistycznej, zasada zachowania masy jest konsekwencją zasady zachowania pędu. Aby uzasadnić to stwierdzenie, rozpatrzmy rozpad spo-



Rys. 3. Rozpad cząstki o masie m wg mechaniki Newtona: a) w układzie, w którym cząstka ta spoczywa, b) w układzie poruszającym się względem poprzedniego z prędkością $\vec{V}$

zasada zachowania masy nierelatywistycznej

czywającej cząstki o masie m na dwie cząstki o masach m_1 i m_2 (rys. 3). Zgodnie z zasadą zachowania pędu

$$\vec{0} = m_1\vec{v}_1 + m_2\vec{v}_2.$$

Rozważmy następnie ten sam proces w układzie odniesienia poruszającym się ze stałą prędkością $-\vec{V}$ względem układu pierwotnego. Zasada zachowania pędu w tym układzie ma postać

$$m\vec{V} = m_1(\vec{v}_1 + \vec{V}) + m_2(\vec{v}_2 + \vec{V}).$$

Z powyższych równań wynika zasada zachowania masy

$$m = m_1 + m_2.$$

W fizyce nierelatywistycznej zasady zachowania masy i pędu są nierozłącznie związane. W fizyce relatywistycznej podobny związek, oparty na einsteino-

wskim prawie dodawania prędkości, obowiązuje dla energii i pędu. Energię liczoną w jednostkach masy nazywamy masą relatywistyczną,

masa relatywistyczna

$$m_{rel} = \frac{\mathcal{E}}{c^2}.$$

Wielkość ta spełnia tę samą funkcję, co masa nierelatywistyczna w równaniu wiążącym pęd cząstki z jej masą

$$\vec{p} = \frac{\mathcal{E}}{c^2}\vec{v}.$$

Masa relatywistyczna, w odróżnieniu od masy nierelatywistycznej, zależy od wyboru układu odniesienia. Z tego powodu bliższa masie nierelatywistycznej jest masa spoczynkowa

$$m = \sqrt{\left(\frac{\mathcal{E}}{c^2}\right)^2 - \left(\frac{\vec{p}}{c}\right)^2}$$

Z zasad zachowania energii i pędu wynika, że masa spoczynkowa rozpadającej się cząstki jest większa niż suma mas spoczynkowych cząstek, na które się ona rozpada, $m > m_1 + m_2$.

W fizyce relatywistycznej mamy o jedną zasadę zachowania (masy) mniej niż w fizyce newtonowskiej. Zasady zachowania energii i pędu są w niej nierozłączne, toteż często nazywa się je zasadami zachowania energii-pędu.

Zasady zachowania w klasycznej teorii pola

III zasada Newtona, $\vec{F}_{ij} = -\vec{F}_{ji}$, jest ściśle związana z zasadami zachowania pędu i momentu pędu w nierelatywistycznej mechanice. Zasada ta wymaga, aby zmiany, którym podlegają siły $\vec{F}_{ij}$ i $\vec{F}_{ji}$, działające na oddalone od siebie cząstki, zachodziły równocześnie. Odrzucenie przez teorię względności koncepcji równoczesności niezależnej od układu odniesienia, pociąga za sobą rezygnację z III zasady Newtona. W związku z tym, suma np. pędów cząstek nie jest na ogół wielkością zachowaną.

III zasada Newtona a teoria względności

W polowym obrazie świata cząstki nie oddziałują na siebie bezpośrednio — na odległość. Oddziaływanie między cząstkami następuje za pośrednictwem pól fizycznych. Typowym polem fizycznym jest pole elektromagnetyczne opisane wektorami natężenia pola elektrycznego $\vec{E}$ i indukcji magnetycznej $\vec{B}$, będącymi funkcjami czasu i położenia.

Podobnie jak cząstkom, polom fizycznym można przypisać energię, pęd i moment pędu. Na przykład, energia pola elektromagnetycznego w próżni przypadająca na jednostkę objętości wynosi:

$$\frac{\mathcal{E}}{V} = \frac{1}{2}\varepsilon_0\vec{E}^2 + \frac{1}{2\mu_0}\vec{B}^2,$$

gdzie μ_0 oznacza przenikalność magnetyczną próżni, natomiast jego pęd na jednostkę objętości jest opisany wzorem

$$\frac{\vec{p}}{V} = \varepsilon_0\vec{E}\times\vec{B}.$$

Moment pędu przypadający na jednostkę objętości określamy tak jak dla cząstki. Dla układu izolowanego, złożonego z pola i cząstek, energia całkowita, będąca sumą energii cząstek i energii pola jest wielkością zachowaną. Podobnie się dzieje dla pędu i momentu pędu.

W elektrodynamice prawa zachowania przybierają formę lokalnych zasad zachowania. Lokalna zasada zachowania np. energii podaje, że jeśli pewna porcja energii pola znika z danego obszaru, to dzieje się tak dlatego, że wypływa ona przez granice tego obszaru. Z lokalnych zasad zachowania wynikają zasady zachowania dla układów izolowanych — globalne zasady zachowania. Wydaje się, że lokalne sformułowanie zasad zachowania jest wręcz koniecznością w teorii relatywistycznej. Energia nie może znikać w pewnym obszarze i, zamiast wypływać przez jego granice, pojawiać się jednocześnie w innym miejscu, gdyż pojęcie jednoczesności zależy w teorii względności od układu odniesienia. Gdyby owa jednoczesność zachodziła w pewnym układzie odniesienia, to w układzie odniesienia poruszającym się względem niego energia pojawiałaby się w innej chwili, wcześniej lub później. A więc w tym innym układzie energia całkowita byłaby przez pewien czas inna niż na początku i na końcu tego energetycznego hokus-pokus, co jest nonsensem.

lokalne zasady zachowania

Dość niespodziewanie, teorią, w której nie można sformułować lokalnych zasad zachowania jest einsteinowska teoria grawitacji, jakkolwiek globalne zasady zachowania w tej teorii obowiązują. Nie znamy zadowalających wyrażeń na gęstość energii i pędu pola grawitacyjnego, jakie istnieją np. dla pola elektromagnetycznego.

Zasady zachowania a symetria praw fizyki

Zasady zachowania wiążą się z symetriami teorii fizycznych. Związek ten dla teorii niekwantowych jest opisany w twierdzeniu E. Noether.

Symetrią teorii fizycznej nazywamy to przekształcenie podstawowych dla niej wielkości fizycznych (jak położenia cząstek, pola fizyczne itp.), które przeprowadzą rozwiązania równań tej teorii w inne ich rozwiązania. Uznaje się także, że teoria jest niezmiennicza względem przekształceń będących symetriami. Rozważmy jednoparametrową rodzinę symetrii (S_α). Rodzinę tę nazywamy jednoparametrową grupą symetrii, gdy złożenie dwóch symetrii z tej rodziny S_α i S_β jest symetrią należącą do tej rodziny odpowiadającą parametrowi $\alpha+\beta$:

$$S_\alpha \cdot S_\beta = S_{\alpha+\beta}.$$

Na przykład, złożenie obrotów wokół ustalonej osi o kąty α i β jest obrotem o kąt $\alpha+\beta$.

Twierdzenie Noether głosi, że każdej jednoparametrowej grupie symetrii teorii fizycznej odpowiada zasada zachowania. Dotyczy to m.in. zasad zachowania energii, pędu i momentu pędu. Jeśli przesunięcia w czasie są symetriami danej teorii, to w jej ramach obowiązuje zasada zachowania energii. Jeśli teoria jest niezmiennicza względem przesunięć w przestrzeni w pewnym kierunku, to składowa pędu w tym kierunku jest zachowana; zasada zachowania wektora pędu obowiązuje w danej teorii, jeśli przesunięcia w przestrzeni we wszystkich kierunkach są jej symetriami. Obroty wokół pewnej osi są związane przez twierdzenie Noether ze składową momentu pędu skierowaną wzdłuż tej osi; jeśli obroty wokół pewnego punktu są symetriami teorii, to wektor momentu pędu względem tego punktu jest w niej zachowany.

twierdzenie Noether

Spróbujmy zbadać sens niezmienniczości teorii względem przesunięć w czasie, rozważając ruch gwiazd o masach m_1 i m_2 (znajdujących się w odległości r) oddziałujących na siebie siłą grawitacyjną

$$F = G\frac{m_1 m_2}{r^2}.$$

Wyniki rozważań dotyczących ruchu tych gwiazd nie będą zależeć od tego, czy ruch odbywał się 10 mln lat temu, odbywa się obecnie, czy też będzie się odbywał za milion lat. Wynika to z uniwersalnego charakteru newtonowskiego prawa grawitacji. Przesunięcia w czasie są więc symetriami newtonowskiej teorii dwóch ciał niebieskich.

To, że niemal wszystkie poważniejsze teorie fizyczne są niezmiennicze względem przesunięć w czasie

i przestrzeni oraz obrotów, jest spowodowane z jednej strony wiarą, iż przekształcenia te są symetriami realnego świata, z drugiej zaś tym, że nie znamy odstępstw od praw zachowania energii, pędu i momentu pędu. Nie jest natomiast prawdą, że byłoby niemożliwe zbudowanie teorii, która nie byłaby niezmiennicza względem wymienionych przekształceń. Przypuśćmy, że astronomowie śledzący ruch planet doszli do wniosku, iż stała grawitacyjna G wolno zmienia się w czasie. Wówczas przesunięcia w czasie nie byłyby symetriami mechaniki ciał niebieskich, a więc energia mechaniczna układu gwiazd nie zostałaby zachowana.. Nikt jednak nie mówiłby wtedy o odkryciu astronomicznego perpetuum mobile. Fizycy staraliby się zbudować teorię, w której G byłaby traktowana nie jako stała, lecz jako pewne pole. Niezmienniczość tej teorii względem przesunięć w czasie zapewniałaby zachowanie całkowitej energii układu, będącej sumą energii cząstek i pola grawitacyjnego, w tym pola G. Warto nadmienić, że istnieje teoria pola grawitacyjnego tego typu, nie mająca wszakże potwierdzenia w obserwacjach astronomicznych.

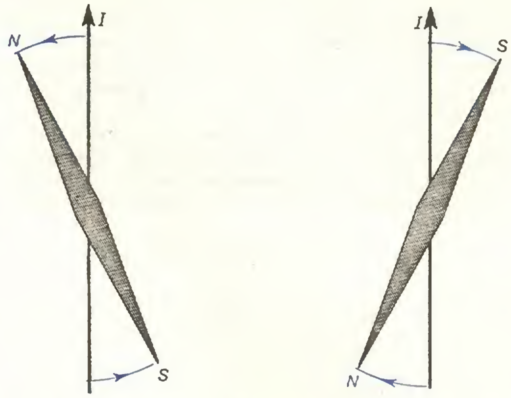
grupa Galileusza i grupa Poincarégo

Przesunięcia w czasie i przestrzeni oraz obroty wchodzą w skład zbioru wszystkich symetrii czasoprzestrzeni, który w fizyce nierelątywistycznej nazywa się grupą Galileusza, a w fizyce relatywistycznej grupą Poincarégo. Do grup tych zalicza się też transformacje mieszające współrzędne czasowe i przestrzenne, noszące odpowiednio nazwy specjalnych transformacji Galileusza, albo specjalnych transformacji Lorentza. Niezmienniczość względem tych przekształceń oznacza, że prawa fizyki są takie same dla dwóch obserwatorów, z których jeden porusza się względem drugiego ruchem jednostajnym i prostoliniowym. Symetriom tym też odpowiadają pewne zasady zachowania, są one jednak konsekwencją zasad rozważanych już poprzednio. Na przykład, w fizyce nierelatywistycznej wynikająca stąd zasada głosi, że prędkość środka masy układu jest zachowana, co jak wiemy, jest wnioskiem wynikającym z zasady zachowania pędu.

Symetriami czasoprzestrzeni są też odbicia w czasie i inwersje — inaczej odbicia względem wyróżnionego punktu. Niezmienniczość teorii względem inwersji i obrotów łącznie jest równoważna niezmienniczości względem odbić zwierciadlanych. Wynika to z tego, że inwersję i obrót można otrzymać jako złożenie odpowiednio trzech lub dwóch odbić zwierciadlanych, natomiast odbicie zwierciadlane można otrzymać jako złożenia obrotu i inwersji. Dlatego niezmienniczość względem inwersji nazywamy symetrią zwierciadlaną.

symetria zwierciadlana

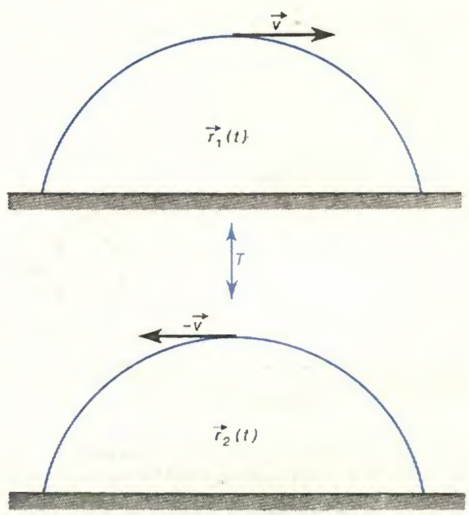
Symetria ta oznacza, że wraz z każdym procesem dozwolonym w ramach teorii jest też możliwy proces będący jego zwierciadlanym odbiciem względem dowolnej płaszczyzny. Zjawiskiem, w którym symetria zwierciadlana wydaje się być łamana, jest odchylenie igły magnetycznej umieszczonej równolegle do przewodnika prostoliniowego, w którym płynie prąd. Płaszczyzna, w której znajdują się igła i przewodnik przed włączeniem prądu jest z makroskopowego punktu widzenia płaszczyzną symetrii. Zatem fakt, że po włączeniu prądu płaszczyzna ta przestaje być płaszczyzną symetrii układu, zdaje się świadczyć o tym, że oddziaływanie elektromagnetyczne pomiędzy prądem w przewodniku a igłą magnetyczną nie jest niezmiennicze względem odbić zwierciadlanych. Aby się przekonać, że tak nie jest, należy wziąć pod uwagę mikroskopową budowę igły magnetycznej. Jej namagnesowanie polega na tym, że wypadkowa momentów pędu elektronów jest skierowana równolegle do osi igły ze zwrotem od bieguna południowego do bieguna północnego. Mikroskopowy, wirowy ruch elektronów i jonów powoduje, iż płaszczyzna przechodząca przez oś igły nie jest jej płaszczyzną symetrii. Odbicie zwierciadlane względem takiej płaszczyzny polega na zamianie miejscami biegunów igły magnetycznej. Igła z biegunami zamienionymi miejscami odchyli się po włączeniu prądu w przewodniku w kierunku przeciwnym w stosunku do igły w pierwotnym położeniu (rys. 4). Świadczy to o tym, że oddziaływania elektromagnetyczne nie łamią symetrii zwierciadlanej.



Rys. 4. Zjawiska przedstawione na rysunku są przed włączeniem prądu nawzajem swymi zwierciadlanymi odbiciami względem płaszczyzny przechodzącej przez oś igły magnetycznej i przewodnik

odbicia w czasie

Odbicia w czasie są symetriami teorii, jeśli wraz z każdym procesem jest możliwy w jej ramach proces zachodzący w odwrotnej kolejności. Na przykład teoria rzutów w ziemskim polu grawitacyjnym bez uwzględnienia oporu powietrza jest niezmiennicza względem odbić w czasie (rys. 5). Powszechnie obserwuje się istnienie zjawisk nieodwracalnych, jakimi są np. rzuty z uwzględnieniem oporu powietrza. Nieodwracalność procesów makroskopowych wiążemy z dużą liczbą obiektów mikroświata biorących w nich udział. Procesy w mikroświecie są natomiast prawie zawsze odwracalne.



Rys. 5. Odbicia w czasie są symetriami teorii rzutów w polu siły ciężkości (z pominięciem oporu powietrza). Ruchy przedstawione na rysunku związane są zależnością $\vec{r}_2(t) = \vec{r}_1(-t)$

Odbicia różnią się od takich symetrii czasoprzestrzeni jak przesunięcia bądź obroty tym, że nie są sparametryzowane wielkościami tego rodzaju co kąt obrotu lub wektor przesunięcia, które mogą przybierać wartości dowolnie małe. Ponieważ nie ma małych odbić, a więc nie wchodzą one w skład jednoparametrowach grup symetrii, nie można tu zatem zastosować twierdzenia Noether. W fizyce klasycznej z odbiciami nie wiążą się żadne zasady zachowania.

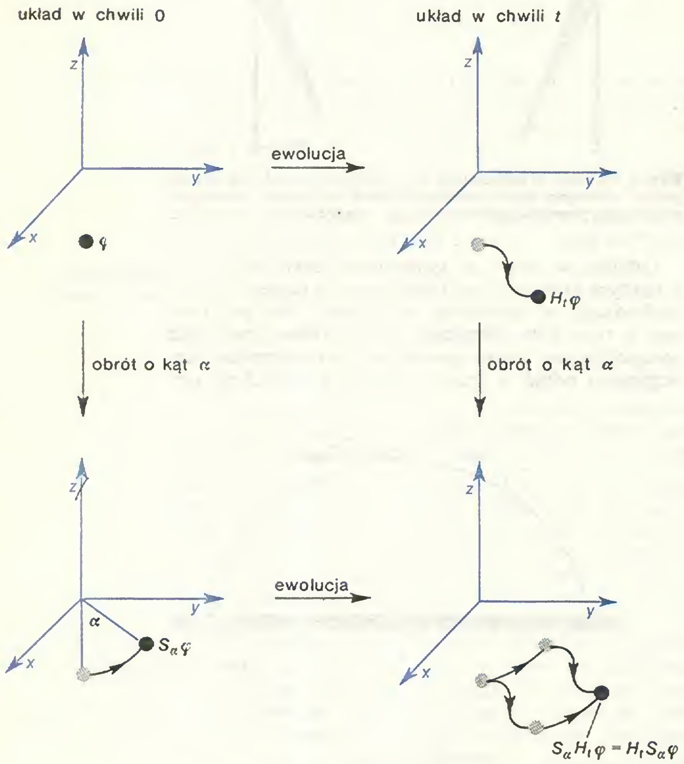
Zasady zachowania w fizyce kwantowej

związek zasad zachowania z symetriami

Związek symetrii z zasadami zachowania tkwi w samych podstawach teorii kwantowych. W fizyce kwantowej stanom układu odpowiadają wektory w pewnej abstrakcyjnej przestrzeni (przestrzeni Hilberta), natomiast mierzalnym wielkościom fizycznym takim jak położenie, pęd itp. — operatory, działające na te wektory. Przekształcenia układu takie jak przesunięcia, obroty itp. są tu także reprezentowane przez operatory. Wśród nich ważną rolę odgrywają operatory przesunięcia w czasie H_t.

Niech φ oznacza pewien stan układu odosobnionego, natomiast $S_\alpha\varphi$ — stan układu po wykonaniu przekształcenia S_α. Przekształceniem S_α może być np. obrót układu o kąt α wokół osi z (rys. 6). Po upływie czasu t stany te przejdą odpowiednio w stany $H_t\varphi$ i $H_tS_\alpha\varphi$. Jeśli S_α jest symetrią, to ten ostatni stan można także otrzymać w wyniku działania S_α na $H_t\varphi$, a zatem $H_tS_\alpha\varphi = S_\alpha H_t\varphi$. Równość ta powinna zachodzić dla wszystkich wektorów φ, a więc warunek na to, aby przekształcenie S_α było symetrią, możemy zapisać następująco:

$$H_tS_\alpha = S_\alpha H_t.$$



Rys. 6. Przemienność operatorów obrotu S_α wokół osi z z operatorami ewolucji ($S_\alpha H_t = H_t S_\alpha$) oznacza, że z-owa składowa momentu pędu J_z jest zachowana

Operatory reprezentujące przekształcenia układów kwantowych nie odpowiadają na ogół żadnym mierzalnym wielkościom fizycznym (wyjątkiem jest tu operator inwersji). Natomiast z każdą jednoparametrową grupą przekształceń (S_α) jest ściśle związany operator S reprezentujący mierzalną wielkość fizyczną. Operator ten jest proporcjonalny do granicy wyrażenia $(S_\alpha - 1)/\alpha$ przy α dążącym do 0. Na przykład, z obrotami wokół osi z jest związany operator z-owej składowej momentu pędu J_z, a z przesunięciami w czasie H_t — operator energii H, zwany też operatorem Hamiltona itd. Warunek przemienności

$$HS = SH,$$

oznacza, że wielkość fizyczna odpowiadająca operatorowi S jest zachowana.

W fizyce kwantowej wiele wielkości fizycznych może przybierać ściśle określone, skwantowane wartości. Do wielkości tych należy przede wszystkim energia i moment pędu, przy czym kwantyzacja energii dotyczy tylko układów związanych, takich jak atomy i cząsteczki, natomiast kwantyzacja momentu pędu ma znaczenie uniwersalne.

kwantyzacja momentu pędu

Moment pędu $\vec{J}$ pojedynczej cząstki jest w mechanice kwantowej sumą orbitalnego momentu pędu $\vec{L} = \vec{r}\times\vec{p}$ i spinu $\vec{S}$:

$$\vec{J} = \vec{L} + \vec{S}.$$

Długość wektora spinu $|\vec{S}|$ nie zależy przy tym ani od stanu ruchu cząstki, ani od wyboru układu odniesienia. Wielkość ta, podobnie jak masa, charakteryzuje samą cząstkę.

Dopuszczalnymi wartościami długości momentu pędu dowolnego układu są

$$|\vec{J}| = \hbar\sqrt{j(j+1)}\text{, gdzie } j = 0, {}^1/_2, 1, {}^3/_2, 2, \ldots$$

Podobna reguła kwantyzacji dotyczy spinu i orbitalnego momentu pędu, z tą różnicą, że liczba kwantowa orbitalnego momentu pędu może przybierać tylko wartości całkowite. Jeśli długość wektora momentu pędu $|\vec{J}|$ jest zadana, to rzut wektora J na wybraną oś — np. oś z — może przybierać wartości

$$J_z = \hbar m\text{, gdzie } m = -j, -j+1, \ldots, j-1, j.$$

Liczba kwantowa j całkowitego momentu pędu i liczba kwantowa m jego rzutu na oś z w pełni charakteryzują moment pędu układu — układ kwantowy nie może mieć jednocześnie określonych wartości rzutów wektora momentu pędu na dwie różne osie, np. J_x i J_z.

Duże znaczenie w interpretacji kwantowej zasady zachowania momentu pędu ma kwantowe prawo dodawania momentów pędu. W fizyce klasycznej moment pędu układu złożonego $\vec{J}$ jest sumą momentów pędu jego podukładów $\vec{J} = \vec{J_1} + \vec{J_2}$. W fizyce kwantowej prawo to dotyczy operatorów momentu pędu. Liczba kwantowa m jest addytywna, $m = m_1 + m_2$, natomiast liczba kwantowa j może przybierać następujące wartości:

$$j = |j_1 - j_2|,\ |j_1 - j_2| + 1, \ldots, j_1 + j_2.$$

Reguła ta dotyczy także dodawania orbitalnego momentu pędu i spinu.

kwantowa zasada zachowania momentu pędu

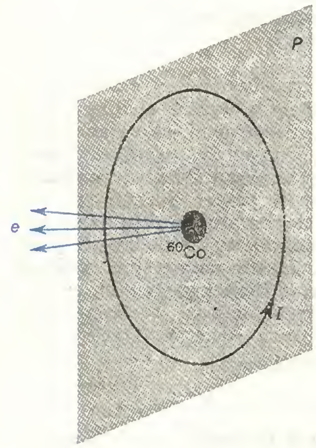
Klasyczna zasada zachowania orbitalnego momentu pędu $\vec{L}$ została w fizyce kwantowej uogólniona na zasadę zachowania całkowitego momentu pędu $\vec{J}$. Właśnie dzięki uwzględnieniu spinu $\vec{S}$, zasada zachowania momentu pędu pozostaje słuszna. Rozpatrzmy jeden z wniosków płynący z kwantowej wersji tej zasady. Ponieważ liczba kwantowa orbitalnego momentu pędu jest całkowita, z reguły dodawania momentów pędu wynika, że j jest liczbą połówkową wtedy i tylko wtedy, gdy układ zawiera nieparzystą liczbę cząstek o spinie połówkowym, zwanych fermionami. Zatem, gdy mamy układ izolowany złożony z samych bozonów, tj. cząstek o spinie całkowitym, fermiony mogą się w nim pojawiać wyłącznie parami. Z tego samego powodu mogą one ginąć (przekształcać się w bozony) tylko parami.

Parzystość

W fizyce kwantowej operatorowi inwersji P odpowiada wielkość fizyczna — parzystość. Dwukrotne poddanie układu fizycznego inwersji daje układ wyjściowy, a zatem $P^2 = 1$. Wynika z tego, że parzystość

wartości parzystości

układów fizycznych może przybierać tylko dwie wartości: +1 lub −1. Parzystość jest wielkością multiplikatywną, gdyż parzystość układu złożonego z nie oddziałujących między sobą podukładów jest iloczynem parzystości podukładów. Właściwość ta w istotny sposób odróżnia parzystość od addytywnych wielkości zachowanych, takich jak energia lub ładunek elektryczny. Określoną parzystość mogą mieć tylko te układy, które zawierają parzystą liczbę fermionów — pojedynczy fermion nie ma określonej parzystości.



Rys. 7. Próbka kobaltu umieszczona w silnym jednorodnym polu magnetycznym w temperaturze bliskiej 0 K emituje prawie wszystkie elektrony na jedną stronę płaszczyzny P

W przyrodzie, jak się okazuje, występują takie procesy, w których symetria zwierciadlana jest łamana, a w konsekwencji parzystość nie jest zachowana. Należy do nich rozpad β. Łamanie symetrii zwierciadlanej w rozpadzie β jest widoczne w doświadczeniu pani C.S.Wu, którego koncepcja jest przedstawiona na rys. 7. W środku przewodnika kołowego umieszczono próbkę radioaktywnego kobaltu. Jądra kobaltu nie są początkowo w żaden sposób uporządkowane, toteż elektrony β są emitowane równomiernie we wszystkich kierunkach. Dopiero pole magnetyczne wytworzone przez prąd elektryczny płynący w kołowym przewodniku ustawia momenty pędów jąder kobaltu, tak jak małe igły magnetyczne — prostopadle do płaszczyzny $\mathcal{P}$, w której znajduje się przewodnik. Na podstawie rys. 4 nietrudno zauważyć, że odbicie zwierciadlane igły magnetycznej względem płaszczyzny prostopadłej do osi igły i przechodzącej przez jej środek pozostawia bieguny na swoich miejscach — jest więc ono symetrią igły magnetycznej. Zatem po uporządkowaniu jąder kobaltu przez pole magnetyczne, $\mathcal{P}$ pozostaje płaszczyzną symetrii całego układu. To, że elektrony powstałe w wyniku rozpadu β są emitowane nierównomiernie na obie strony płaszczyzny $\mathcal{P}$, łamiąc w ten sposób zwierciadlaną symetrię układu względem $\mathcal{P}$, dowodzi, iż oddziaływania słabe wywołujące rozpad β nie są niezmiennicze względem inwersji.

doświadczenie Wu

Próbą obrony uniwersalności symetrii zwierciadlanej jest hipoteza, że geometryczną operację inwersji P należy składać z operacją sprzężenia cząstka-antycząstka C, aby otrzymać „fizyczną" inwersję CP, która byłaby symetrią wszystkich oddziaływań, w tym oddziaływań słabych. „Fizycznym" zwierciadlanym odbiciem układu przedstawionego na rys. 7 byłby wtedy układ, w którym kobalt byłby zastąpiony antykobaltem, natomiast elektrony płynące w przewodniku — pozytonami. Emisja pozytonów, powstałych w wyniku rozpadu antykobaltu w kierunku przeciwnym niż zaznaczony na rysunku kierunek emisji elektronów świadczyłaby o tym, że CP jest symetrią oddziaływań słabych.

Przekształcenia C i CP mają właściwości analogiczne do właściwości inwersji P, toteż odpowiadające im wielkości fizyczne, noszące odpowiednio nazwy parzystości ładunkowej i parzystości kombinowanej, mają właściwości analogiczne do właściwości parzystości zwykłej. Liczne doświadczenia przez długi czas potwierdzały, iż CP jest uniwersalną symetrią przyrody, a więc, że parzystość kombinowana jest bezwzględnie zachowana. Ponieważ parzystość nie jest zachowana w oddziaływaniach słabych, oznacza to, że w oddziaływaniach tych nie jest też zachowana parzystość ładunkowa. Analiza rozpadu mezonów K doprowadziła jednak do stwierdzenia, że również parzystość kombinowana nie jest, co prawda nieznacznie, zachowana w oddziaływaniach słabych.

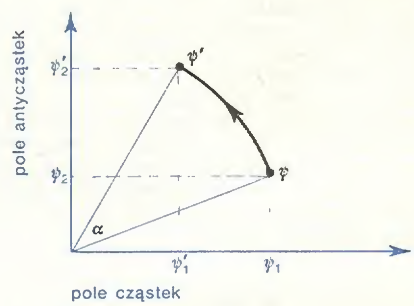
parzystość ładunkowa i kombinowana

Wielkości zachowane typu ładunku elektrycznego

Zasada zachowania ładunku elektrycznego jest najwcześniej poznaną zasadą zachowania nie związaną z symetriami czasoprzestrzeni. Zasada ta, inaczej niż zasady omawiane poprzednio, powstała nie dzięki rozważaniom teoretycznym, lecz została stwierdzona doświadczalnie. Ładunek elektryczny jest wielkością addytywną; jego cechą charakterystyczną jest ziarnistość — występuje on zawsze w ilościach będących całkowitą wielokrotnością ładunku elementarnego.

zachowanie ładunku elektrycznego

Zasada zachowania ładunku jest związana, za pomocą twierdzenia Noether, z transformacjami cechowania. W klasycznej teorii pola cząstki określonego rodzaju (np. elektrony) opisujemy za pomocą pola $\psi_1 = \psi_1(\vec{x}, t)$, ich antycząstki zaś (np. pozytony) za pomocą pola $\psi_2 = \psi_2(\vec{x}, t)$. Pole $\psi = (\psi_1, \psi_2)$ opisujące układ złożony z cząstek i ich antycząstek możemy sobie wyobrazić jako punkt na płaszczyźnie (rys. 8); pola ψ_1, ψ_2 są przy tym odłożone na osiach



Rys. 8. Transformacja cechowania

prostopadłych. Transformacją cechowania (pierwszego rodzaju) nazywa się obrót o pewien kąt α wokół punktu przecięcia tych osi. Niezmienniczość teorii względem wszystkich transformacji cechowania zapewnia zachowanie ładunku elektrycznego. W teorii kwantowej, operator ładunku elektrycznego jest związany z operatorami cechowania — obrotu w płaszczyźnie par cząstka–antycząstka — w podobny sposób jak operator J_z z obrotami w płaszczyźnie prostopadłej do osi z. Wynika z tego, że skwantowanie ładunku elektrycznego jest analogiczne do skwantowania z-owej składowej momentu pędu. Zatem, powiązanie zachowania ładunku elektrycznego z transformacjami cechowania, tłumaczy jego ziarnistość.

transformacje cechowania

Niezmienniczość względem transformacji cechowania zależnych od punktu czasoprzestrzeni, $\alpha = \alpha(\vec{x}, t)$, zwanych transformacjami cechowania drugiego rodzaju, jest możliwa tylko wtedy, gdy uwzględni się oddziaływanie pola ψ z polem elektromagnetycznym. Przy tym charakter tego oddziaływania, jak i niektóre właściwości pola elektromagnetycznego — np. prawo indukcji Faradaya — są jej konsekwencją.

Liczbę barionową B różną od zera przypisujemy tym cząstkom elementarnym, które są fermionami i oddziaływają silnie. Przy czym cząstki o liczbie barionowej $+1$, zwane barionami, są związane procesami rozpadu ze swym najlżejszym reprezentantem — protonem. Antybariony mają liczbę barionową -1. Liczba barionowa jest wielkością addytywną. Liczba barionowa atomów, będąca sumą liczby protonów i neutronów zawartych w ich jądrach, pokrywa się z ich liczbą masową. Zasada zachowania liczby barionowej jest zatem uogólnieniem zasady zachowania liczby masowej, obowiązującej w reakcjach chemicznych. Liczne obserwacje i doświadczenia doprowadziły do przekonania, że liczba barionowa jest wielkością zachowaną bezwzględnie. Przemawiają za tym przede wszystkim dwa następujące fakty: po pierwsze, nie zaobserwowano nigdy żadnego z ciężkich barionów — neutronu, hiperonów Λ^0, Σ^0, Σ^+, Σ^-, Ξ^0, Ξ^-, lub rezonansów barionowych — który nie rozpadłby się w końcu na proton (albo związany w jądrze neutron) i cząstki lżejsze mające liczbę barionową 0. Po drugie, ze względu na obfitość protonów, rozpad protonu byłby zjawiskiem nietrudnym do zauważenia, tymczasem zjawiska takiego nigdy nie zaobserwowano.

zachowanie liczby barionowej

Fermiony, które nie biorą udziału w oddziaływaniach silnych, nazywa się leptonami lub antyleptonami. Liczbę leptonową $L = +1$ mają leptony: elektron e^-, neutrino elektronowe ν_e, mion μ^- i neutrino mionowe ν_μ. Ich antycząstki mają liczbę leptonową $L = -1$. Liczba leptonowa jest wielkością addytywną. Zasada zachowania liczby leptonowej jest wynikiem analizy reakcji z udziałem leptonów i antyleptonów. Jak wiadomo, fermiony mogą powstawać i znikać wyłącznie parami. Z zasad zachowania liczby barionowej i liczby leptonowej wynika, że pary te muszą składać się z barionu i antybarionu, albo z leptonu i antyleptonu.

zachowanie liczby leptonowej

Wspólnymi cechami liczby barionowej, liczby leptonowej i ładunku elektrycznego są: addytywność, ziarnistość oraz zmiana znaku przy zastąpieniu cząstki antycząstką. Dlatego zasady zachowania liczby barionowej i leptonowej są, podobnie jak zasada zachowania ładunku, związane z transformacjami cechowania. Nie można ich jednak wiązać z transformacjami cechowania drugiego rodzaju, gdyż liczby barionowa i leptonowa, w odróżnieniu od ładunku elektrycznego, nie są bezpośrednimi źródłami pól o właściwościach analogicznych do właściwości pola elektromagnetycznego.

Zachowanie izospinu

Przeglądając tabelę cząstek elementarnych (str. 84), widzimy że mezony i bariony, a więc cząstki oddziaływające silnie, tworzą grupy o bliskich sobie masach i jednakowych spinach. Grupy te noszą nazwę multipletów izospinowych. Multiplet izospinowy tworzy np. grupa pionów: π^-, π^0 i π^+; inny przykład to multiplet nukleonów, złożony z protonu p i neutronu n. Liczne dane doświadczalne świadczą o tym, że oddziaływania silne cząstek należących do jednego multipletu są jednakowe. Uzasadnia to formalne traktowanie tych cząstek jako jednej cząstki mogącej znajdować się w różnych stanach. Stany te numerujemy liczbą I_3, przybierającą $2I+1$ wartości ($2I+1$ jest więc liczbą cząstek w multiplecie), $I_3 = -I, -I+1, \ldots, I$, dbając przy tym o to, aby najmniejsza wartość I_3 odpowiadała cząstce o najmniejszym ładunku. Tak określone liczby kwantowe I i I_3 mają właściwości analogiczne do liczb kwantowych całkowitego spinu i jego trzeciej składowej. Z tego powodu nazywa się je odpowiednio izospinem i trzecią składową izospinu.

izospin a spin

Okazuje się, że analogia pomiędzy izospinem a spinem jest bliższa. Jeśli bowiem umówimy się, aby obliczać I i I_3 dla układów cząstek zgodnie z kwantowym prawem dodawania momentów pędu, to izospin i jego trzecia składowa będą zachowane w reakcjach wywoływanych przez oddziaływania silne. Izospin jest zachowany wyłącznie w tych reakcjach, toteż cząstkom nie oddziaływającym silnie nie przypisuje się żadnej wartości izospinu. Zasadę zachowania izospinu tłumaczymy niezmienniczością teorii silnych oddziaływań względem obrotów w abstrakcyjnej, trójwymiarowej przestrzeni, zwanej przestrzenią izospinową.

Znaczenie zasad zachowania w fizyce cząstek elementarnych

wykorzystanie zasad zachowania

Zasady zachowania liczby barionowej, liczby leptonowej i izospinu powstały w wyniku analizy reakcji pomiędzy cząstkami elementarnymi, jako usankcjonowanie występowania jednych a niewystępowania innych reakcji. Zasady te wraz z zasadami poznanymi wcześniej wprowadziły pewien porządek w licznej rodzinie cząstek elementarnych. Pozwoliły one m.in. sklasyfikować cząstki wg posiadanych przez nie liczb kwantowych. Zasady zachowania zawierają też przepis na określanie liczb kwantowych cząstek nowo odkrywanych. Przypuśćmy np., że chcemy określić liczby kwantowe rezonansu $\Lambda(1815)$, który ulega tzw. rozpadowi szybkiemu

$$\Lambda(1815) \rightarrow p + K^-.$$

Znając liczbę barionową, ładunek i izospin protonu: $B = 1$, $Q = e$, $I = 1/2$, $I_3 = -1/2$ oraz mezonu K^-: $B = 0$, $Q = -e$, $I = 1/2$, $I_3 = -1/2$, bez trudu dochodzimy do wniosku, że $\Lambda(1815)$ ma $B = 1$, $Q = 0$, $I_3 = 0$. Izospin tego rezonansu, który powinien być zachowany w tym procesie wywołanym przez oddziaływania silne, mógłby przybierać wartości $I = 0$ lub $I = 1$. Gdyby $I = 1$, rezonans ten musiałby mieć dwóch towarzyszy o bliskich masach, czego jednak nie zaobserwowano; a zatem $I = 0$. Masę tego rezonansu $m \approx 1815$ MeV można wyznaczyć, mierząc energię i pędy protonu i mezonu K^-, a następnie stosując zasadę zachowania energii-pędu. Wyznaczenie spinu wymaga subtelniejszych metod.

symetria a zasady zachowania

Związane z zasadami zachowania symetrie praw fizyki prowadzą często do wniosków wykraczających poza same zasady zachowania. Nie wszystkie zresztą

Zasada zachowania i symetrie w zależności od typu oddziaływania

Wielkość fizyczna	Symetria	Czy zasada zachowania obowiązuje w oddziaływaniach silnych?	elektromagnetycznych?	słabych?
Energia	przesunięcie w czasie	tak	tak	tak
Pęd	przesunięcia w przestrzeni	tak	tak	tak
Moment pędu	obroty	tak	tak	tak
Ładunek elektryczny	transformacje cechowania	tak	tak	tak
Liczba barionowa	transformacje cechowania	tak	tak	tak
Liczba leptonowa	transformacje cechowania	tak	tak	tak
Parzystość	inwersja	tak	tak	nie
Parzystość ładunkowa	sprzężenie cząstka-antycząstka	tak	tak	nie
Parzystość kombinowana	inwersja kombinowana	tak	tak	tak
3-cia składowa izospinu	obroty wokół 3-ciej osi w przestrzeni izospinu	tak	tak	nie
Izospin	obroty w przestrzeni izospinu	tak	nie	nie

zasady zachowania a oddziaływania

symetrie są związane z zasadami zachowania. Na przykład niezmienniczości względem odbicia w czasie T, zapewniającej odwracalność zjawisk na poziomie mikroskopowym, nie odpowiada żadna zasada zachowania. To samo dotyczy złożenia inwersji kombinowanej i odbicia w czasie CPT. Wykazano teoretycznie, że symetria CPT powinna obowiązywać w każdej teorii relatywistycznej. Symetria ta prowadzi do wielu sprawdzanych doświadczalnie wniosków, m.in. zapewnia ona równość mas cząstki i antycząstki.

Jest wiele zasad zachowania, które obowiązują nie przy wszystkich, a tylko przy niektórych oddziaływaniach. Najwięcej wielkości jest zachowanych w oddziaływaniach silnych, nieco mniej w elektromagnetycznych, najmniej w słabych (tabela). A więc, im oddziaływanie jest słabsze, tym mniej zasad zachowania obowiązuje w wywoływanych przez nie zjawiskach. Nie jest jeszcze znane zadowalające wytłumaczenie tej prawidłowości. Nie wiadomo także, czy dotyczy ona najsłabszego z oddziaływań, mianowicie grawitacyjnego, gdyż nie zaobserwowano żadnych reakcji między cząstkami elementarnymi, których by ono było przyczyną.

Badania teoretyczne wskazują jednak na to, że w bardzo silnych polach grawitacyjnych, w pobliżu czarnych dziur, są możliwe procesy kreacji cząstek łamiące zasady zachowania liczby barionowej i liczby leptonowej.

G. Białkowski, R. Sosnowski *Cząstki elementarne*, Warszawa 1971; W. Rubinowicz, W. Królikowski *Mechanika teoretyczna*, Warszawa 1980.

Termodynamika fenomenologiczna

Stanisław Piasecki

Przedmiotem badań termodynamiki są właściwości układów makroskopowych rozpatrywane z punktu widzenia ich zależności od temperatury. Ten olbrzymi dział fizyki rozwinął się początkowo jako teoria fenomenologiczna, nie uwzględniająca atomowej (mikroskopowej) struktury materii. Zasady termodynamiki fenomenologicznej, sformułowane w XIX w., przetrwały po dzień dzisiejszy. Stanowią one uogólnienie wyników niezliczonych obserwacji zachowania się makroskopowych ilości materii. Za typowy układ makroskopowy można przyjąć 1 mol określonej substancji. Zawiera on niewyobrażalnie wielką liczbę odpowiednich składników mikroskopowych (np. dla wody — cząsteczek H_2O, dla gazu elektronowego — elektronów), równą liczbie atomów tworzących 0,012 kg węgla ^{12}C. Liczba ta, zwana liczbą Avogadra, wynosi $N_A = (6{,}022169 \pm 0{,}000040) \cdot 10^{23}$ mol^{-1}. W zależności od struktury układu rolę składników mikroskopowych mogą odgrywać cząstki elementarne, atomy, cząsteczki, jony itp. W opisie termodynamicznym tak ogromnych zbiorowisk oddziałujących mikrocząstek uwzględnia się jedynie te właściwości, które dają się zaobserwować na poziomie makroskopowym. Charakteryzują je parametry makroskopowe, takie jak objętość, ciśnienie, temperatura. Powiązaniem tych wielkości ze strukturą mikroskopową materii zajmuje się termodynamika statystyczna (→ Termodynamika statystyczna). Należy podkreślić, że pojęcia oraz prawa termodynamiki fenomenologicznej tworzą spójną całość, dostateczną do rozwiązania wielu praktycznych problemów. Ten klasyczny dział fizyki nadal się żywo rozwija, obejmując swym zakresem coraz więcej zjawisk. Na przykład współcześnie są podejmowane próby zastosowania pojęć termodynamicznych do opisu właściwości czarnych dziur (→ Czarne dziury i grawitacyjne zapadanie). Odkryto bowiem niezmiernie interesujące właściwości brzegu czarnej dziury, zwanego horyzontem zdarzeń. Okazuje się, że w procesach pochłaniania materii lub promieniowania pole powierzchni horyzontu zdarzeń zawsze rośnie. Nasunęło to myśl o możliwości interpretowania go jako entropii czarnej dziury (analogia z II zasadą termodynamiki). Aby tę tezę rozwinąć, należało jednak najpierw zbadać, czy jest możliwe występowanie stanów równowagi termodynamicznej czarnej dziury z otoczeniem (entropię definiuje się w termodynamice dla stanów równowagi). Problem ten okazał się trudny, gdyż zgodnie z prawami fizyki klasycznej czarna dziura może jedynie pochłaniać promieniowanie, jest zaś niezdolna do wysyłania go, co wyklucza równowagę z otoczeniem. Możliwość emisji cząstek przez czarne dziury została jednak udowodniona przez S.W. Hawkinga na podstawie praw mechaniki kwantowej. W ten sposób zostały stworzone podstawy do rozwoju dalszych badań nad możliwością termodynamicznego opisu tych niezwykłych układów.

układ makroskopowy

termodynamika czarnej dziury

Innym przykładem dynamicznie rozwijającego się działu termodynamiki jest teoria przemian fazowych. Dotyczy ona wielu rozmaitych zjawisk, takich jak nadciekłość, nadprzewodnictwo, zmiany stanu skupienia, zmiany struktury krystalicznej, przejście od stanu paramagnetycznego do ferro- lub antyferromagnetycznego i wielu innych (→ Przejścia fazowe i zjawiska krytyczne). Wszystkie one zależą w sposób istotny od temperatury i stanowią naturalny przedmiot badań termodynamiki.

teoria przemian fazowych

Stwierdzono doświadczalnie, że układy makroskopowe izolowane od wpływu otoczenia osiągają po dostatecznie długim czasie szczególnie proste stany, w pełni scharakteryzowane przez wartości niewielkiego zespołu parametrów makroskopowych. Stany te, w razie braku jakichkolwiel systematycznych zmian, nazywa się stanami równowagi termodynamicznej. W opisie ich szczególną rolę odgrywa odpowiadająca im energia układu. Energię zgromadzoną w układzie będącym w stanie równowagi nazywa się w termodynamice energią wewnętrzną i oznacza przez U. Ponieważ sens fizyczny mają jedynie różnice energii, określa się ją przyjmując, że w wybranym stanie $U = 0$. Energia wewnętrzna jest parametrem ekstensywnym. Oznacza to, że np. w układzie złożonym z dwu podukładów makroskopowych o energiach U_1 i U_2 całkowita energia wewnętrzna wynosi $U = U_1 + U_2$. Wzór ten jest ścisły, gdy nie ma oddziaływania między podukładami. Stanowi on jednak również bardzo dobre przybliżenie przy oddziaływaniach krótkozasięgowych. Energia oddziaływania jest wówczas efektem powierzchniowym, który dla dostatecznie dużych (makroskopowych) podukładów jest pomijalny. Przykładami parametrów ekstensywnych są również objętość układu V i liczby moli N_1, N_2,...N_r tworzących go składników. Jakkolwiek termodynamika stosuje się do układów o złożonych właściwościach mechanicznych, elektromagnetycznych i cieplnych, jej istotny sens można wyjaśnić, rozpatrując tzw. proste układy termodynamiczne. Z definicji ich stany równowagi są całkowicie określone przez wartości zespołu parametrów $(U, V, N_1, ..., N_r)$. Przykładem fizycznym prostego układu jednoskładnikowego może być wypełniona N molami gazu szlachetnego objętość V. Stany równowagi takiego układu można przedstawiać za pomocą punktów o współrzędnych (U, V, N) w trójwymiarowej przestrzeni parametrów stanu.

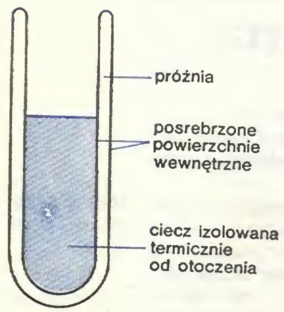
stan równowagi termodynamicznej

prosty układ termodynamiczny

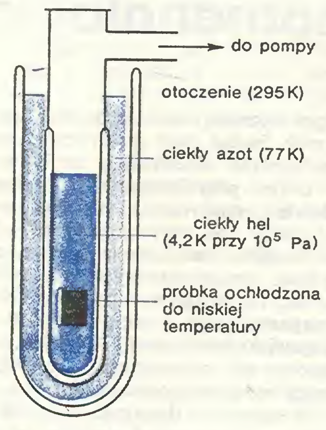
Stan równowagi termodynamicznej ustala się zawsze w określonych warunkach zewnętrznych. Oddziaływanie układu z otoczeniem charakteryzuje się

kontakt termodynamiczny układu z otoczeniem

przez wprowadzenie pojęcia odpowiednich ścianek. Ściankę, która wyklucza jakąkolwiek formę oddziaływania układu z otoczeniem, nazywamy izolującą. Układ izolowany jest oczywiście daleko posuniętą idealizacją. Łatwiej uzyskać osłonięcie nie dopuszczające do wymiany cząstek z otoczeniem, wówczas układ nazywamy zamkniętym. Z kolei układ otwarty to taki, który może wymieniać cząstki z otoczeniem. Osłonięciu ściankami adiabatycznymi towarzyszy ograniczenie wpływu otoczenia na układ do możliwości wykonania nad nim pracy mechanicznej. Ich bardzo dobrą przybliżoną realizacją są ścianki naczynia Dewara, np. w postaci dwu posrebrzanych płytek szklanych przedzielonych obszarem wysokiej próżni (rys. 1, 2). Z doświadczeń wiadomo, że poza przepływem cząstek i wykonaniem pracy mechanicznej występują jeszcze inne formy oddziaływania powodujące wymianę energii między układami i zmianę ich



Rys. 1. Naczynie Dewara używane do pracy w niskich temperaturach. Naczynia takie (nazwa pochodzi od nazwiska J. Dewara, który pierwszy skroplił wodór w 1898 r.) są podobne do termosów, izolują termicznie od otoczenia zawarty w nich płyn. Mogą być wykonane ze szkła, albo metalu, np. stali nierdzewnej. Izolację stanowi próżnia panująca wewnątrz podwójnej ścianki. W szklanych naczyniach Dewara zwykle pokrywa się wewnętrzne powierzchnie ścianek odbijającą warstwą srebra, aby zmniejszyć straty ciepła na promieniowanie



Rys. 2. Podwójne naczynie Dewara używane do pracy w temperaturze rzędu 1 K. Wewnętrzne naczynie Dewara, wypełnione ciekłym helem, jest zanurzone — w celu zmniejszenia strat ciepła — w drugim naczyniu Dewara, wypełnionym ciekłym azotem

parametrów stanu. Określa się je mianem oddziaływania termicznego. Ścianki dopuszczające jedynie kontakt termiczny z otoczeniem nazywają się ściankami diatermicznymi.

I zasada termodynamiki

I zasada termodynamiki jest formułowana zazwyczaj jako zasada zachowania energii dla układów osłoniętych adiabatycznie. Jej treść stanowi stwierdzenie, że praca nad układem osłoniętym adiabatycznie wykonana w procesie przejścia od początkowego stanu równowagi A do stanu końcowego B jest całkowicie określona przez te stany, niezależnie od przebiegu procesu. Oznaczając tę pracę symbolem W_{ad} (A, B) możemy napisać, że

$$W_{ad}(A,B) = U(B) - U(A),$$

przy czym z prawej strony występuje różnica między energią stanu końcowego i początkowego. Jeżeli proces nie przebiega adiabatycznie, to zasada zachowania energii dla układów zamkniętych przybiera postać:

$$U(B) - U(A) = W(A, B) + Q(A, B),$$

gdzie $Q(A,B)$ przedstawia zmianę energii związaną z kontaktem termicznym układu z otoczeniem. Wielkość $Q(A,B)$, zwana ciepłem przekazanym układowi w procesie przejścia ze stanu A do B, jest więc tego samego typu wielkością co praca mechaniczna. Mierzy się ją w dżulach. Zarówno praca jak i ciepło zależą nie tylko od stanów A i B ale również od konkretnej realizacji procesu powodującego zmianę stanu układu. Różnią się tym zasadniczo od energii wewnętrznej będącej funkcją stanu.

Fakt, że jedynie część procesów zgodnych z zasadą zachowania energii zachodzi w rzeczywistości, ujmuje precyzyjnie II zasada termodynamiki. Stwierdza ona, że dla każdego układu termodynamicznego istnieje funkcja stanu S, zwana entropią, o następujących właściwościach:

II zasada termodynamiki

— S jest wielkością ekstensywną,

— w procesach zachodzących w układach izolowanych entropia nigdy nie maleje.

Jeśli więc układ izolowany przechodzi ze stanu A do B, to $S(B) \geqslant S(A)$. W nierówności tej, wyrażającej prawo wzrostu entropii, jest zawarta informacja o nieodwracalnym charakterze procesów przebiegających w warunkach izolacji od otoczenia (procesy samorzutne). W szczególności wynika z niej, że spośród wszystkich możliwych wartości entropii w obecności ścianek wewnętrznych — stanowi równowagi bez ścianek odpowiada wartość maksymalna (usunięcie ścianki powoduje ewentualnie samorzutny proces prowadzący do wzrostu entropii). Znajomość zależności entropii od parametrów ekstensywnych charakteryzujących stany równowagi oznacza pełną informację o właściwościach termodynamicznych układu.

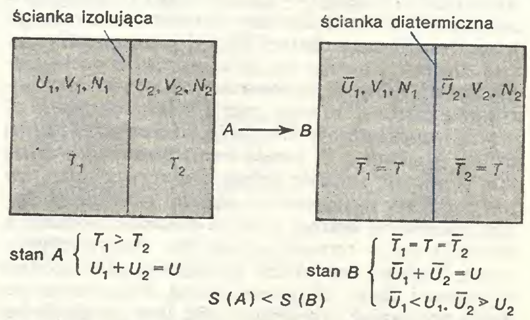
Na przykład dla prostego jednoskładnikowego układu termodynamicznego (zwanego płynem prostym) entropia jest funkcją trzech parametrów $S = S(U,V,N)$. Związek tego typu nazywa się równaniem podstawowym. Dla rzeczywistych układów fizycznych entropia jest funkcją rosnącą energii wewnętrznej. Można więc rozpatrywać równoważne równanie postaci $U = U(S, V, N)$. Przy niewielkich zmianach dS, dV, dN parametrów S, V, N zmiana energii wewnętrznej dU jest określona przez różniczkę $dU = (\partial U/\partial S)_{V,N}\, dS + (\partial U/\partial V)_{S,N}\, dV + (\partial U/\partial N)_{S,V}\, dN$. Wielkości $T = (\partial U/\partial S)_{V,N}$, $-p = (\partial U/\partial V)_{S,N}$, $\mu = (\partial U/\partial N)_{S,V}$ nazywane są parametrami intensywnymi układu. Charakteryzują one układ lokalnie i nie zależą od jego rozmiarów, przy czym T nazywamy temperaturą termodynamiczną, p — ciśnieniem, μ — potencjałem chemicznym. Z definicji pochodnej cząstkowej potencjał chemiczny obliczamy ze wzoru

$$\mu = \lim_{dN \to 0} [U(S, V, N+dN) - U(S, V, N)] / dN.$$

Określa on szybkość wzrostu energii wewnętrznej przy wzroście liczby cząstek układu (entropia i objętość pozostają ustalone). Podobnie ciśnienie mierzy stosunek zmian energii do zmian objętości (przy ustalonej entropii i liczbie cząstek), temperatura zaś określa szybkość wzrostu energii przy wzroście entropii, gdy stała jest liczba cząstek i objętość. Znaczenie fizyczne parametrów T, p i μ można zrozumieć lepiej na podstawie II zasady termodynamiki. Wiąże się ono bezpośrednio z warunkami równowagi układu.

warunki równowagi termodynamicznej

Rozważmy np. płyn prosty z izolującą ścianką, rozdzielającą go na dwa podukłady o liczbach moli N_1, N_2, objętościach V_1, V_2 i energiach U_1, U_2. Zastąpienie ścianki izolującej ścianką diatermiczną wprowadzi podukłady w kontakt termiczny (rys. 3).



Rys. 3. W układzie izolowanym złożonym z dwu podukładów rozdzielonych ścianką izolującą (stan A) po zastąpieniu tej ścianki ścianką diatermiczną zachodzi nieodwracalny proces przepływu energii w postaci ciepła od podukładu cieplejszego do zimniejszego. Końcowy podział energii między podukłady odpowiada równości ich temperatur termodynamicznych (stan B)

Zgodnie z zasadą wzrostu entropii, w wyniku przepływu ciepła ustali się taki podział energii między podukłady, któremu odpowiada wartość maksymalna entropii całkowitej przy ustalonych parametrach N_1, N_2, V_1, V_2 oraz $U = U_1+U_2$. Ze wzoru na różniczkę energii wewnętrznej wynika, że przy ustalonej objętości i liczbie cząstek różniczki entropii podukładów mają postać:

$$dS_1|_{V_1,N_1} = dU_1/T_1, \quad dS_2|_{V_2,N_2} = dU_2/T_2.$$

Ponieważ energia całkowita jest również stała, więc zmiany energii podukładów muszą się wzajemnie kompensować, tzn.

$$dU = dU_1+dU_2 = 0.$$

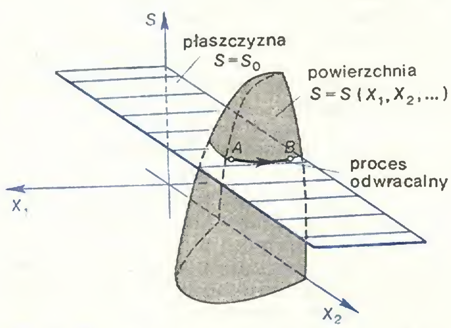
Dla entropii całkowitej, będącej sumą entropii podukładów, znajdujemy zatem

$$d(S_1+S_2)|_{V_1,N_1,V_2,N_2,U=U_1+U_2} = dU_1/T_1+(-dU_1)/T_2.$$

Warunkiem koniecznym osiągania przez entropię wartości maksymalnej jest znikanie jej różniczki. Równowaga termiczna pociąga więc za sobą równość $(1/T_1-1/T_2)dU_1 = 0$, oznaczającą równość temperatur termodynamicznych $T_1 = T_2$. W procesie samorzutnego wyrównywania się temperatur energia w postaci ciepła przepływa od ciała o wyższej temperaturze (cieplejszego) do ciała o niższej temperaturze (zimniejszego), np. przy małych różnicach temperatur, zgodnie z uzyskanym wzorem, zmiana entropii wynosi $dS = (1/T_1-1/T_2)dU_1$. Ponieważ $dS>0$, więc jeśli $T_1>T_2$, to $dU_1<0$ (ciało cieplejsze traci energię). Definicja termodynamiczna jest zatem zgodna z intuicyjnym pojęciem temperatury. W podobny sposób stwierdzamy, że dwa układy, między którymi jest możliwy przepływ cząstek, mogą pozostawać w równowadze jedynie wówczas, gdy ich potencjały chemiczne są równe $\mu_1 = \mu_2$. Jeżeli $T_1 = T_2$, zaś $\mu_1>\mu_2$, materia przepływa z obszaru o potencjale μ_1 do obszaru o potencjale μ_2. W różniczce dU występuje też znany z mechaniki parametr intensywny — ciśnienie p. Zasada wzrostu entropii prowadzi do wniosku, że warunkiem równowagi mechanicznej płynu jest stałość ciśnienia w całej objętości V.

procesy odwracalne i nieodwracalne

II zasada termodynamiki wprowadza doniosłe rozróżnienie między procesami odwracalnymi i nieodwracalnymi. Procesy odwracalne w układach izolowanych to takie, w których układ przechodzi przez



Rys. 4. Proces odwracalny w układzie izolowanym. Kwazistatyczna zmiana parametrów ekstensywnych $(X_1, X_2,...)$ podukładów powoduje przejście złożonego z nich układu izolowanego przez kontinuum stanów równowagi tworzących krzywą na powierzchni stałej entropii

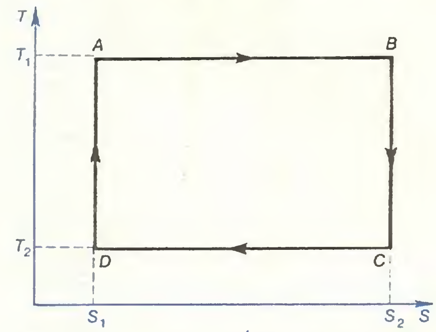
ciąg stanów równowagi przy stałej entropii całkowitej. Jest to idealizacja, którą jedynie w przybliżeniu można realizować doświadczalnie. Proces odwracalny musi mieć kwazistatyczny przebieg, tzn. na każdym jego etapie poszczególne makroskopowe części układu pozostają ze sobą w równowadze. W opisie takiego procesu nie trzeba brać pod uwagę czasu, wystarcza podanie kontinuum stanów równowagi, reprezentowanego przez krzywą w przestrzeni parametrów stanu (rys. 4). Dzięki stałości entropii proces może przebiegać w dowolnym kierunku (odwracalność).

Rozważmy przykładowo prosty układ termodynamiczny w kontakcie termicznym i mechanicznym z otoczeniem o temperaturze T_0 i ciśnieniu p_0. W procesie odwracalnym ciśnienie p i temperatura T tego układu spełniają warunki równowagi $p = p_0$, $T = T_0$. Zgodnie z prawami mechaniki praca elementarna związana ze zmianą objętości wynosi $đW = -p_0dV = -pdV$. Symbol $đ$ został tu użyty w celu podkreślenia, że $đW$ nie jest różniczką funkcji stanu. Praca związana ze zmianą stanu układu jest bowiem zależna nie tylko od stanów początkowego i końcowego, ale i od sposobu realizacji procesu. Ponieważ $dN = 0$ (liczba cząstek jest stała), więc $dU = TdS-pdV$, a uwzględniając I zasadę termodynamiki otrzymujemy $đQ+đW = TdS-pdV$. Zatem w procesach odwracalnych ciepło $đQ$ przekazane przez otoczenie do układu jest dane wzorem

$$đQ = TdS.$$

pojemność cieplna

Jest to podstawowy związek wiążący zmiany entropii z energią wymienianą z otoczeniem przez kontakt termiczny. Kwazistatyczne dostarczenie ciepła do układu przy stałej objętości powoduje zmianę temperatury zgodnie ze wzorem $(đQ)_V = (TdS)_V = T(\partial S/\partial T)_V dT$. Współczynnik $C_V = T(\partial S/\partial T)_V$ nazywamy pojemnością cieplną, $c_V = C_V/N$ zaś molowym ciepłem właściwym przy stałej objętości. Podobnie określa się ciepło właściwe przy stałym ciśnieniu $c_p = T(\partial S/\partial T)_p$. Jeżeli proces jest cykliczny,



Rys. 5. Cykl Carnota na płaszczyźnie (T, S)

to stan początkowy pokrywa się ze stanem końcowym — wówczas całkowita zmiana entropii $\Delta S = \oint đQ/T = 0$, przy czym całka jest obliczana wzdłuż krzywej zamkniętej, reprezentującej proces w przestrzeni parametrów stanu. Zastosujmy ten wzór do cyklu Carnota (rys. 5), w którym układ podlega kolejno przemianie izotermicznej przy temperaturze T_1 od stanu A do B, ochłodzeniu adiabatycznemu do stanu C o temperaturze $T_2< T_1$, przemianie izotermicznej do stanu D, po czym wraca na drodze adiabatycznej do wyjściowego stanu A. Wykorzystując fakt, że odwracalny proces adiabatyczny zachodzi bez zmiany entropii, otrzymujemy związek

cykl Carnota

$$\Delta S = Q(A, B)/T_1+Q(C,D)/T_2 = 0,$$

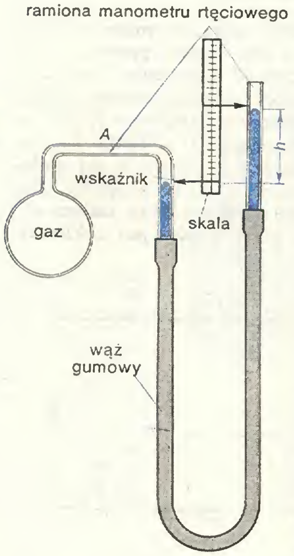
w którym $Q(A,B)$, $Q(C,D)$ oznaczają ciepło przekazane układowi w procesach izotermicznych. Wynika z niego, że ciepło $Q(A,B)$ i ciepło $Q(C,D)$ są zawsze przeciwnego znaku. Ponieważ zmiana energii wewnętrznej w cyklu wynosi zero, więc praca wykonana przez układ jest równa $[Q(A,B)+Q(C,D)]$. Jej stosunek do ciepła pobranego określa sprawność cyklu η. Zakładając, że $Q(A,B) > 0$, otrzymujemy $\eta = (1-T_2/T_1)$. Pomiar sprawności odwracalnego cyklu Carnota oznacza więc pomiar stosunku temperatur termodynamicznych, odpowiadających procesom izotermicznym. Wynik nie zależy od układu, który podlega przemianie cyklicznej.

bezwzględna skala temperatur Kelvina

Przypisując określonemu stanowi fizycznemu pewną wartość temperatury, określamy tym samym skalę temperatur. Powszechnie stosowana bezwzględna skala Kelvina przypisuje punktowi potrójnemu wody (stan, w którym współistnieją w równowadze lód, woda i para wodna) temperaturę 273,16 K. Jednostka tej skali jest zwana kelwinem (oznaczenie K). Iloczyn TS ma wymiar energii. Wprowadzając stałą Boltzmanna $k = (1{,}380622 \pm 0{,}000059) \cdot 10^{-23}\,\mathrm{JK^{-1}}$, określającą stosunek dżula do kelwina, możemy więc entropię dowolnego układu zapisać w postaci $S = \bar{S} \cdot k$, gdzie $\bar{S}$ jest wielkością bezwymiarową.

Iloczyn $k \cdot N_A = R = (8{,}31434 \pm 0{,}00035)\,\mathrm{JK^{-1}mol^{-1}}$ jest zwany stałą gazową. Pojawia się on np. w równaniach stanu jednoatomowego gazu doskonałego: $U = 3NRT/2$, $pV = 2U/3 = NRT$. Doświadczalnie stwierdzono, że związki te stanowią bardzo dobre przybliżenie dla realnych gazów jednoatomowych w obszarze wysokich temperatur i małych gęstości.

termometr gazowy



Rys. 6. Schemat termometru gazowego o stałej objętości. Gaz wypełnia zbiornik będący w kontakcie termicznym ze środowiskiem o mierzonej temperaturze T. Menisk rtęci w ramieniu A manometru doprowadza się do poziomu wskaźnika, tak aby mierzyć ciśnienie p zawsze tej samej objętości gazu termometrycznego (hel, azot lub wodór). Wartość p znajduje się z pomiaru różnicy poziomów rtęci h przy uwzględnieniu ciśnienia atmosferycznego. Znając ciśnienie p_0 odpowiadające znanej temperaturze $T_0 = 273{,}16$ K punktu potrójnego wody możemy obliczyć wartość T ze wzoru $T = T_0(p/p_0)$. Dokładne pomiary wymagają uwzględnienia odchyleń własności gazu termometrycznego od własności gazu doskonałego, zmian objętości zbiornika itp.

Wykorzystuje się je do pomiaru temperatury bezwzględnej (termodynamicznej) za pomocą termometrów gazowych (rys. 6). Odpowiadający im wzór na różnicę entropii między stanami B i A można znaleźć następująco: przy małych zmianach energii i objętości (przy ustalonej liczbie moli) zmiana entropii ma postać $dS = dU/T + pdV/T$, jest to wynik przekształcenia wzoru na różniczkę energii wewnętrznej; zatem w rozważanym przykładzie

$$dS = \frac{3}{2}NR \cdot \frac{dU}{U} + NR \cdot \frac{dV}{V} =$$

$$= NRd\left(\frac{3}{2}\ln U + \ln V\right) = NRd(\ln VU^{3/2}).$$

Przy przejściu ze stanu A do B przyrost entropii wynosi więc

$$S(B) - S(A) = NR\ln\{[U(B)/U(A)]^{3/2}[V(B)/V(A)]\}.$$

zagadnienie pracy maksymalnej

Z II zasady termodynamiki wynika wniosek o dużym znaczeniu praktycznym: w procesie przejścia ze stanu A do B układ wykonuje maksymalną pracę ($W_{\max}A,B$) wówczas, gdy proces jest odwracalny. W tym wypadku zmianę entropii układu $\Delta S = S(B) - S(A)$ kompensuje zmiana entropii otoczenia ($-\Delta S$). Natomiast gdy proces jest nieodwracalny, przyrost entropii układu jest taki sam, przyrost entropii otoczenia zaś jest większy (zasada wzrostu entropii). Ponieważ zmiana energii układu jest w obu wypadkach taka sama, równa $[U(B) - U(A)]$, więc większa entropia końcowa otoczenia oznacza, iż więcej wpłynęło do niego energii z układu w postaci ciepła, czyli że układ wykonał mniejszą pracę. Praca jest więc maksymalna, gdy proces jest odwracalny. W szczególności maksymalna sprawność cyklu jest osiągana przy jego odwracalnym przebiegu, toteż dla cyklu Carnota nie może ona przekroczyć wartości $(1 - T_2/T_1)$. Wiążą się z tym wprost dawne sformułowania II zasady termodynamiki stwierdzające niemożliwość zachodzenia procesów, których jedynym rezultatem byłoby przekazanie ciepła od ciała zimniejszego do cieplejszego (sformułowanie Clausiusa) lub całkowite zamienienie na pracę ciepła pobranego z układu o stałej temperaturze (sformułowanie Kelvina).

Jeśli otoczenie zachowuje w trakcie procesu stałą temperaturę T_0 (nazywa się je wówczas termostatem) i stałe ciśnienie p_0, to otrzymujemy prosty wzór na pracę maksymalną $W_{\max}(A, B) = T_0\Delta S - \Delta U$, obowiązuje on dla układów zamkniętych ($\Delta N = 0$). Wielkość $(T_0\Delta S - \Delta U - p_0\Delta V)$ jest to maksymalna praca użyteczna (z pominięciem pracy objętościowej $p_0\Delta V$). Postać tego wyrażenia pozwala w ciekawy sposób sformułować warunki stabilności stanów równowagi. Fizyczna treść tych warunków sprowadza się do stwierdzenia, że zaburzenie stanu równowagi wymaga wykonania pracy nad układem, co wyraża nierówność $(T_0\Delta S - \Delta U - p_0\Delta V) < 0$. Wynikają stąd podstawowe zasady ekstremalne termodynamiki.

zasady ekstremalne dla potencjałów termodynamicznych

Przy stałych S, V i N z odchyleniem od stanu równowagi układu wiąże się wzrost energii $\Delta U > 0$. Energia wewnętrzna osiąga więc w tych warunkach minimum w stanie równowagi. Zasada minimum energii wewnętrznej (przy ustalonej entropii) jest równoważna zasadzie maksimum entropii (przy ustalonej energii). Z kolei, porównując stany układu o tej samej temperaturze $T = T_0$ oraz V i N stwierdzamy, że odchyleniu od stanu równowagi odpowiada nierówność $\Delta(U - TS) > 0$. Zatem w stanie równowagi minimum osiąga potencjał Helmholtza (energia swobodna) $F = U - TS$. Podobnie przy stałych T, p, N obowiązuje zasada minimum dla potencjału Gibbsa (entalpia swobodna) $G = U - TS + pV$, zaś przy stałych p, V, N dla entalpii $H = U + pV$. Posługując się którąkolwiek z powyższych zasad minimum, można wyznaczyć wszystkie własności równowagowe układu termodynamicznego. Funkcje U, F, G, H nazywa się w związku z tym potencjałami termodynamicznymi (analogia z potencjałem w mechanice). Wszystkie wymienione zasady ekstremalne są zawarte w podstawowej nierówności $(\Delta U - T_0\Delta S + p_0\Delta V) > 0$. Gdy układ jest w równowadze, jego temperatura T i ciśnienie p przyjmują wartości odpowiednich parametrów otoczenia $T = T_0$, $p = p_0$. Zatem dla małych odchyleń entropii i objętości od wartości równowagowych jest spełniony wzór

$$\Delta U - T\Delta S + p\Delta V = \frac{1}{2}\left\{\left(\frac{\partial^2 U}{\partial S^2}\right)(\Delta S)^2 + \right.$$

$$\left. + 2\left(\frac{\partial^2 U}{\partial S \partial V}\right)\Delta S\Delta V + \left(\frac{\partial^2 U}{\partial V^2}\right)(\Delta V)^2\right\}.$$

Z prawej strony występuje tu różniczka drugiego rzędu energii wewnętrznej obliczona dla odchyleń ΔS i ΔV, różniczka pierwszego rzędu $(T\Delta S - p\Delta V)$ znosi się z wyrazem $(-T\Delta S + p\Delta V)$. Przy ustalonej objętości warunek $(\Delta U - T\Delta S + p\Delta V) > 0$ prowadzi więc do nierówności $(\partial^2 U/\partial S^2)_{V,N} > 0$. Ponieważ $(\partial U/\partial S)_{V,N} = T$, więc $(\partial^2 U/\partial S^2)_{V,N} = (\partial T/\partial S)_{V,N} = T[T(\partial S/\partial T)_{V,N}]^{-1} = T/C_V$.

stabilność stanu równowagi

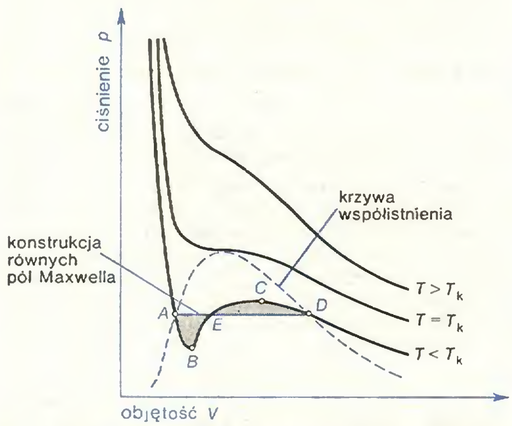
Stan równowagi jest zatem stabilny, gdy ciepło właściwe przy stałej objętości jest dodatnie $c_V > 0$, ($c_V = C_V/N$). W podobny sposób można wykazać, że warunkiem stabilności mechanicznej jest dodatnia wartość ściśliwości izotermicznej $K_T = -(\partial V/\partial p)_T V^{-1} > 0$. Reakcją układu na izotermiczny wzrost ciśnienia jest zmniejszenie objętości. W zakresie wartości parametrów stanu, w którym warunki stabilności nie mogą być spełnione, układ nie może występować w postaci fazy jednorodnej. Wniosek ten potwierdza do-

świadczenie. Równania stanu postaci $T = T(U, V, N)$, $p = p(U, V, N)$, $\mu = \mu(U, V, N)$, podobnie jak wynikające z nich równanie $p = p(T, V, N)$, muszą być zgodne z warunkami stabilności. Rozważmy przykładowo równanie stanu van der Waalsa

$$(p+aN^2/V^2)(V-bN) = NRT.$$

Wynika z niego, iż małej izotermicznej zmianie ciśnienia $dp|_T$ towarzyszy zmiana objętości $dV|_T$ zgodnie ze wzorem

$$dp|_T = d\big(NRT(V-bN)^{-1} - aN^2V^{-2}\big)|_T =$$
$$= \big(-NRT(V-bN)^{-2} + 2aN^2V^{-3}\big)\cdot dV|_T.$$



Rys. 7. Izotermy van der Waalsa. Do wyznaczenia punktów A i D izoterm poniżej temperatury krytycznej T_k i do otrzymania krzywej współistnienia stosuje się konstrukcję równych pól Maxwella (pole ABE = pole ECD). Wprowadza się przy tym część płaską izoterm (odcinek AD), której punkty reprezentują stabilne stany równowagi, odpowiadające współistnieniu pary i cieczy. Oryginalne izotermy van der Waalsa zawierają części o ujemnej ściśliwości izotermicznej (odcinek BC), co narusza warunek stabilności mechanicznej

Ściśliwość izotermiczna zatem ma postać

$$K_T = -1/V(\partial V/\partial p)_T =$$
$$= [NVRT(V-bN)^{-2} - 2aN^2V^{-2}]^{-1} =$$
$$= [nRT(1-bn^{-2}) - 2an^2]^{-1},$$

gdzie $n = N/V$ jest gęstością molową gazu. Nierówność $K_T > 0$ jest tu równoważna warunkowi $RT > 2an(1-bn)^2$. Maksymalna wartość funkcji $f(n) = 2an(1-bn)^2$ w przedziale $1/b > n \geqslant 0$ (zakres zmienności gęstości gazu) wynosi $8a/27b$. Jest ona osiągana w punkcie $n = 1/3b$, toteż dla dostatecznie niskich temperatur ($RT < 8a/27b$), izotermy zawierają części o $K_T < 0$ (naruszenie warunku stabilności mechanicznej, rys. 7). Interpretacja tego obszaru, jako odpowiadającego współistnieniu cieczy i pary (układ niejednorodny), stanowi podstawę teorii Maxwella–van der Waalsa przejścia fazowego gaz–ciecz.

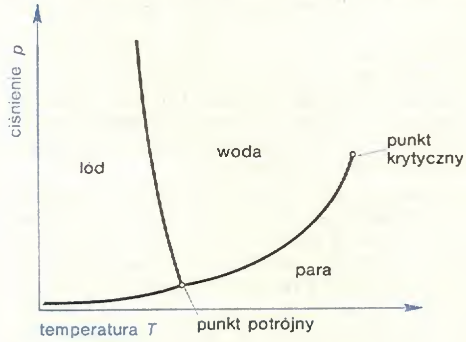
równowaga układów wielofazowych

Jeżeli układ r-składnikowy jest niejednorodny i rozpada się na α faz, równowaga termodynamiczna, poza stałością temperatury i ciśnienia w całym układzie wymaga dla każdego składnika równości jego potencjałów chemicznych we wszystkich fazach

$$\mu_j^1 = \ldots = \mu_j^\alpha = \mu_j, \quad j = 1, \ldots, r.$$

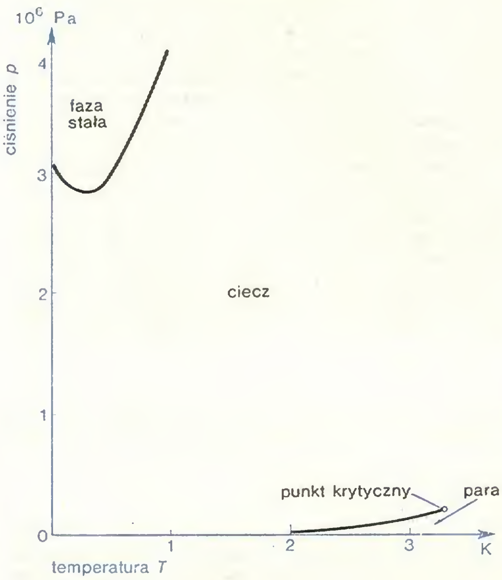
Po uwzględnieniu tych związków okazuje się, że liczba termodynamicznych stopni swobody (liczba parametrów intensywnych mogących się zmieniać niezależnie) wynosi $f = r+2-\alpha$. Ponieważ $f \geqslant 0$, więc w stanie równowagi może współistnieć co najwyżej $(r+2)$ faz (reguła faz Gibbsa). Dla układu jednoskładnikowego liczba ta wynosi 3 (rys. 8). Fakt, iż wówczas $f = 0$, oznacza, że stan taki występuje przy ściśle określonych wartościach parametrów intensywnych. Znajdujemy je rozwiązując równania $\mu^1(p,T) = \mu^2(p,T) = \mu^3(p,T)$, gdzie μ^1, μ^2, μ^3 oznaczają potencjały chemiczne trzech współistniejących faz. Uzyskujemy w ten sposób tzw. punkty potrójne danej substancji. Współistnieniu jedynie dwu faz odpowiada krzywa $\mu^1(p,T) = \mu^2(p,T)$. Zbiór krzywych współistnienia na płaszczyźnie (p,T) tworzy wykres fazowy, który, w zależności od właściwości danej substancji, może mieć bardzo różny przebieg. Niezwykła jest np. jego postać dla helu, mogącego występować w stanie ciekłym w dowolnie niskich temperaturach. Nie pojawia się tu bowiem w ogóle punkt potrójny faz gazowej, ciekłej i stałej (rys. 9), natomiast ciekły hel ^{4}He dla temperatur $T < 2{,}172$ K podlega przejściu fazowemu do stanu nadciekłego (→ Nadpłynność).

punkty potrójne

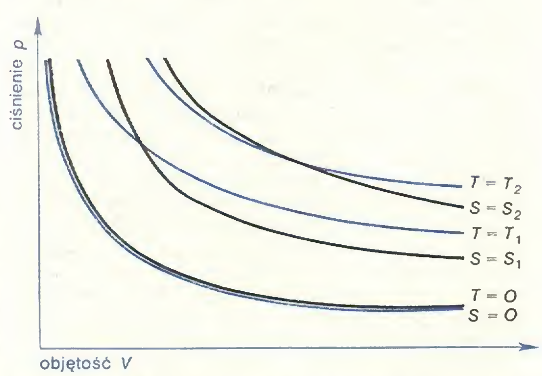


Rys. 8. Schematyczny wykres fazowy wody

wykresy fazowe



Rys. 9. Wykres fazowy helu ^{3}He



Rys. 10. Przebieg izoterm i adiabat w niskich temperaturach. Izoterma $T = 0$ pokrywa się z adiabatą $S = 0$. Zatem żaden proces adiabatyczny, rozpoczęty przy temperaturze $T > 0$, nie może doprowadzić do stanu o $T = 0$

III zasada termodynamiki

Podstawowe właściwości układów makroskopowych w niskich temperaturach ujmuje III zasada termodynamiki. W sformułowaniu Plancka stwierdza ona, że w stanach równowagi o zerowej temperaturze termodynamicznej wartość entropii wynosi zero (rys. 10). Wynika stąd, że zmiana entropii w dowolnym procesie izotermicznym dąży do zera, jeżeli temperatura, przy której proces przebiega, dąży do zera (twierdzenie Nernsta). III zasada pozwala jednoznacznie przypisać stanom równowagi układu określone wartości entropii (II zasada wyznacza entropię z dokładnością do stałej addytywnej). Wynika też z niej dążenie do zera w granicy $T \to 0$ ciepła właściwego c_V i c_p, współczynnika rozszerzalności termicznej $\alpha = (\partial V/\partial T)_p V^{-1}$ oraz współczynnika temperaturowego ciśnienia $(\partial p/\partial T)_V$. Wyjaśnienie tych faktów i w ogóle zjawisk występujących w niskich temperaturach, wymaga zastosowania teorii kwantowej. Interpretację fizyczną III zasady termodynamiki uzyskuje się na gruncie kwantowej fizyki statystycznej (→ Termodynamika statystyczna).

termodynamiczny opis procesów nieodwracalnych

Omówione trzy zasady termodynamiki są uniwersalnym narzędziem do badania właściwości układów makroskopowych w stanie równowagi. Wynika też z nich wiele istotnych wniosków na temat procesów termodynamicznych. Tak więc II zasada orzeka, iż są możliwe wyłącznie te procesy, które są zgodne z zasadą wzrostu entropii. W wypadku wyidealizowanych, kwazistatycznie przebiegających procesów odwracalnych stany pośrednie można uznać za stany równowagi, można więc je przedstawić w przestrzeni parametrów stanu (rys. 4). Zasady termodynamiki nie wystarczają jednak do opisu przebiegu rzeczywistych procesów nieodwracalnych. Dlatego zbudowana na ich podstawie teoria jest często nazywana termostatyką (nie występuje w niej w ogóle czas). Ażeby móc badać ewolucję w czasie stanu układu, rozważa się procesy, w czasie trwania których o każdej dostatecznie małej (ale makroskopowej) części układu można założyć, że jest w równowadze. Wtedy parametry intensywne, takie jak temperatura, ciśnienie czy potencjał chemiczny, mają w różnych częściach układu na ogół różne wartości, podlegające zmianom w czasie. Lokalnie, w określonej chwili czasu, parametry te wiążą się ze sobą przez równowagowe równania stanu. Okazuje się, że tego rodzaju założenie o lokalnej równowadze doskonale stosuje się do różnorodnych zjawisk (jest ono np. istotnym elementem hydrodynamiki). Określenie zmian w czasie lokalnych wartości parametrów makroskopowych wymaga sformułowania nowych praw. Jako przykład rozważymy tu proces przewodnictwa cieplnego w układzie, którego różne części mają w chwili początkowej różne temperatury. Z II zasady termodynamiki wynika, że przepływ energii wewnętrznej z obszarów o wyższej temperaturze do obszarów o niższej temperaturze doprowadzi układ do stanu równowagi, który charakteryzuje się stałą temperaturą w całej objętości układu. „Siłą termodynamiczną" powodującą przepływ energii (przewodnictwo cieplne) jest więc niejednorodność rozkładu temperatury. W punkcie $\vec{r}$ w chwili t miarą tej niejednorodności jest wektor grad $T(\vec{r}, t) = (\partial T/\partial x, \partial T/\partial y, \partial T/\partial z)$. Wektor ten znika, gdy temperatura nie zależy od $\vec{r}$. Przepływ energii w układzie można scharakteryzować za pomocą gęstości strumienia energii $\vec{J}_u(\vec{r}, t)$. Jest to wektor, którego długość określa ilość energii przepływająca w jednostce czasu przez jednostkową powierzchnię do niego prostopadłą. Doświadczalnie stwierdzono, że w procesach przewodnictwa cieplnego w ciałach izotropowych (przewodzących tak samo we wszystkich kierunkach) gęstość strumienia energii jest proporcjonalna do gradientu temperatury $\vec{J}_u(\vec{r}, t) = -\lambda \,\mathrm{grad}\, T(\vec{r}, t)$. Współczynnik λ jest zwany współczynnikiem przewodnictwa cieplnego, a prawo powyższe — prawem Fouriera. Jest ono przykładem liniowych praw fenomenologicznych, ustalających liniowe związki między gęstościami strumieni różnych wielkości fizycznych a odpowiednimi siłami termodynamicznymi. Powiązanie tych praw z prawami zachowania (energii, pędu, ładunku elektrycznego itp.) oraz z założeniem o lokalnej równowadze pozwala uzyskać pełny opis ewolucji w czasie stanu układu podlegającego procesowi nieodwracalnemu.

przewodnictwo cieplne

Sz. Szczeniowski, *Fizyka doświadczalna cz. II. Ciepło i fizyka cząsteczkowa*, Warszawa 1976; J. Werle, *Termodynamika fenomenologiczna*, Warszawa 1957; K. Zalewski, *Wykłady z termodynamiki fenomenologicznej i statystycznej*, Warszawa 1976.

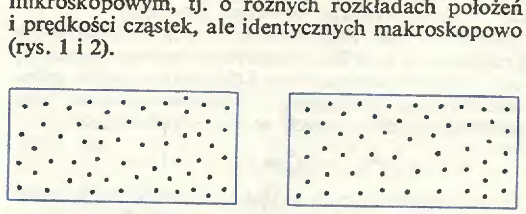
Termodynamika statyczna

Jerzy Czerwonko

Termodynamika statystyczna jest działem termodynamiki, który uzupełnia i uzasadnia termodynamikę fenomenologiczną. Na czym polega to uzupełnienie? Po pierwsze, na możliwości otrzymania, na podstawie założeń o budowie ciał, ich równania stanu, np. równania stanu gazu doskonałego, które do termodynamiki fenomenologicznej wchodziło jako prawo empiryczne. Po drugie, na wprowadzeniu fluktuacji, tj. odchyleń wielkości fizycznych od ich wartości średnich, rozpatrywanych w termodynamice fenomenologicznej. Skoncentrujemy się tu raczej na omówieniu termodynamiki statystycznej stanów równowagi, ze względu na to, że jej podstawy teoretyczne tworzą logiczną całość oraz fakt, że stanowi ona punkt wyjścia dla termodynamiki procesów nieodwracalnych. Natomiast uzasadnieniem termodynamiki statystycznej są prawa dynamiczne opisujące ruch cząstek, a więc równania Newtona lub równania Schrödingera dla układów niekwantowych bądź kwantowych przy rezygnacji z dokładnego opisu układu fizycznego.

uzasadnienie termodynamiki statystycznej

Aby zrozumieć o co chodzi w ostatnim zdaniu, zwróćmy uwagę, że termodynamika zajmuje się z reguły układami makroskopowymi, zawierającymi bardzo wiele atomów. Interesujemy się przy tym niewieloma parametrami charakteryzującymi układ, tj. takimi, które mają sens makroskopowy jak: objętość, energia, ciśnienie, polaryzacja dielektryczna, nie zaś położeniami czy prędkościami poszczególnych atomów. Dlatego też istnieje bardzo wiele układów fizycznych o tej samej strukturze, istotnie różnych w opisie mikroskopowym, tj. o różnych rozkładach położeń i prędkości cząstek, ale identycznych makroskopowo (rys. 1 i 2).

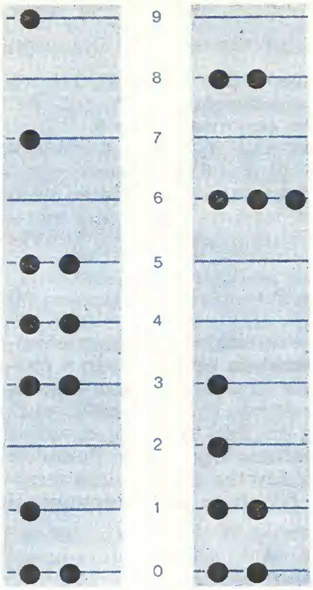


Rys. 1. Dwa różne układy 50 cząstek umieszczonych na tej samej powierzchni

zespół statystyczny

Pozwala to na wprowadzenie ważnego pojęcia zespołu statystycznego, który jest zbiorem wszystkich układów o identycznej budowie i identycznej wartości określonych parametrów makroskopowych. Parametry te i ich wartości służą jakby za szyld zespołu statystycznego, możemy mówić np. o zespole statystycznym 1 mola neonu (liczba atomów równa

liczbie Avogadra) znajdującym się w objętości 10 l. Ponieważ określony stan makroskopowy odpowiada wielu stanom mikroskopowym, więc sensowne staje



Rys. 2. Dwa różne układy 11 cząstek nie oddziałujących ze sobą, o tej samej energii całkowitej. Poziomy energetyczne mają energie od 0 do 9 jednostek, energia całkowita wynosi 41 jednostek. Liczba kółek na poziomach odpowiada liczbie cząstek o danej energii

się pytanie o prawdopodobieństwo występowania określonego stanu mikroskopowego w danym stanie makroskopowym. Do kwestii tej powrócimy jeszcze przy omawianiu konkretnych przykładów.

pomiar wielkości makroskopowych

Zobaczmy teraz, jak wygląda pomiar wielkości makroskopowych w warunkach równowagi — jako przykładowy parametr weźmy objętość układu. Cząstki układu znajdują się w bezustannym ruchu, tym samym zmienia się również bezustannie objętość układu, nawet układu znajdującego się w równowadze, która oznacza jedynie brak systematycznych zmian parametrów makroskopowych. Przyrządy stosowane do pomiaru parametrów makroskopowych mają określoną czułość i określoną bezwładność, tak że z reguły nie są w stanie śledzić chaotycznych zmian, którym podlegają parametry makroskopowe. Stąd, przyrządy pomiarowe podają nam nie chwilowe, ale uśrednione w czasie wartości parametrów makroskopowych, przy czym wzór na wartość średnią ma postać

$$\overline{A} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau A(t)\,dt, \tag{1}$$

gdzie $A(t)$ oznacza chwilową, a $\overline{A}$ — średnią wartość parametru A; czas $t = 0$ oznacza czas początku pomiaru. Granica $\tau \to \infty$ nie musi być traktowana zbyt dosłownie, chodzi tu wyłącznie o uśrednienie po okresie znacznie dłuższym od okresu charakterystycznego dla chaotycznych zmian parametru A.

wartość średnia w zespole statystycznym układów kwantowych

Jeśli wprowadzić prawdopodobieństwo w zespole statystycznym, to można również mówić o wartościach średnich w tym zespole. Pojęcie wartości średniej szczególnie łatwo wprowadzić dla układów opisywanych przez mechanikę kwantową. Niech zespół statystyczny stanowi zbiór stanów kwantowych M. Każdemu stanowi kwantowemu l należącemu do M ($l \in M$) przypisujemy prawdopodobieństwo P_l, przy czym

$$\sum_{l \in M} P_l = 1,$$

co oznacza, że prawdopodobieństwo znalezienia układu gdziekolwiek w zespole jest równe jedności. Równość ta jest prostą konsekwencją tego, że do zespołu statystycznego wchodzą wszystkie układy fizyczne o identycznej wartości określonych parametrów makroskopowych. W każdym stanie kwantowym $l \in M$ dowolna wielkość fizyczna A ma określoną wartość średnią $\overline{A_l}$. Występowanie tej wartości średniej nie jest związane z istnieniem zespołu statystycznego, lecz z kwantowym charakterem rozpatrywanego układu. (Jeśli wziąć pod uwagę np. paczkę fal odpowiadających jednej cząstce, to ze względu na rozmycie przestrzenne, położenie cząstki nie będzie ściśle określone, można będzie natomiast mówić o wartości średniej składowych wektora położenia cząstki). Zgodnie z powyższym, średnia wartość wielkości A w zespole statystycznym M wyniesie

$$\langle A \rangle = \sum_{l \in M} \overline{A_l} P_l. \tag{2}$$

wartość średnia w zespole statystycznym układów klasycznych

Nieco trudniej wprowadzić pojęcie średniej w zespole statystycznym układów opisywanych przez mechanikę klasyczną, mimo że mechanika klasyczna opisuje układy fizyczne zawsze w sposób mniej dokładny od mechaniki kwantowej. Stan układu klasycznego można określić znając wektory położeń $\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N$ i pędów $\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_N$ wszystkich N cząstek. Uwzględniając fakt, że wektor w przestrzeni K-wymiarowej to uporządkowany zbiór K liczb, można uznać, że stan układu klasycznego jest określony przez wektor w przestrzeni $6N$-wymiarowej, zwanej przestrzenią fazową lub przestrzenią układu N cząstek. Oznaczamy ten wektor przez $\vec{R}$, jego składowymi są współrzędne wektorów położenia i pędu wszystkich N cząstek w prostokątnym układzie współrzędnych, będziemy je wszystkie nazywali składowymi $R_k (k = 1, 2, \ldots, 6N)$ wektora $\vec{R}$. Zmienne R_k zmieniają się w sposób ciągły i można wprowadzić dla nich funkcję gęstości prawdopodobieństwa. Prawdopodobieństwo znalezienia układu w nieskończenie małej $6N$-wymiarowej kostce określonej przez nierówność $R_{k0} < R_k < R_{k0} + dR_k$, $k = 1, 2, \ldots, 6N$, będzie dane przez $f(\vec{R})\,d\Gamma$, gdzie $d\Gamma$ jest objętością owej kostki, a nieujemna funkcja $f(\vec{R})$ — gęstością prawdopodobieństwa. Wielkość $d\Gamma$ można zapisać jako iloczyn $dR_1 dR_2 \ldots dR_{6N}$, czyli $dx_1 dy_1 dz_1 \ldots dx_N dy_N dz_N dp_{x_1} dp_{y_1} dp_{z_1} \ldots dp_{x_N} dp_{y_N} dp_{z_N}$, gdzie $x_i, y_i, z_i; p_{x_i}, p_{y_i}, p_{z_i}$ oznaczają współrzędne wektorów położenia i pędu i-tej cząstki w prostokątnym układzie współrzędnych.

gęstość prawdopodobieństwa w przestrzeni gazowej

Aby funkcja $f(\vec{R})$ miała sens gęstości prawdopodobieństwa musi spełnić warunek

$$\int_\Gamma f(\vec{R})\,d\Gamma = 1,$$

w którym występuje całka $6N$-krotna po całej przestrzeni Γ. Wielkości makroskopowe, stanowiące przedmiot zainteresowania termodynamiki statystycznej, są zazwyczaj funkcjami wektora R w przestrzeni Γ. Na przykład, energię kinetyczną układu określa wzór

$$E_{\text{kin}} = \sum_{i=1}^{N} \frac{1}{2m} (p_{x_i}^2 + p_{y_i}^2 + p_{z_i}^2),$$

w którym m jest masą cząstek, przy czym energia całkowita to

$$E_{\text{kin}} + V(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N),$$

gdzie V jest energią potencjalną.

Teraz możemy już określić średnią wartość w zespole statystycznym układów klasycznych dowolnej wielkości fizycznej A będącej funkcją $\vec{R}$:

$$\langle A(\vec{R}) \rangle = \int A(\vec{R}) f(\vec{R})\,d\Gamma.$$

Opis układów kwantowych i klasycznych można formalnie ujednolicić, dzieląc przestrzeń fazową na komórki o równej objętości Ω, na tyle małe, aby zarówno gęstość prawdopodobieństwa, jak i rozpatrywane wielkości makroskopowe zmieniały się znikomo mało w granicach jednej komórki. Wtedy,

jeśli ponumerować komórki wskaźnikiem l i wybrać po jednym wektorze $\vec{R}_l$ z każdej komórki, to prawdopodobieństwo znalezienia układu w l-tej komórce będzie równe $P_l = f(\vec{R}_l)\,\Omega$, wartość średnia zaś będzie określona następująco

wartość średnia

$$\langle A(\vec{R})\rangle = \sum_{l\in M} A(\vec{R}_l)f(\vec{R}_l)\,\Omega = \sum_{l\in M} \bar{A}_l P_l.$$

Konkretyzacja pojęcia zespołu statystycznego

Poprzednie rozważania były z konieczności dość abstrakcyjne. Obecnie dostosujemy pojęcie zespołu statystycznego do warunków oddziaływania makroskopowego układu z otoczeniem. Zacznijmy od przypadku najprostszego — układu izolowanego od otoczenia i zajmującego określoną objętość V. Taki układ ma energię E ustaloną z dokładnością do błędu pomiaru δE, przy założeniu, że $\delta E \ll |E|$. Wskaźnik l oznacza tym razem różne stany kwantowe układu o energii $E < E_l < E+\delta E$, bądź też numeruje komórki przestrzeni fazowej, których energie spełniają taką nierówność. Zakładamy, że δE jest na tyle duże, iż w przedziale E, $E+\delta E$ znajduje się bardzo wiele stanów o energii E_l. Jest to łatwe do spełnienia przy podziale przestrzeni fazowej na odpowiednio małe komórki, albo przy odpowiednio dużych układach kwantowych, których poziomy energetyczne są prawie ciągłe. Sprawdźmy to na przykładzie pojedynczej cząstki zamkniętej w pudle sześciennym o krawędzi L, gdy pęd cząstki $\vec{p}$ spełnia warunki Borna-Kármána, tj. $\vec{p} = h\vec{n}/L$, gdzie h jest stałą Plancka, a $\vec{n}$ — wektorem o całkowitych współrzędnych. Dla dużych L siatka wektorów $\vec{p} = h\vec{n}/L$ staje się bardzo gęsta, podobnie jak i zbiór wartości energii $\mathcal{E}(\vec{p})$, jeśli tylko jest to funkcja ciągła. Podobnie będzie dla układu wielu cząstek, co szczególnie łatwo uzasadnić, jeśli cząstki ze sobą nie oddziałują i energia ich układu jest sumą energii poszczególnych cząstek.

prawdopodobieństwo znalezienia układu w l-tym stanie

Przyjmijmy chwilowo, że układ niekoniecznie znajduje się w równowadze; wtedy prawdopodobieństwo znalezienia go w l-tym stanie powinno zależeć od czasu, oznaczamy je $P_l(t)$. Zakładamy wtedy, że prawdopodobieństwo to spełnia zależność

$$P_l(t+\Delta) = \sum_{k\in M} P_k(t) A_{k\to l}(\Delta), \tag{3}$$

w której $A_{k\to l}(\Delta)$ oznacza prawdopodobieństwo przejścia ze stanu k do stanu l po upływie czasu Δ. Zależność ta oznacza, że jeśli układ znalazł się w stanie k w chwili t, to z prawdopodobieństwem $A_{k\to l}(\Delta)$ znajdzie się w chwili $t+\Delta$ w stanie l, że zaś w stanie k układ znalazł się z prawdopodobieństwem $P_k(t)$, więc pomnożenie $A_{k\to l}(\Delta)$ przez nie i zsumowanie takich wyrażeń dla wszystkich różnych wartości k daje pełne prawdopodobieństwo znalezienia układu w stanie l w chwili $t+\Delta$. Zależność ta ma więc sens bilansu prawdopodobieństwa, przy czym jedynym istotnym założeniem przyjętym przy jej wprowadzaniu jest założenie o niezależności zdarzeń polegających na znalezieniu się układu w stanie k i na jego przejściu z k do l, co prowadzi do mnożenia prawdopodobieństw. Prawdopodobieństwa $A_{k\to l}(\Delta)$ muszą spełniać związek

$$\sum_{l\in M} A_{k\to l}(\Delta) = 1 \tag{4}$$

oznaczający, że jeśli układ znalazł się w stanie k, to jest pewne, że znajdzie się w jakimkolwiek innym stanie po chwili Δ. Przeprowadzając sumowanie zależności (3) po wszystkich $l \in M$ i korzystając z równości (4), łatwo udowodnić, że jeżeli suma wszystkich prawdopodobieństw $P_l(t)$ była równa jedności, to będzie to spełnione automatycznie również w chwili $t+\Delta$. Podobnie, z równości (3) i (4), w prosty sposób wynika, że jeśli prawdopodobieństwa wszystkich stanów k są w pewnej chwili sobie równe (to znaczy niezależne od k), to będą również spełniały ten związek dowolnie długo, co oznacza stan równowagi. Przy stanie równowagi — z równości prawdopodobieństw wszystkich stanów $l \in M$ — otrzymujemy zatem

$$P_l = M^{-1}(E, \delta E), \tag{5}$$

gdzie M jest liczbą stanów (lub komórek), których energia (E_l) spełnia nierówność $E < E_l < E+\delta E$. Dla pozostałych stanów, oczywiście, $P_k = 0$.

Można wykazać, że jeśli $A_{k\to l}(\Delta) > 0$ dla dowolnej pary stanów k, l i $\Delta > 0$, co oznacza możliwość przejścia od każdego stanu k do każdego stanu l, to wyrażenie (5) jest jedynym wyrażeniem niezależnym od czasu spełniającym zależność (3). W takiej sytuacji mamy również

$$\lim_{t\to\infty} P_l(t) = P_l = M^{-1}(E, \delta E), \tag{6}$$

co oznacza, że przy $t\to\infty$ rozkład prawdopodobieństwa stanów dąży do rozkładu występującego w równowadze, i to niezależnie od tego, jakie prawdopodobieństwa $P_l(t')$ były w pewnej chwili początkowej t'. Równość (6) oznacza, że $P_l(t)$ różni się dowolnie mało od P_l po odpowiednio długim czasie. Powoduje to, że dowolna wielkość fizyczna A, uśredniona w czasie zgodnie ze wzorem (1), równa jest równowagowej średniej

$$\langle A\rangle = \frac{1}{M(E, \delta E)} \sum_l A_l, \tag{7}$$

przy czym sumowanie ogranicza się do takich l, że $E < E_l < E+\delta E$.

zespół mikrokanoniczny

W ten sposób określiliśmy zespół statystyczny układu izolowanego od otoczenia, dla którego wszystkie prawdopodobieństwa stanów kwantowych o energiach w podanym przedziale są równe. Zespół taki nazywa się zespołem mikrokanonicznym. Wykazaliśmy również, że średnia po czasie (1) jest równa średniej w takim zespole (7). Wniosek ten znany jest pod nazwą twierdzenia ergodycznego. Zostało ono wprowadzone przez L.E. Boltzmanna dla układów opisywanych przez mechanikę Newtona. Boltzmann podał wiele istotnych argumentów przemawiających za prawdziwością twierdzenia ergodycznego, ścisły zaś matematyczny dowód tego twierdzenia został podany przez G. Birkhoffa i J. von Neumanna dopiero w latach trzydziestych naszego wieku. Rozkład mikrokanoniczny nie wyróżnia ani stanów kwantowych ani też punktów w przestrzeni fazowej. Taki stan często określa się mianem „chaosu molekularnego". Dla układów klasycznych tor wektora $\vec{R}$ w przestrzeni fazowej przy $t\to\infty$ pokrywa wówczas powierzchnię ustalonej energii w sposób ciągły i równomierny.

twierdzenie ergodyczne

Jak już wyjaśniono, $\lim_{t\to\infty} P_l(t) = P_l$ niezależnie od początkowego rozkładu prawdopodobieństwa. Jest to stwierdzenie istnienia relaksacji, tj. procesu dochodzenia do równowagi. Zjawisko to ma nieodwracalny charakter, ponieważ stan równowagi raz osiągnięty trwać będzie nieograniczenie długo i nie ma możliwości powrotu do wyjściowego rozkładu prawdopodobieństw. Do faktu tego doszliśmy już jako do wniosku zależności (3) przy uwzględnieniu warunku (4). Zwróćmy uwagę, że wniosek ten jest prawdziwy nawet wtedy, gdy $A_{k\to l}(\Delta) = A_{l\to k}(\Delta)$, co oznacza, że prawdopodobieństwo przejścia ze stanu k do l jest równe prawdopodobieństwu przejścia z l do k, i co można by nazwać warunkiem odwracalności mikroskopowej.

Należy podkreślić, że nie sposób traktować przytoczonego wyprowadzenia równości wartości średnich po czasie i średnich w zespole mikrokanonicznym jako dowodu twierdzenia ergodycznego dla układów kwantowych. Zależność (3) została bowiem tylko przyjęta, nie zaś wyprowadzona na podstawie m.in. równania Schrödingera. Wyprowadzenie takie

wymaga z reguły dodatkowych założeń o układzie fizycznym. Dlatego też rozważania niniejsze mają, przynajmniej częściowo, charakter tylko wyjaśniający pojawianie się takich własności procesów fizycznych, jak relaksacja, nieodwracalność itp. Rozważania te pozwalają również zrozumieć statystyczny charakter podstaw termodynamiki.

Entropia i granica termodynamiczna

Na podstawie wyrażenia (7) można znaleźć wartości obserwowane tych wielkości fizycznych, które mają określone wartości średnie bądź w stanie kwantowym, bądź też dla funkcji $\vec{R}$, tj. $\vec{r_i}$ oraz $\vec{p_i}$ w układzie klasycznym. Do wielkości tego rodzaju nie należy ani wymiana ciepła ΔQ, ani też entropia, zdefiniowana w termodynamice przez równanie $\Delta S = \Delta Q/T$, gdzie T jest temperaturą w skali Kelvina. Podana definicja jest słuszna przy nieskończenie małych różnicach ΔQ i ΔS. Wymianę ciepła określa się jako różnicę energii stanu końcowego i początkowego układu, jeśli nie wykonano nad nim pracy. Taka sama różnica energii może powstać w wyniku jednoczesnej wymiany ciepła z otoczeniem i wykonania pracy nad układem. Dlatego można mówić raczej o wymianie ciepła niż o cieple i nie można wyrazić tej wielkości w funkcji $\vec{R}$ ani też przypisać jej wartości stanowi kwantowemu. Z tego także powodu nie można wyrazić entropii jako wartości średniej pewnej funkcji od $\vec{R}$ lub też funkcji wskaźnika stanu kwantowego, tylko jest konieczne podanie osobnej statystycznej definicji entropii. Zostało to dokonane przez Boltzmanna.

granica termodynamiczna

Ograniczymy się teraz do wprowadzenia entropii dla układów makroskopowych, a nawet tylko dla układów w tzw. granicy termodynamicznej, tj. przy objętości V i liczbie N dążących do nieskończoności tak, aby iloraz N/V (gęstość cząstek) pozostawał skończony i niezerowy. Założymy również, że istotne w opisie układu makroskopowego wielkości dzielą się na dwie klasy: wielkości ekstensywne i wielkości intensywne.

Pierwsze z nich, przy wzroście liczby cząstek dużego układu, rosną wraz z N prawie liniowo, drugie — prawie nie zmieniają swojej wartości. Uściślając to matematycznie, można uznać, że wielkości ekstensywne są rozbieżne przy $N \to \infty$ tak, że dla dowolnej takiej wielkości E, granica E/N istnieje i jest niezerowa. Wielkości intensywne są natomiast przy $N \to \infty$ zbieżne do liczby. Do wielkości ekstensywnych należą np.: energia, objętość i ładunek elektryczny, do intensywnych zaś: ciśnienie, temperatura, stała dielektryczna itd. W granicy termodynamicznej — podanie jednej wielkości ekstensywnej i wielu wielkości intensywnych pozwala określić każdą inną wielkość ekstensywną, np. zamiast wielkości ekstensywnych wymienionych wyżej można podać: objętość, gęstość energii (E/V) oraz gęstość ładunku. Przyjęcie układów w granicy termodynamicznej prowadzi do rezygnacji z rozpatrywania efektów powierzchniowych. Dlatego też, w granicy termodynamicznej, wielkości ekstensywne stają się makroskopowo addytywne, co oznacza, że np. energia ciała makroskopowego jest równa sumie energii jego makroskopowych części. Wniosek ten, oczywisty w odniesieniu do ilości cząstek, ładunku, czy objętości, dla energii wymaga pewnego uzasadnienia. Jeśli w układzie działają siły o skończonym zasięgu rzędu n odległości międzycząsteczkowych, to energia oddziaływania makroskopowych podukładów kontaktujących się przez powierzchnię powinna być rozbieżna wraz z N jak $nN^{2/3}$, nie zaś jak N. Poprawka do energii, a więc wielkości ekstensywnej, rzędu $nN^{2/3}$ jest w granicy termodynamicznej pomijalna, co uzasadnia makroskopową addytywność energii. Warto dodać, że ekranowanie ładunku przez elektrony w ośrodkach ciągłych sprawia, że siły Coulomba, działające między elektronami i jonami, stają się również siłami o skończonym zasięgu.

wielkości ekstensywne i intensywne

Dodajmy, że wnioski, do których poprzednio doszliśmy, dotyczące nieodwracalności i dochodzenia do równowagi są słuszne jedynie w granicy termodynamicznej. W dużych ale skończonych układach mamy do czynienia z ruchem zbliżonym do okresowego, o niezwykle dużych okresach. Dlatego też, przy doświadczeniach np. nad molem gazu, trwających dowolnie długo w ludzkiej skali czasu, możemy śmiało potraktować procesy tam zachodzące jako procesy nieodwracalne. Na kłopoty związane z przybliżoną okresowością ruchu układów skończonych zwracano uwagę już w dziewiętnastym stuleciu.

entropia

Ponieważ wymiana ciepła jest określona przez różnicę energii, więc jest również wielkością ekstensywną. Stąd, jak i z definicji — entropia jest również wielkością ekstensywną. Entropia układów opisywanych przez mechanikę kwantową będzie zdefiniowana następująco:

$$S = k \ln M(E, \delta E), \tag{8}$$

gdzie k jest stałą Boltzmanna, $M(E, \delta E)$ zaś zostało określone wzorem (5). Wzór (8) jest jedynie definicją entropii. Później uzasadnimy, że „entropia statystyczna" (8) ma wszelkie własności entropii termodynamicznej. W podanym wcześniej termodynamicznym określeniu entropii, $\Delta S = \Delta Q/T$, występuje wyłącznie różnica entropii, a więc można ją określić jedynie z dokładnością do stałej; statystyczna definicja określa entropię bez jej dowolności. Z drugiej strony, niepokoi zależność S nie tylko od E i innych parametrów, od których M zależy w sposób niejawny, ale również od δE. Można wykazać, że zależność ta jest nieistotna w granicy termodynamicznej.

Jak można zdefiniować entropię dla układów opisywanych przez mechanikę klasyczną? Objętości przestrzeni fazowej obszaru $E < E(\vec{R}) < E + \delta E$, gdzie $E(\vec{R})$ jest energią układu klasycznego, nie można traktować jako odpowiednika $M(E, \delta E)$, ponieważ jest to wielkość wymiarowa i jej logarytm jest nieokreślony poza wybranym układem jednostek. Ale jeśli tę objętość fazową podzielić przez h^{3N}, gdzie h jest stałą Plancka, to otrzymamy wielkość bezwymiarową.

paradoks Gibbsa

Gibbs zauważył ponadto, że jeśli objętości fazowej nie podzielić przez $N!$, to S będzie zależała od N silniej niż liniowo przy $N \to \infty$, a więc nie będzie wielkością ekstensywną. Co gorsze, powstaje przy tym niemożność jednoznacznego określenia entropii. Wykażemy to na podstawie następującego rozumowania: wyobraźmy sobie dwa naczynia, wypełnione identycznym gazem, o tym samym ciśnieniu i temperaturze. Niech S_1 i S_2 będą entropią gazu znajdującego się w pierwszym i drugim naczyniu. Entropia S układu naczyń połączonych ze sobą, równa jest $S_1 + S_2$ ze względu na odwracalność mieszania gazu w takich warunkach. Jeśli S_1 i S_2 są ekstensywne, to $S_1 = N_1 s_0$, $S_2 = N_2 s_0$, gdzie wielkość s_0 oznacza entropię na jedną cząstkę i jest intensywna. Wtedy $S = (N_1 + N_2) s_0 = N s_0$ i entropia naczyń połączonych zależy jedynie od stanu układu. Jeśli jednak $S_{1,2} \neq N_{1,2} s_0$, to do określenia S jest jeszcze potrzebne podanie N_1 lub N_2, a więc historii układu.

Dlatego też entropia połączonego układu zależałaby od miejsca, z którego wyciągnięto przegródkę



Rys. 3. Entropia połączonego układu zależy od miejsca, z którego wyciągnięto przegródkę A oddzielającą jednakowe i znajdujące się w jednakowych warunkach gazy w naczyniach 1 i 2

A (rys. 3) oddzielającą dwa jednakowe i znajdujące się w jednakowych warunkach gazy w naczyniach *1* i *2*.

liczba stanów w warstwie energetycznej

Z niniejszych rozważań wynika, że właściwa postać wyrażenia na liczbę stanów w warstwie energetycznej o grubości δE ma postać

$$M(E, \delta E) = \frac{1}{N!} \int \frac{d\Gamma}{h^{3N}}, \quad E < E(\vec{R}) < E + \delta E. \qquad (9)$$

Wyrażenie to można również otrzymać przez przejście graniczne od mechaniki kwantowej do klasycznej. Ponieważ $N!$ to liczba wszystkich możliwych permutacji N elementów, więc podzielenie przez $N!$ oznacza odpowiednie zredukowanie liczby stanów z uwagi na to, że permutacja położeń i pędów cząstek jednakowych nie powoduje zmiany sytuacji fizycznej, co stanowi jakby klasyczną zapowiedź nierozróżnialności jednakowych cząstek w mechanice kwantowej.

Pokażemy teraz, jak znaleźć własności termodynamiczne układu, jeśli znamy S w funkcji E i V. Załóżmy dla uproszczenia, że nad układem można wykonywać jedynie pracę objętościową. Jeśli więc objętość V będzie stała, to praca nie będzie wykonywana i $\Delta Q = \Delta E$. Stąd oraz z termodynamicznej definicji entropii wynika, że $(\partial S/\partial E) = 1/T$, co pozwala wyrazić energię (a tym samym — entropię) przez objętość i temperaturę. Ponadto, jeśli E jest stałe, to praca objętościowa układu $p\Delta V$, gdzie p — ciśnienie, równa się zmianie ciepła. Powstaje to stąd, że $(\partial S/\partial V) = p/T$ i jeśli podstawić tu E w funkcji objętości i temperatury, to otrzymamy zależność typu $p = f(V, T)$, a więc równanie stanu, co byłoby niemożliwe bez użycia metod termodynamiki statystycznej.

prawo wzrostu entropii

Statystyczna definicja entropii (8) pozwala łatwo uzasadnić prawo jej wzrostu. Weźmy w tym celu dwa układy o energiach E_1 i E_2 i liczbie stanów $M_1(E_1, \delta E)$ oraz $M_2(E_2, \delta E)$. Jeśli układy te nie mogą wymieniać między sobą energii, to liczba stanów całego układu będzie po prostu iloczynem $M_1 M_2$, bowiem stany każdego z podukładów określają stan układu. Natomiast jeżeli podukłady mogą wymieniać energię między sobą, to aby otrzymać liczbę stanów całego układu, trzeba jeszcze wysumować iloczyn $M_1(E_1 - m\delta E, \delta E)\, M_2(E_2 + m\delta E, \delta E)$ po wszystkich całkowitych k takich, że $E_1 - m\delta E$ oraz $E_2 + m\delta E$ są dopuszczalnymi energiami obydwu podukładów. Zauważmy, że w podanym wyżej iloczynie suma energii podukładów musi być równa $E_1 + E_2$, tzn. uwzględniono tu prawo zachowania energii. Gdy podany iloczyn osiąga maksimum dla $m = 0$, to w granicy termodynamicznej można wykazać, że entropia układu będzie równa sumie entropii podukładów. Jeśli jednak maksimum tego iloczynu jest osiągnięte przy $m \neq 0$, to wspomniana suma iloczynów jest większa od $M_1(E_1, \delta E)\, M_2(E_2, \delta E)$, co oznacza wzrost entropii połączonego układu. Nietrudno również zauważyć, że warunek na maksimum iloczynu jest równoważny warunkowi równości temperatury w obydwu podukładach.

Należy przy tym podkreślić nieco inny charakter prawa wzrostu entropii w termodynamice statystycznej niż w fenomenologicznej. Polega to na możliwości nawet przejściowego zmniejszania się entropii w procesie wymiany energii między podukładami, ponieważ $M_1(E_1 - m\delta E, \delta E)\, M_2(E_2 + m\delta E, \delta E)$ może być mniejsze od $M_1(E_1, \delta E)\, M_2(E_2, \delta E)$ przy odpowiednim m. Zmiana energii właśnie o $m\delta E$ jest — dla dużych układów — wysoce nieprawdopodobna, nie jest jednak całkowicie wykluczona. Widzimy więc, że prawo wzrostu entropii traci, przy statystycznym podejściu do termodynamiki, swój absolutny charakter. Ujmując rzecz ogólnie, podejście statystyczne charakteryzuje występowanie fluktuacji — odchyleń parametrów fizycznych od ich wartości średnich $A(t) - \bar{A}$. Miarą fluktuacji nie może być wartość średnia $\bar{A}(t) - A$, wielkość ta bowiem znika; za miarę tę przyjmuje się pierwiastek ze średniego kwadratu $(A(t) - \bar{A})$, tj. zgodnie z (1)

$$\sqrt{\lim_{\tau\to\infty} \frac{1}{\tau} \int_0^{\tau} (A(t) - \bar{A})^2 dt},$$

czyli średnie odchylenie kwadratowe wielkości $A(t)$. Ze względu na równość średniej w czasie i średniej w zespole statystycznym, miarę fluktuacji można wyrazić również w postaci średniej w zespole statystycznym. Doświadczalne obserwacje fluktuacji wielkości fizycznych potwierdziły statystyczny charakter termodynamiki. Wielki wkład do tego zagadnienia wniósł polski fizyk Marian Smoluchowski.

zespół kanoniczny i makrokanoniczny

Z wyrażeń na entropię oraz prawdopodobieństwo w zespole mikrokanonicznym można, w stosunkowo prosty sposób, otrzymać prawdopodobieństwa stanów makroskopowego podukładu należącego do znacznie większego od niego układu i wymieniającego z nim energię. Zespół statystyczny opisujący taki podukład nazywa się zespołem kanonicznym lub zespołem kanonicznym Gibbsa. Funkcja prawdopodobieństwa dla podukładu kwantowego ma postać

$$e^{-E_l/kT}/Z, \qquad (10)$$

suma stanów

gdzie Z otrzymuje się z warunku, że prawdopodobieństwo zsumowane po l daje jedność, czyli

$$Z = \sum_l e^{-E_l/kT}.$$

Wielkość Z nazywa się sumą stanów. Znalezienie jej, co jest z reguły łatwiejsze niż określenie $M(E, \delta E)$, pozwala obliczyć wszystkie funkcje termodynamiczne układu. Jeśli zamiast E_l podstawić $E(\vec{R})$, a zamiast sumy po l — całkę po $d\Gamma$, to otrzymamy odpowiednio gęstość prawdopodobieństwa oraz sumę stanów dla układu klasycznego pomnożoną przez $h^{3N}N!$. Równie często używa się w termodynamice statystycznej zespołu statystycznego podukładu makroskopowego wymieniającego zarówno energię, jak i cząstki, ze znacznie większą od siebie resztą układu. Zespół taki nazywany jest zespołem makrokanonicznym.

Ostatnie zdania wydają się sugerować, że istnieje osobna termodynamika „mikro" i „makro" oraz po prostu „kanoniczna". Okazuje się jednak, że niezależnie od ustalenia energii, liczby cząstek czy też możliwości ich wymiany, wielkości te są i tak praktycznie stałe, ponieważ ich średnie odchylenia kwadratowe są rzędu $N^{1/2}$, dzięki czemu względne odchylenia energii czy liczby cząstek są rzędu $N^{-1/2}$, co jest pomijalne w granicy termodynamicznej.

III zasada termodynamiki

Z postaci wzoru (10) widać, że przy $T \to 0$ prawdopodobieństwo stanu o najniższej energii dąży do jedności, pozostałe zaś prawdopodobieństwa dążą do zera. Zatem przy $T = 0$ układ będzie znajdował się w stanie podstawowym. Stan ten jest bądź niezwyrodniały, bądź też jego zwyrodnienie jest skończone przy $N \to \infty$. Powoduje to, że przy $T = 0$ entropia jest zerowa w granicy termodynamicznej, co stanowi treść trzeciej zasady termodynamiki. Była ona uzasadniana początkowo wyłącznie empirycznie, widzimy jednak, że wynika ona w nie dający się zaprzeczyć sposób z definicji entropii (8) i kwantowości układu.

Gazy klasyczne i kwantowe

Po raczej abstrakcyjnych rozważaniach przejdźmy do zastosowań, czyli do teorii gazów doskonałych — klasycznych i kwantowych. Przez gaz doskonały rozumie się układ, którego energię całkowitą można zapisać w postaci

$$E = \sum_l E_l N_l,$$

gdzie N_l jest liczbą cząstek w stanie o energii E_l. Stan o energii E_l wygodniej będzie potraktować nie

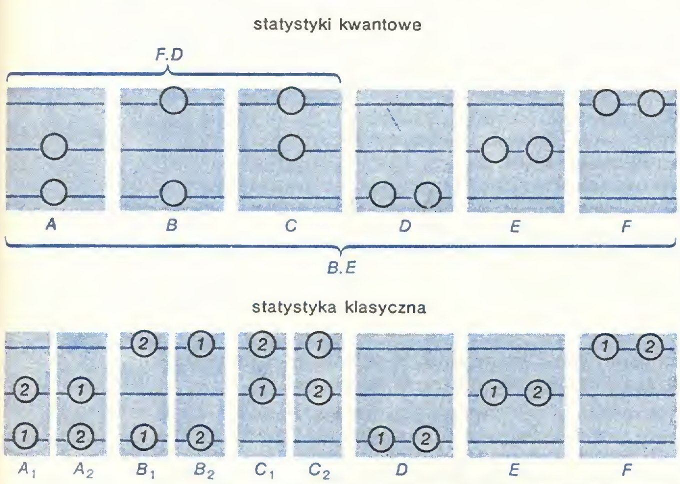
jako oddzielny stan kwantowy, ale jako zbiór liczby stanów wielkości ekstensywnej o bliskich energiach. E_l zaś będzie oznaczać pewną energię średnią. Dla odpowiednio dużych układów kwantowych bądź też dla układów klasycznych, jeśli tylko przedział energii δE wokół E_l jest wielkością — tym razem — intensywną, liczba stanów jest wielkością ekstensywną. Nałóżmy na pęd warunki Borna–Kármána (założenie przyjmowane zwykle w teorii ciała stałego), tj. przyjmijmy, że pęd spełnia warunki $\vec{p} = h\vec{n}/L$, gdzie $\vec{n}$ jest wektorem o współrzędnych równych liczbom całkowitym, a L — długością krawędzi sześcianu zajmowanego przez układ. W tym wypadku różnym stanom kwantowym odpowiadają różne wektory $\vec{n}$, czemu z kolei odpowiada sieć regularna prosta w przestrzeni pędów o długości krawędzi h/L. Dlatego właśnie gęstość stanów kwantowych w przestrzeni pędu wynosi V/h^3. Zatem pomnożenie objętości obszaru przestrzeni pędów przez V/h^3 daje liczbę stanów w tym obszarze, przy czym będzie to wielkość ekstensywna. Taki sam rezultat daje wyrażenie (9) na liczbę stanów układu klasycznego przy $N = 1$, gdyż występująca tam całka po zmiennych przestrzennych daje czynnik V. Widzimy więc, że podzielenie klasycznej objętości fazowej przez h^{3N}, zastosowane ze względów wymiarowych, ma znacznie głębsze uzasadnienie.

Oznaczmy liczbę stanów o średniej energii E_l (wielkość ekstensywna) przez g_l. Zastanówmy się, na ile sposobów można rozmieścić N_l cząstek w g_l stanach. W tym celu trzeba ustalić, które rozmieszczenia uznajemy za różne i jakie dodatkowe własności muszą spełniać rozmieszczenia. W mechanice klasycznej cząstki jednakowe są rozróżnialne, ponieważ można prześledzić tor każdej z nich, nie wpływając na ruch cząstki. Nie ma przy tym żadnych ograniczeń na liczbę cząstek w poszczególnych stanach. W takim wypadku będziemy mówili, że cząstki podlegają statystyce Maxwella–Boltzmanna (M.B.). W mechanice kwantowej cząstki są nierozróżnialne, ale możliwe są dwa przypadki: albo nie ma ograniczeń na liczbę cząstek w danym stanie kwantowym, albo też liczba cząstek w danym stanie kwantowym nie może przewyższać jedności. W pierwszym z nich mówimy o statystyce Bosego–Einsteina (B.E.) lub po prostu Bosego, a cząstki nazywamy bozonami, w drugim — o statystyce Fermiego–Diraca (F.D.) lub po prostu Fermiego, a cząstki nazywamy fermionami. Aby zrozumieć różnicę między tymi trzema możliwościami, popatrzmy na rys. 4, gdzie przedstawiono rozmieszczenie dwóch cząstek w trzech stanach. Dla cząstek podlegających statastyce F.D. możliwe są tylko trzy stany A, B, C, dla cząstek podlegających

statystyka Maxwella–Boltzmanna

statystyka Bosego–Einsteina

statystyka Fermiego–Diraca



Rys. 4. Możliwe rozmieszczenia dwóch cząstek w trzech stanach zależą od statystyki, której te cząsteczki podlegają

statystyce B.E. — wszystkie stany A–F. Jak widzimy, każdemu ze stanów A, B, C odpowiadają po dwa stany układu klasycznego. Razem mamy 3 różne rozmieszczenia w wypadku statystyki F.D., 6 w wypadku — B.E. i 9 w razie — M.B. Ogólne wyrażenia na liczbę W_l rozmieszczeń cząstek w zależności od g_l i N_l mają postać:

$$W_l = \begin{cases} \dfrac{g_l!}{N_l!(g_l-N_l)!} & (\text{dla } g_l \geqslant N_l) - \text{F.D.} \\ \dfrac{(g_l+N_l-1)!}{N_l!(g_l-1)!} & - \text{B.E.} \\ g_l^{N_l} & - \text{M.B.} \end{cases}$$

Wyprowadzenie tych wzorów można znaleźć w każdej niemalże książce poświęconej kombinatoryce czy rachunkowi prawdopodobieństwa, dlatego też pozwolimy je sobie opuścić.

Ogólna liczba sposobów rozmieszczenia cząstek nierozróżnialnych jest równa iloczynowi wszystkich W_l dla wszystkich różnych wartości l, natomiast dla cząstek rozróżnialnych trzeba taki iloczyn jeszcze pomnożyć przez czynnik $N!/N_1!N_2!\ldots N_l!\ldots$ ($N = N_1+N_2,\ldots$), określający na ile sposobów można wydzielić z N rozróżnialnych cząstek grupy po $N_1, N_2, \ldots, N_l \ldots$ cząstek. Zgodnie z tym co zauważył Gibbs, rezultat dla cząstek rozróżnialnych należy poza tym podzielić przez $N!$. Zatem wzory na liczbę możliwych stanów będą miały postać:

liczba możliwych stanów

$$M = \begin{cases} \prod\limits_l \dfrac{g_l!}{N_l!(g_l-N_l)!} & - \text{F.D.} \\ \prod\limits_l \dfrac{(g_l+N_l-1)!}{N_l!(g_l-1)!} & - \text{B.E.} \\ \prod\limits_l \dfrac{g_l^{N_l}}{N_l!} & - \text{M.B.} \end{cases}$$

przy spełnieniu warunku, że $\sum\limits_l E_l N_l = E$ oraz że $\sum\limits_l N_l = N$. Wielkość M oznacza liczbę stanów przy określonych liczbach cząstek $N_1, N_2, \ldots, N_l \ldots$. Aby otrzymać całkowitą liczbę stanów trzeba zsumować takie wyrażenia na M po wszystkich możliwych N_l spełniających dwa wypisane warunki. Skorzystamy teraz ze wzoru Stirlinga, który ma postać:

wzór Stirlinga

$$\ln(x!) \approx x(\ln x - 1) \text{ przy } x \to \infty,$$

co oznacza, że iloraz obydwu stron dąży do jedności przy $x \to \infty$. Związek taki zachodzi dla ekstensywnych N_l i g_l ($N_l, g_l \to \infty$): ze wzoru Stirlinga zastosowanego do M podanych wyżej wynika, że $\ln M$ jest również wielkością ekstensywną. Aby to udowodnić zauważmy, że nawet gdy M' jest równe M to $\ln(M+M') = \ln M + \ln 2 \approx \ln M$.

Warunek na maksimum M (równoważny warunkowi na maksimum $\ln M$) przy ustalonej energii i całkowitej liczbie cząstek można zapisać jako znikanie różniczki (stosujemy tutaj metodę mnożników Lagrange'a):

$$d\Big(\ln M - \beta \sum_l E_l N_l + \beta\mu \sum_l N_l\Big) = 0,$$

gdzie zmienne N_l można traktować obecnie jako niezależne oraz ze względu na ich ekstensywność — ciągłe. Wielkość $\ln M$ osiąga maksimum przy $N_l = \overline{N}_l$, gdzie

$$\overline{N}_l = \frac{g_l}{e^{-\beta\mu}e^{\beta E_l} \pm 1}, \begin{pmatrix}\text{F.D.}\\ \text{B.E.}\end{pmatrix},$$

$$\overline{N}_l = g_l e^{\beta\mu} e^{-\beta E_l}, \text{ M.B.}$$

Stałe μ i β należy określić z warunków

$$\sum_l E_l \bar{N}_l = E, \quad \sum_l \bar{N}_l = N.$$

Pomnożenie $\ln M$ przez stałą Boltzmanna i podstawienie $\bar{N}_l$ zamiast N_l daje wartość entropii. Jeśli posłużyć się układem otrzymanych równości i obliczyć $(\partial S/\partial E)$ przy stałym N, to dostaniemy $k\beta$, stąd $\beta = 1/kT$.

zapełnienie jednego poziomu

Wielkości $\bar{N}_l/g_l$ mają sens najbardziej prawdopodobnego i, ze względu na makroskopowość $\bar{N}_l$ oraz wielkość ich fluktuacji $\sim\sqrt{\bar{N}_l}$, również średniego zapełnienia jednego poziomu. Na to, ażeby N_l dla statystyki B.E. były dodatnie, jeśli — jak zazwyczaj — przyjąć, że najniższa wartość $E_l = 0$, potrzeba aby $e^{-\beta\mu} \leqslant 1$, tj. $\mu \leqslant 0$. Jeśli nie obowiązuje prawo zachowania liczby bozonów, jak np. dla fotonów, które podlegają emisji i absorpcji, to warunku $\sum_l N_l = N$ nie trzeba wprowadzać, dzięki czemu $\mu = 0$. Podobnie jak można się przekonać, anharmoniczność drgań sieci prowadzi do niezachowania liczby fononów, dzięki czemu dla nich również $\mu = 0$. Jeśli $e^{-\beta\mu} e^{\beta E_l} \gg 1$, co odpowiada wysokim temperaturom i małym gęstościom, to $\bar{N}_l$ dla obydwu statystyk kwantowych dążą do wartości klasycznej statystyki M.B. Nie jest to bynajmniej własność wyróżniająca gazy doskonałe, bowiem wszystkie ciała makroskopowe wykazują odstępstwa od klasycznego opisu w granicy niskich temperatur i dużych gęstości oraz zachowują się klasycznie w sytuacji odwrotnej. Granica ta, dla różnych zjawisk i ciał, może odpowiadać różnym zakresom temperatur i gęstości.

Należy odróżnić rozkład kanoniczny Gibbsa, w którym E_l jest ekstensywną energią układu, od rozkładów F.D., B.E. i M.B., w których E_l jest intensywną energią jednej cząstki. Dzięki temu nie ma sprzeczności między rozkładami F.D. i B.E. a rozkładem kanonicznym. Jeśli zamiast E_l podstawić $p^2/2m$, uwzględnić fakt, że

$$\sum_l g_l f(E_l) = \frac{V}{h^3} \int d^3p f(p^2/2m)$$

i obliczyć wielkość $e^{\beta\mu}$ dla rozkładu M.B., to można otrzymać rozkład Maxwella w postaci

$$f(p^2/2m) = \frac{N}{V} h^3 \left(\frac{\pi}{2mkT}\right)^{3/2} e^{-p^2/2mkT}.$$

Ponadto, jeśli dla rozkładu M.B. $\ln M$ jest wielkością ekstensywną, to $\ln(N!M)$, a więc przed podzieleniem przez $N!$, zawiera składnik $N(\ln N - 1)$ i wielkością ekstensywną nie jest. Fakt ten uprawomocnia paradoks Gibbsa, rozważany poprzednio.

Aktualne problemy termodynamiki statystycznej

Trudno byłoby stwierdzić, że udało się nam tutaj przedstawić współczesne oblicze termodynamiki statystycznej, zajęliśmy się raczej zapoznaniem Czytelnika z niezbędnym arsenałem środków najpowszechniej w niej używanych. Przejdziemy teraz do zagadnień bardziej aktualnych. Ograniczymy się na razie do termodynamiki statystycznej stanów równowagi. Do otrzymania równowagowych funkcji termodynamicznych wystarcza obliczenie sumy stanów Z lub też $M(E, \delta E)$. Niestety, dla układów cząstek oddziałujących jest to, ogólnie biorąc, niemożliwe. Dlatego też oblicza się te wielkości w sposób przybliżony dla pewnego zakresu temperatur lub gęstości bądź też dla pewnych szczególnych modeli oddziaływania. Rozwinięto ponadto metody pozwalające obliczyć funkcje termodynamiczne, gdy oddziaływanie międzycząstkowe jest słabe.

Omówimy pokrótce przybliżone metody obliczania funkcji termodynamicznych. Dla niskich temperatur, a więc w układzie słabo wzbudzonym, jego wzbudzenia można opisać w języku kwazicząstek (→ Wzbudzenia elementarne w ciałach stałych). Kwazicząstki te podlegają określonej statystyce kwantowej — B.E. lub F.D. Stąd obliczenie pewnych funkcji termodynamicznych w granicy niskotemperaturowej daje się sprowadzić do obliczenia analogicznych funkcji dla gazu kwazicząstek. Dlatego właśnie gaz kwantowy doskonały ma tak doniosłe znaczenie w wyjaśnianiu niskotemperaturowych własności układów z silnym oddziaływaniem międzycząstkowym. Z kolei, przy dostatecznie wysokich temperaturach, kwantowość przestaje odgrywać istotną rolę, co pozwala znacznie uprościć obliczenie funkcji termodynamicznych. Jeśli przyjąć normalną postać energii, tj. $E(\vec{R}) = E_{\text{kin}} + U$, to wykonanie w sumie stanów Z (zob. 10) całki po pędach staje się trywialne i obliczenie Z sprowadza się do obliczenia całki z funkcji $e^{-V/kT}$ po współrzędnych położenia cząstek. Energia potencjalna gazów V jest z dobrą dokładnością sumą energii potencjalnej oddziaływania wszystkich par cząstek. Dla takich V znaleziono bardzo efektywne metody obliczania Z, zwane metodami rozwinięć grupowych. Nie pozwalają one wprawdzie obliczyć Z dowolnej funkcji oddziaływania dwóch cząstek, ale umożliwiają, poza dowodem pewnych własności ogólnych, np. rozwiazanie zagadnienia przy granicznie małej gęstości. W tym wypadku zasięg oddziaływania dwucząstkowego jest znacznie mniejszy od średniej odległości między cząstkami.

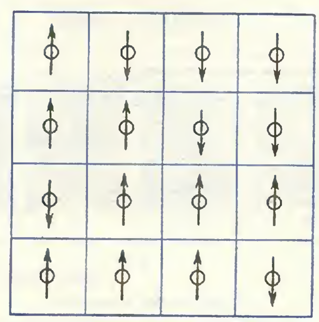
metody rozwinięć grupowych

Dzięki temu dowolna wybrana cząstka oddziałuje z pozostałymi stosunkowo rzadko, co daje możliwość rozwiazania zagadnienia przy granicznie małych wartościach N/V. Zagadnienie obliczania Z w sytuacjach granicznych nie zostało tu wyczerpane. Chodziło bowiem tylko o uwidocznienie podstaw fizycznych takich obliczeń.

Przejdźmy do omówienia pewnych modeli oddziaływania. Oczekuje się wówczas, że $\ln Z$ da się obliczyć asymptotycznie dokładnie w granicy termodynamicznej. Takich modeli może można znaleźć i dużo, niestety niewiele z nich ma coś wspólnego z rzeczywistością fizyczną. Wśród tych chlubnych wyjątków wymieńmy międzyelektronowe oddziaływanie Bardeena–Coopera–Schrieffera, opisujące całkiem poprawnie termodynamikę nadprzewodników oraz model Isinga. Model ten może opisywać własności magnetyczne niektórych substancji lub też stop dwuskładnikowy. Omówimy niektóre cechy tego modelu.

model Isinga

Wyobraźmy sobie regularną sieć krystaliczną. W każdym jej węźle j znajduje się moment magnetyczny S_j, przyjmujący dwie wartości ± 1 (rys. 5).



Rys. 5. Układ 16 momentów magnetycznych; strzałka do góry oznacza wartość $S_j = +1$, w dół $S_j = -1$. Węzły oznaczone kółkami. Energia układu wynosi $-H-4J$

Oddziaływanie momentów ze sobą oraz z zewnętrznym polem magnetycznym H ma postać

$$E = -\sum_j HS_j - J\sum_{(jk)} S_j S_k,$$

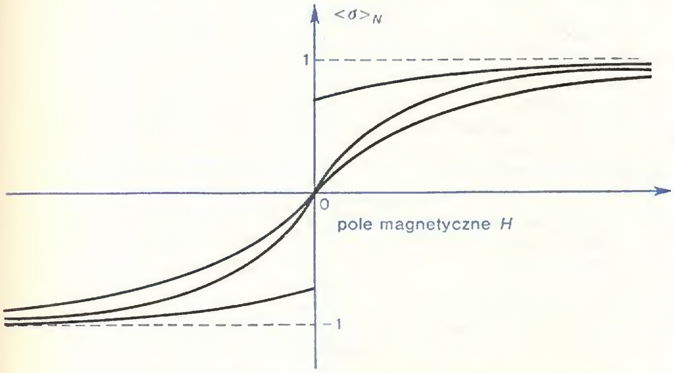
gdzie pierwsza suma przebiega wszystkie N węzłów sieci krystalicznej, druga zaś — wszystkie pary (jk) najbliższych sąsiadów w sieci. Aby obliczyć sumę

stanów Z należy E podstawić na miejsce E_l i wysumować po wszystkich możliwościach $S_j = \pm 1$, liczba tych możliwości jest dana przez 2^N. Jeżeli $H > 0$ i $J > 0$, to najniższa wartość E jest osiągana, gdy wszystkie $S_j = 1$, natomiast przy $H = 0$ również wtedy, gdy wszystkie $S_j = -1$. Obliczmy średni moment sieci przypadający na jeden węzeł (wielkość intensywna). W tym celu wielkość $\sigma = (\sum_j S_j)/N$ trzeba pomnożyć przez $Z^{-1}e^{-E/kT}$ i wysumować po wszystkich $S_j = \pm 1$. Jeśli $H = 0$, to przy skończonych N wielkość ta znika, ponieważ prawdopodobieństwo dowolnego układu momentów S_j jest wtedy dokładnie równe prawdopodobieństwu układu S_j zorientowanych przeciwnie. Stąd przy skończonych N $\lim_{H \to 0} \langle\sigma\rangle_N = 0$ (gdzie $\langle ... \rangle_N$ oznacza wartość średnią przy ustalonym N) ze względu na ciągłą zależność prawdopodobieństwa od H. Może się jednak zdarzyć, że granica termodynamiczna i granica $H \to 0$ są nieprzemienne, tj. że

średni moment sieci

$$\lim_{H \to 0} \lim_{N \to \infty} \langle\sigma\rangle_N \neq \lim_{N \to \infty} \lim_{H \to 0} \langle\sigma\rangle_N = 0,$$

przy $T < T_k$ zwanej temperaturą krytyczną. Przykład zależności $\langle\sigma\rangle_N$ od H, przy różnych N i przy ustalonej temperaturze w zakresie $0 < T < T_k$, pokazano na rys. 6. Mniejszym N odpowiadają krzywe położone bliżej osi poziomej. Krzywa najdalsza, nieciągła w punkcie $H = 0$ granica ciągu funkcji ciągłych $\langle\sigma\rangle_N$, odpowiada granicy termodynamicznej.



Rys. 6. Zależność średniego momentu sieci przypadającego na jeden węzeł $\langle\sigma\rangle_N$ od pola magnetycznego H

Jeśli wielkość $\lim_{H \to 0} \lim_{N \to \infty} \langle\sigma\rangle_N$ nie znika, to nazywamy ją momentem spontanicznym na jeden węzeł. Pojawienie się momentu spontanicznego przy $T = T_k$ oznacza przejście fazowe od fazy paramagnetycznej do ferromagnetycznej. W punkcie $T = T_k$ drugie pochodne $\ln Z$ względem H i T, związane z podatnością magnetyczną i ciepłem właściwym, mają osobliwość. Dla skończonych N jest to niemożliwe, bowiem Z jest wtedy sumą 2^N dodatnich funkcji zmiennych H i T, ciągłych i różniczkowalnych dowolną ilość razy. Widzimy z tego, że poza granicą termodynamiczną przejście fazowe nie istnieje. W rzeczywistości mamy do czynienia z bardzo dużą, ale skończoną liczbą cząstek. W tym wypadku np. biegun funkcji termodynamicznej przechodzi w bardzo ostre maksimum, o wartości zależnej od N. Zbadanie tej zależności jest interesujące zarówno z teoretycznego jak i doświadczalnego punktu widzenia. Niestety, dokładność pomiarów jest na ogół niewystarczająca do jej wyjaśnienia.

W latach dwudziestych E. Ising wykazał, że łańcuch jednowymiarowy nie ma przejścia fazowego. W latach czterdziestych L. Onsager otrzymał dokładne funkcje termodynamiczne modelu Isinga dla sieci kwadratowej płaskiej. Okazało się, że przejście fazowe występuje przy $kT_k = 2J[\ln(\sqrt{2}+1)]^{-1} \approx 2{,}269185J$. Było to pierwsze przejście fazowe obliczone w sposób ścisły na podstawie modelu oddziaływania.

Wiadomo również, że przejście fazowe występuje w wypadku sieci trójwymiarowych, choć dokładnych funkcji termodynamicznych i T_k dotychczas dla nich nie znaleziono.

Osobliwości funkcji termodynamicznych prowadzą do wzrostu fluktuacji momentu magnetycznego przy $T \to T_k$. Ponadto w otoczeniu T_k momenty magnetyczne zaczynają się porządkować. Stąd, jeśli w jakimkolwiek miejscu pojawi się fluktuacja momentu, to spowoduje ona uporządkowanie w dużym obszarze wokół tego miejsca. Można więc stwierdzić, że promień korelacji fluktuacji momentu rośnie przy $T \to T_k$. Ze względu na wzrost fluktuacji i ich promienia korelacji, fluktuacje stają się makroskopowe przy $T \to T_k$. Daje to możliwość niekwantowego potraktowania pola fluktuacji i operowania nim jako obiektem wyjściowym dla teorii. Pozwoliło to wyznaczyć postać asymptotyczną funkcji termodynamicznych przy $T \to T_k$ i stało się źródłem szybkiego rozwoju teorii przejść fazowych w ostatnim dziesięcioleciu, mimo braku wyraźnego postępu w zakresie modeli dokładnie rozwiązalnych. Zasadnicza rola klasycznego pola fluktuacji w pobliżu T_k powoduje, że przejścia fazowe, niezależnie od różnicy ich mechanizmów, stają się do siebie bardzo podobne.

Omawialiśmy dotychczas termodynamikę statystyczną stanów równowagi. Na zakończenie dołączmy parę słów o nierównowagowej termodynamice statystycznej. Wyobraźmy sobie, że na znajdujący się w równowadze układ makroskopowy podziałaliśmy polem zewnętrznym, wywołującym zaburzenia stanu równowagi. Prawdopodobieństwa znalezienia stanu kwantowego l w zespole zaczynają zależeć wtedy od czasu, podobnie jak i wartości średnie wielkości fizycznych. Równania opisujące zmiany tych wielkości pod wpływem przyłożonych pól zewnętrznych wywodzą się, podobnie jak równowagowa termodynamika statystyczna, z równań ruchu oraz podejścia statystycznego, związanego z niepełnym opisem układu. Ostatnia kwestia wynika z zainteresowania jedynie wielkościami makroskopowymi opisującymi układ. Otrzymane równania nie mają, z reguły, zamkniętej postaci układu skończonej liczby n równań dla n niewiadomych funkcji. Układ taki udaje się doprowadzić do postaci zamkniętej w szczególnych warunkach fizycznych. Dotyczy to przede wszystkim gazów rozrzedzonych, tj. takich, dla których promień oddziaływania międzycząsteczkowego jest znacznie mniejszy od średniej odległości między cząsteczkami. Równania dla takiego układu, w przypadku klasycznym, zostały podane przez Boltzmanna w latach siedemdziesiątych ubiegłego stulecia. Mają one kształt równań bilansu prawdopodobieństwa w przestrzeni fazowej jednej cząstki, przy czym zderzenia międzycząstkowe grają istotną rolę w bilansie. Podobnie można z reguły podać zamknięty układ równań dla układów słabowzbudzonych, kiedy możemy posłużyć się obrazem kwazicząstek, przy czym własności kwantowe ich gazu są bardzo istotne. Można wreszcie mówić o zamknięciu układu równań dla niektórych modeli oddziaływania. Podobnie można przewidzieć nierównowagowe własności układów w pobliżu punktu krytycznego T_k. Dodajmy jeszcze, że listę tę trudno jest wyczerpać.

nierównowagowa termodynamika statystyczna

We wszystkich sytuacjach równania istotnie upraszczają się, gdy pole zaburzające jest na tyle słabe, że odchylenie od równowagi można uznać za liniową funkcję pola. Prawa Ohma czy Fouriera wynikają z nierównowagowej termodynamiki statystycznej właśnie jako rezultat takiego zlinearyzowanego podejścia. Podobnie, jeśli pole zmienia się bardzo słabo na średniej odległości międzycząstkowej, to reakcję układu na pole można potraktować w sposób makroskopowy, co znakomicie upraszcza zarówno równania jak i ich rozwiązania.

słabe pole zaburzające

Należy stwierdzić, że bardziej ścisłe lub dokładne rozwiązanie zagadnienia wielu oddziałujących cząstek jest bardzo często trudne lub wręcz niemożliwe. Jest to przyczyną rozwoju wszelkiego rodzaju metod ekstrapolacyjnych w termodynamice statystycznej, których dokładność dość trudno oszacować. Wspomniane trudności sprawiają, że bada się również nierówności dla funkcji termodynamicznych.

L. Landu, E. Lifszic *Fizyka teoretyczna, Fizyka statystyczna*, Warszawa 1959; F. Reif *Fizyka statystyczna*, Warszawa 1971.

Przejścia fazowe i zjawiska krytyczne

Bogusław Mrygoń

Układy fizyczne złożone z wielu cząstek (atomów, molekuł) możemy podzielić na dwa rodzaje. Do pierwszej kategorii zaklasyfikujemy takie układy, które można opisać traktując je jako zbiory cząstek nieoddziałujących ze sobą. Przykładem takiego układu jest gaz idealny złożony z nieoddziałujących atomów lub paramagnetyk, w którym oddziaływanie momentów magnetycznych atomów jest na tyle słabe, że nie wpływa na właściwości magnetyczne układu. Drugą grupę stanowią układy, których nawet w przybliżeniu nie można opisać teoretycznie, nie uwzględniając mikroskopowych oddziaływań. Jeżeli układ taki znajduje się w odpowiednich warunkach zewnętrznych, to mikroskopowe oddziaływania mają charakter zjawiska kolektywnego, które w zależności od rodzaju oddziaływań przejawia się jako określone właściwości makroskopowe układu. Dla zilustrowania tego wymienimy kilka najbardziej typowych układów, których specyficzne właściwości są wynikiem kolektywnych oddziaływań mikroskopowych. Spontaniczne namagnesowanie kryształu ferromagnetycznego jest rezultatem oddziaływania wymiany spinów atomowych.

układy cząstek nieoddziałujących

układy cząstek oddziałujących

W metalach w niskich temperaturach pojawia się stan nadprzewodnictwa, który jest wynikiem specyficznego (niekulombowskiego) oddziaływania elektronów. Krystaliczny stan skupienia wynika z oddziaływań międzyatomowych.

W wymienionych tutaj przykładach kolektywne oddziaływania mikroskopowe powodują powstawanie określonych stanów układu, którym ogólnie można przypisać pewien rodzaj mikroskopowego uporządkowania. Wyjątek stanowi kondensacja idealnego gazu Bosego. Stan układu — lub jego części — charakteryzujący się przestrzenną jednorodnością makroskopową i określonym mikroskopowym uporządkowaniem nazywamy fazą. Układ może się znajdować w danej fazie tylko w ściśle określonych przedziałach wartości zmiennych termodynamicznych, takich jak: temperatura, ciśnienie, natężenie pola magnetycznego. W ogólności możliwe jest istnienie kilku faz danego układu. Przejście układu od jednej do drugiej fazy, nazywane ogólnie przejściem (przemianą) fazowym, zachodzi w określonych warunkach termodynamicznych, np. gdy temperatura osiąga wartość krytyczną T_k, nazywaną też punktem krytycznym układu. Przejścia fazowe należą do zjawisk występujących powszechnie w różnych układach fizycznych. Najważniejsze z nich to: zmiany stanu skupienia, przejścia magnetyczne typu ferromagnetyk–paramagnetyk, przejścia typu porządek–nieporządek w stopach podwójnych oraz przejścia do stanu nadprzewodnictwa. Lista ta nie wyczerpuje wszystkich przejść fazowych, które występują szczególnie w ciałach stałych. Z przejściami fazowymi związane są zjawiska nazywane krytycznymi, których przejawem jest charakterystyczne zachowanie się w pobliżu punktu krytycznego niektórych właściwości układu, np. ciepła właściwego. Mimo ogromnych różnic między naturą mikroskopowych oddziaływań w różnych układach, przejścia fazowe i zjawiska krytyczne zachodzące w tych układach wykazują wiele wspólnych cech i analogii. Fakt ten umożliwia podanie ogólnego, jednolitego opisu wszystkich przejść fazowych i zjawisk krytycznych.

faza układu

termodynamiczny opis przejścia fazowego

W fenomenologicznym opisie zachowania się układu fizycznego posługujemy się zbiorem zmiennych termodynamicznych. Dla każdego szczególnego układu możemy wyodrębnić wiele zmiennych termodynamicznych, ale do kompletnego opisu jednoskładnikowego układu znajdującego się w jednej fazie wystarczą trzy zmienne termodynamiczne, przy czym tylko dwie z nich są zmiennymi niezależnymi. Liczbę zmiennych niezależnych n zwaną liczbą termodynamicznych stopni swobody wyraża się wzorem

$$n = c+2-r, \tag{1}$$

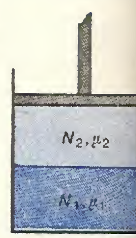
gdzie c jest liczbą niezależnych składników układu, a r — liczbą faz współistniejących w równowadze. W zależności od wyboru zmiennych niezależnych określona jest charakterystyczna funkcja stanu. Jeżeli do opisu gazu złożonego z N atomów wybierzemy jako zmienne niezależne temperaturę T i ciśnienie p, to wszystkie informacje o zachowaniu się i właściwościach układu zawarte są w funkcji stanu $G(T, p)$ i możemy je otrzymać z odpowiednich związków termodynamicznych. Funkcja $G(T, p)$ jest często nazywana potencjałem Gibbsa. Jeżeli parę zmiennych niezależnych tworzy temperatura T i objętość V układu, to funkcją stanu jest $F(T, V)$ — zwana energią swobodną. Wszystkie możliwe termodynamiczne funkcje stanu są ogólnie nazywane potencjałami termodynamicznymi, ponieważ pracę związaną z przejściem układu od stanu a do stanu b można wyrazić przez spowodowany tym przejściem przyrost funkcji stanu.

Warunkiem równowagi termodynamicznej układu jest minimum potencjałów termodynamicznych (→ Termodynamika fenomenologiczna) przy ustalonych wartościach odpowiednich zmiennych niezależnych. Każdy układ zamknięty osiąga po pewnym czasie stan równowagi termodynamicznej. Jeżeli przejście fazowe zachodzi w dostatecznie długim czasie, to możemy założyć, że wszystkie stany pośrednie układu w otoczeniu punktu krytycznego są stanami równowagi, a więc do opisu przejścia fazowego możemy stosować formalizm termodynamiki procesów odwracalnych.

układ dwufazowy

Rozpatrzmy przy ustalonej temperaturze T i ciśnieniu p jednoskładnikowy układ (rys. 1), który znajduje się w stanie równowagi dwu faz (układ dwufazowy). Może to być faza ciekła w kontakcie z fazą gazową. Ciecz i gaz mogą współistnieć w równowadze, jeżeli temperatura i ciśnienie obydwu faz są sobie równe. Przyjmijmy, że N_1 i N_2 oznaczają odpowiednio liczby atomów, a μ_1 i μ_2 — wartości potencjału Gibbsa na jeden atom w obydwu fazach ($\mu = G/N$ jest potencjałem chemicznym). Funkcję $G(T, p)$ możemy więc wyrazić wzorem:

$$G(T, p) = N_1\mu_1+N_2\mu_2. \tag{2}$$



Rys. 1.

Ponieważ ciśnienie i temperatura układu są stałe, warunek na minimum potencjału termodynamicznego (w tym wypadku $dG = 0$) zredukuje się do postaci:

$$\mu_1 dN_1+\mu_2 dN_2 = 0. \tag{3}$$

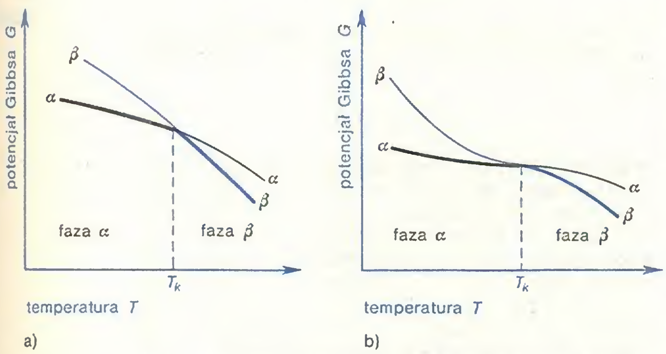
W rozpatrywanym układzie całkowita liczba atomów jest zachowana ($dN_1+dN_2 = 0$) zatem

$$\mu_1 = \mu_2. \tag{4}$$

Jest to jeden z ważniejszych wniosków termodynamiki przejść fazowych: w obszarze współistnienia, a tym samym również w punkcie krytycznym, potencjały chemiczne obydwu faz są sobie równe.

przejścia fazowe n-tego rodzaju

Analizując fakty doświadczalne, Ehrenfest wprowadził klasyfikację przejść fazowych; przemianę nazywa się przejściem fazowym n-tego rodzaju, jeżeli kolejne pochodne potencjału Gibbsa aż do $(n-1)$-szej włącznie są funkcjami ciągłymi, natomiast n-ta pochodna ma w punkcie przejścia nieciągłość skokową. Zgodnie z takim określeniem, potencjał G i jego pochodne będą miały następujący przebieg względem każdej zmiennej niezależnej: przy przejściu I rodzaju funkcja G jest ciągła załamana (rys. 2a), natomiast jej pochodne są nieciągłe; przy przejściu II rodzaju funkcja G jest ciągła gładka (rys. 2b), jej pierwsza pochodna jest ciągła załamana, natomiast wyższe pochodne są nieciągłe. Opis przejść fazowych zastosowany w definicjach Ehrenfesta należy traktować jako



Rys. 2. Przykłady możliwych przebiegów potencjału termodynamicznego w obszarze przejścia fazowego: a) I rodzaju, b) II rodzaju

idealizację rzeczywistych przebiegów funkcji G i jej pochodnych. Pochodne te mają określony sens fizyczny, np. ciepło właściwe przy stałym ciśnieniu wyraża się przez drugą pochodną funkcji G względem temperatury:

$$C_p = -T\left(\frac{\partial^2 G}{\partial T^2}\right)_p. \qquad (5)$$

przejścia fazowe I i II rodzaju

Obserwowane w otoczeniu punktu krytycznego przebiegi wielkości fizycznych w układach rzeczywistych o skończonych rozmiarach są co najwyżej zbliżone do teoretycznie przewidywanych w klasyfikacji Ehrenfesta. Za typowe przejścia I rodzaju uważa się zmiany stanu skupienia oraz zmiany struktury krystalicznej. Najczęściej wymienianymi przykładami przejść II rodzaju są przejścia od stanu ferromagnetycznego do paramagnetycznego w punkcie Curie oraz przejście od stanu nadprzewodnictwa do stanu zwykłego przewodzenia, zachodzące w temperaturze krytycznej przy nieobecności zewnętrznego pola magnetycznego. Wiele przejść fazowych nie mieści się w ogóle w klasyfikacji Ehrenfesta. Znane są również inne klasyfikacje przejść fazowych, np. oparta na przebiegu ciepła właściwego w pobliżu temperatury krytycznej. Istnieje przy tym dość duża dowolność zaszeregowania konkretnych przejść fazowych do określonego rodzaju.

zmiany stanu skupienia

Na rys. 3 przedstawiono schematycznie wykres fazowy wszystkich, z wyjątkiem helu (→ Nadpłynność), substancji jednoskładnikowych. Krzywe *1*, *2* i *3* oddzielają obszary, w których substancja występuje w określonym stanie skupienia i nazywane są krzywymi równowagi faz. Dla wartości zmiennych termodynamicznych, określonych przez te krzywe, możliwe jest współistnienie odpowiednich faz w równowadze. W punkcie wspólnym trzech krzywych równowagi mogą zatem współistnieć trzy fazy. Punkt ten nazywany jest punktem potrójnym.

punkt potrójny

Jak już wspomniano, zmiany stanu skupienia są przejściami fazowymi I rodzaju. Przejścia te połączone są z wydzieleniem lub pochłonięciem ciepła (tzw. utajone ciepło przemiany: ciepło parowania, ciepło topnienia, ciepło sublimacji). Wyjątek stanowi przejście fazowe w punkcie K, które jest ciągłe, tzn. w punkcie K można przemienić ciecz w gaz i odwrotnie w sposób ciągły. Ciepło przemiany w punkcie K, zwanym punktem krytycznym, znika. W obszarach fazy ciała stałego różnych substancji obserwujemy wiele innych przejść fazowych, których przykłady podano na wstępie. Zachowanie się substancji w pobliżu punktów krytycznych zlokalizowanych w fazie stałej (np. krytyczne zachowanie magnetyków, ferroelektryków, nadprzewodników) wykazuje wiele analogii do zachowania się gazu w pobliżu punktu krytycznego K. Z tego względu omówimy tu bardziej szczegółowo krzywą równowagi faz w układzie ciecz–gaz.

przejście fazowe w gazie rzeczywistym

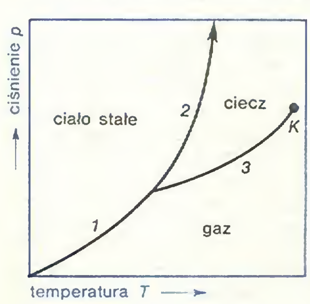
Przejście ciągłe w układzie ciecz–gaz odkryte zostało w 1869 r. przez Andrewsa, który badał zachowanie się krytyczne dwutlenku węgla. Zachowanie się gazu opisane jest przez równanie stanu. Zakładając, że znamy potencjał oddziaływania $U(r)$ między atomami, możemy dla gazu złożonego z N atomów w objętości V napisać ścisłe równanie stanu w postaci:

$$p = \frac{NkT}{V} + \frac{\overline{W}}{3V}, \qquad (6)$$

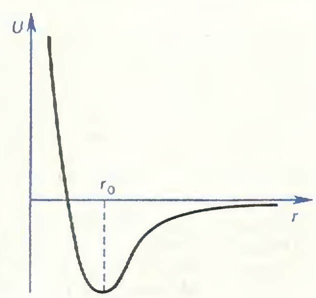
gdzie k oznacza stałą Boltzmanna, a $\overline{W}$ jest wielkością związaną z siłami międzyatomowych oddziaływań, nazywaną wiriałem sił wewnętrznych:

$$\overline{W} = \overline{\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} F_{ij}|\vec{r_i}-\vec{r_j}|}. \qquad (7)$$

Wektory $\vec{r_i}$ i $\vec{r_j}$ oznaczają położenia atomów, a F_{ij} jest siłą, z jaką te atomy oddziałują. Wyrażenie (7) uwzględnia oddziaływania wszystkich par atomów w układzie, przy czym aby każda para była liczona tylko raz, przed znakiem sumy występuje współczynnik $^1/_2$. Siła F zależy od potencjału $U(r)$, którego kształt jest przedstawiony na rys. 4, przy czym $F = -\partial U(r)/\partial r$. Pierwszy wyraz równania (6), tzw. ciśnienie kinetyczne, jest związane z energią kinetyczną atomów i opisuje zachowanie się gazu idealnego, w którym przejście fazowe nie zachodzi. Drugi wyraz — ciśnienie statyczne — stanowi poprawkę do równania stanu związaną z energią po-



Rys. 3. Schematyczny wykres fazowy substancji jednoskładnikowej



Rys. 4. Potencjał oddziaływania międzyatomowego w gazie rzeczywistym. Widać, że przy $r>r_0$ $\partial U/\partial r>0$, zatem działają siły przyciągające, natomiast przy $r<r_0$ działają siły odpychające

równanie van der Waalsa

tencjalną wzajemnego oddziaływania atomów. W opisie gazu rzeczywistego posługujemy się różnymi przybliżeniami równania stanu (6). Rozpatrzmy jedno z najbardziej znanych przybliżeń tego równania, historycznie znacznie od niego wcześniejsze, półempiryczne równanie van der Waalsa

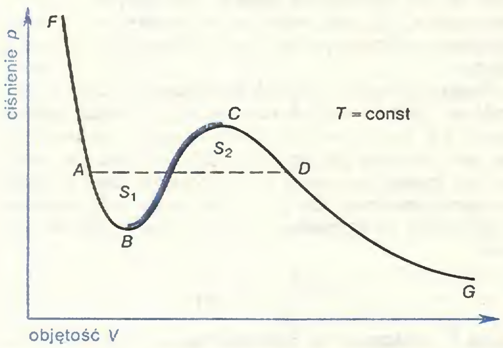
$$\left(p+\frac{N^2a}{V^2}\right)(V-Nb) = NkT, \tag{8}$$

gdzie stałe a i b określane są dla każdego gazu eksperymentalnie. Stała a związana jest z oddziaływaniem między atomami (molekułami), b — ze skończonymi rozmiarami atomów. Na rys. 5 przedstawiona jest typowa izoterma otrzymana z równania (8). Warunek równowagi (minimum potencjału termodynamicznego) można, wykorzystując tożsamościowe związki termodynamiki, zapisać w postaci

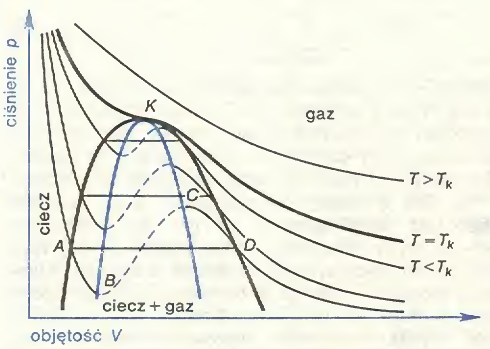
$$\left(\frac{\partial V}{\partial p}\right)_T < 0. \tag{9}$$

warunek stabilności mechanicznej

Nierówność ta, nazywana warunkiem stabilności mechanicznej, oznacza, że jeżeli przy stałej temperaturze ciśnienie układu wzrasta, to jego objętość musi maleć. Z rys. 5 widać, że na odcinku BC izoterma van der Waalsa nie spełnia warunku (9). Oznacza to, że izotermy otrzymane z równania (8) są przynajmniej



Rys. 5. Izoterma opisana równaniem van der Waalsa



Rys. 6. Wykres fazowy gazu rzeczywistego

w części niefizyczne. Wynika to z faktu, że równanie (8) jest tylko przybliżeniem równania (6). Niefizyczną izotermę van der Waalsa można poprawić, posługując się tzw. konstrukcją Maxwella. Polega ona na wykreśleniu poziomego odcinka AD (linia przerywana), który łączy dwie gałęzie izotermy FA i DG odpowiadające fazie ciekłej i fazie gazowej układu. Z warunku (4) wynika, że odcinek AD należy tak wykreślić, aby powierzchnie S_1 i S_2 były sobie równe. Między punktem A i D układ jest niejednorodny, tzn. podzielony na dwie fazy współistniejące w równowadze. Odcinkom AB i CD izotermy van der Waalsa odpowiadają stany cieczy przegrzanej i gazu przechłodzonego — możliwe w szczególnych warunkach.

konstrukcja Maxwella

Rys. 6 przedstawia rodzinę izoterm (8) wykreślonych przy wykorzystaniu konstrukcji Maxwella. Miejscem geometrycznym końcowych punktów fizycznych gałęzi izoterm jest krzywa BKC (rys. 6) ograniczająca obszar, w którym układ nie może w żadnym wypadku występować jako jednofazowy. Podobnie otrzymuje się krzywą równowagi faz AKD ograniczającą obszar, w którym dwie fazy współistnieją w równowadze. W obszarach między krzywą AKD i BKC możliwe jest występowanie cieczy przegrzanej lub gazu przechłodzonego. Punkt K, w którym krzywe BKC i AKD są styczne, tzn. punkt określony współrzędnymi p_k, V_k, T_k, któremu odpowiada nieskończenie mała długość odcinka AD — nazywamy punktem krytycznym. W punkcie krytycznym

punkt krytyczny

$$\left(\frac{\partial V}{\partial p}\right)_T = 0. \tag{10}$$

Aby stan układu w punkcie krytycznym był stanem równowagi, musi znikać również druga pochodna, tzn.

$$\left(\frac{\partial^2 V}{\partial p^2}\right)_T = 0. \tag{11}$$

Warunek ten wynika z ogólniejszych rozważań termodynamicznych. Spełnienie obu warunków zostało potwierdzone doświadczalnie, w granicach dokładności pomiarów. Z warunków równowagi stanu krytycznego wynika, że ciepło właściwe gazu przy stałym ciśnieniu dąży w punkcie krytycznym do nieskończoności. Równania (8), (10) i (11) pozwalają wyrazić temperaturę krytyczną T_k, ciśnienie krytyczne p_k i objętość krytyczną V_k przez stałe a i b:

$$T_k = \frac{8}{27}\frac{a}{bk}, \quad V_k = 3Nb, \quad p_k = \frac{1}{27}\frac{a}{b^2}. \tag{12}$$

Wprowadzając następnie bezwymiarowe zmienne zredukowane

$$\bar{p} = p/p_k; \quad \bar{V} = V/V_k, \quad \bar{T} = T/T_k, \tag{13}$$

otrzymujemy zredukowane równanie van der Waalsa:

$$\left(\bar{p}+\frac{3}{\bar{V}^2}\right)\left(\bar{V}-\frac{1}{3}\right) = \frac{8}{3}\bar{T}. \tag{14}$$

W równaniu tym nie ma żadnych stałych charakteryzujących dany gaz. Zatem jest to równanie stanu wszystkich układów, do których można stosować równanie van der Waalsa. Stany dwóch układów, których zmienne zredukowane mają takie same wartości, nazywamy stanami odpowiednimi lub stanami odpowiadającymi sobie. Zredukowane izotermy określone równaniem (14), są jednakowe dla wszystkich gazów. Wynik taki jest znany jako prawo stanów odpowiednich. Stany krytyczne wszystkich gazów też są oczywiście stanami odpowiednimi. Na podstawie wyznaczonych wartości T_k, p_k i V_k znajdujemy, że dla jednego mola gazu zachodzi

stany odpowiadające sobie

$$\frac{p_k V_k}{RT_k} = \frac{3}{8}, \tag{15}$$

gdzie R jest stałą gazową niezależną od rodzaju gazu. Zostało stwierdzone doświadczalnie, że stosunek (15), nazywany krytycznym, rzeczywiście jest prawie taki sam dla różnych gazów. Np. wartość stosunku krytycznego wyznaczona doświadczalnie dla neonu wynosi 0,295, dla argonu 0,290, ksenonu 0,293, azotu 0,291, tlenu 0,292, etanu 0,288. Średnia dla wielu gazów wynosi 0,292. Wprawdzie niezgodność między teoretycznym wynikiem (15) i wartością doświadczalną stosunku krytycznego jest dość duża (ok. 25%), ale trzeba pamiętać, że równanie van der Waalsa jest tylko przybliżeniem ścisłego równania

stosunek krytyczny

stanu. Niemniej jednak równanie to daje ogólny i zgodny jakościowo z faktami doświadczalnymi opis przejścia fazowego w gazie rzeczywistym.

Można podać analogiczne opisy przejść fazowych w innych układach fizycznych. Dla każdego rodzaju układu fizycznego istnieją charakterystyczne zmienne termodynamiczne, ale ogólna postać związków termodynamicznych jest wspólna dla wszystkich układów. Jeżeli dokonamy np. następującej zamiany zmiennych: $V \rightarrow -M, p \rightarrow H$ (M oznacza namagnesowanie układu, a H — natężenie pola magnetycznego), to funkcje termodynamiczne wyrażone w nowych zmiennych będą opisywały układ magnetyczny. Potencjał termodynamiczny G układu magnetycznego jest więc funkcją temperatury T i natężenia pola H. Wzór (5) określa w tym wypadku magnetyczne ciepło właściwe. Dla każdego układu możemy również wyprowadzić odpowiednie równanie stanu. Poniżej podamy tylko ogólne charakterystyki różnych przejść fazowych w wybranych układach fizycznych.

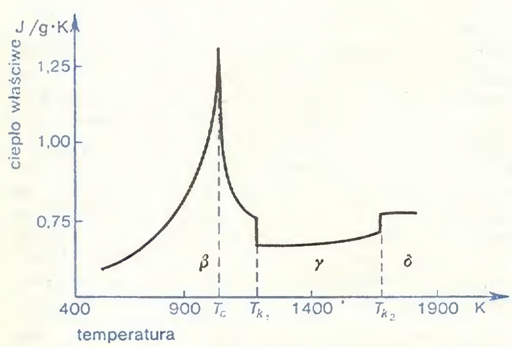
przejścia magnetyczne

W temperaturze Curie w ferromagnetykach lub temperaturze Neéla w antyferromagnetykach układ magnetyczny przechodzi od stanu paramagnetycznego do ferro- lub antyferromagnetycznego (→ Teoria magnetyzmu). Spiny atomów zorientowane chaotycznie w fazie para- tworzą poniżej punktu krytycznego odpowiednie uporządkowanie magnetyczne. W ferromagnetyku powstaje spontaniczne namagnesowanie układu. W antyferromagnetyku wypadkowy moment magnetyczny układu równy jest zeru, można natomiast mówić o spontanicznym namagnesowaniu podsieci. Są również znane przejścia typu paramagnetyk–ferrimagnetyk (ferryty), w których faza odpowiadająca uporządkowaniu magnetycznemu spinów ma wypadkowe namagnesowanie wynikające z niepełnej kompensacji momentów magnetycznych dwu podsieci.

W punktach przejścia magnetycznego niektóre wielkości termodynamiczne wykazują anomalie (rys. 7, punkt T_C), mające ten sam charakter w różnych magnetykach. Bardzo podobne przejścia fazowe obserwuje się w ferroelektrykach.

przejścia strukturalne w kryształach

W wielu kryształach występują przejścia fazowe polegające na zmianie symetrii struktury krystalicznej. Zachodzą one przy zmianie temperatury lub ciśnienia i są najczęściej przejściami fazowymi I rodzaju. Towarzyszy im wydzielenie lub pochłonięcie ciepła (rys. 7, punkty T_{k1} i T_{k2}).



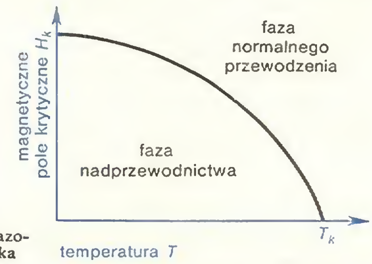
Rys. 7. Ciepło właściwe żelaza w funkcji temperatury. Punkt T_C odpowiada temperaturze Curie (przejście magnetyczne). W punktach T_{k1} i T_{k2} zachodzą przejścia strukturalne. Żelazo w fazie β i δ tworzy sieć regularną przestrzennie centrowaną, a w fazie γ — płasko centrowaną

uporządkowanie stopów

W dwuskładnikowych stopach o składzie stechiometrycznym powstają w określonych temperaturach stany mające cechy związku chemicznego. Oznacza to, że poniżej temperatury przejścia fazowego powstaje faza o całkowitym, wzajemnym uporządkowaniu dwóch rodzajów atomów A i B, np. $ABABAB\ldots$. Wzajemne położenia atomów A i B powyżej temperatury krytycznej są zupełnie przypadkowe.

przejścia do stanu nadprzewodnictwa

Wiele metali i stopów ochłodzonych do dostatecznie niskich temperatur staje się nadprzewodnikami. Zjawisko to stwierdzono w próbkach ponad 30% pierwiastków i więcej niż 1000 stopów. W fazie nadprzewodzącej znika opór elektryczny (→ Nadprzewodnictwo), przy czym przejście od stanu przewodzenia normalnego do stanu nadprzewodnictwa jest bardzo ostre. Ponadto nadprzewodnik umieszczony w słabym polu magnetycznym zachowuje się jak doskonały diamagnetyk, tzn. indukcja magnetyczna wewnątrz nadprzewodzącego kryształu jest równa zeru (zjawisko Meissnera). Przejście od stanu nadprzewodnictwa do stanu przewodzenia normalnego może być spowodowane wzrostem temperatury powyżej wartości krytycznej T_k lub zewnętrznym polem magnetycznym o natężeniu większym od wartości krytycznej H_k. Temperatury krytyczne zbadanych nadprzewodników znajdują się w przedziale 0,1–20 K. Natężenie pola krytycznego H_k jest funkcją temperatury (rys. 8).



Rys. 8. Wykres fazowy nadprzewodnika

wykładniki krytyczne

Charakterystyczną, doświadczalnie stwierdzoną cechą wszystkich przejść fazowych jest anomalne zachowanie się niektórych wielkości termodynamicznych, gdy układ zbliża się do punktu krytycznego. Charakter anomalnego przebiegu wielkości termodynamicznej $f(\varepsilon)$ w otoczeniu punktu krytycznego można przedstawić w ogólnej postaci:

$$f(\varepsilon) \sim \varepsilon^{\nu}, \quad \text{gdy } \varepsilon \to 0, \tag{16}$$

gdzie $\varepsilon = |T - T_k|/T_k$ oznacza zredukowaną odległość od punktu krytycznego na osi temperatury, a ν jest tzw. wykładnikiem krytycznym. Zależność (16) nie oznacza, że wartość wielkości $f(\varepsilon)$ jest proporcjonalna do prostego wyrażenia potęgowego ε^{ν}, a określa jedynie charakter zachowania się tej wielkości, gdy $T \to T_k$. Przy założeniu, że $f(\varepsilon)$ jest funkcją dodatnią, wykładnik krytyczny jest zdefiniowany następująco:

$$\nu = \lim_{\varepsilon \to 0} \frac{\ln f(\varepsilon)}{\ln \varepsilon}. \tag{17}$$

Z zależności (16) widać, że jeżeli $\nu > 0$, to funkcja $f(\varepsilon)$ dąży do zera, gdy $T \to T_k$. Dla $\nu < 0$ funkcja $f(\varepsilon)$ w punkcie krytycznym jest rozbieżna do nieskończoności. Gdy funkcja $f(\varepsilon)$ jest w temperaturze krytycznej rozbieżna logarytmicznie lub ma wartość skończoną, to $\nu = 0$. Dla danego przejścia fazowego można wprowadzić zbiór wykładników krytycznych, przypisując każdej zachowującej się krytycznie wielkości termodynamicznej odpowiedni wykładnik krytyczny. Wartości wykładników krytycznych wyznacza się albo doświadczalnie, albo teoretycznie, rozpatrując odpowiedni model oddziaływań w danym układzie fizycznym. Tak więc za pomocą jednej ogólnej relacji typu (16) można w sposób przybliżony opisać w otoczeniu punktu krytycznego przebieg wszystkich wielkości termodynamicznych zachowujących się krytycznie, niezależnie od rodzaju układu fizycznego i typu przejścia fazowego.

hipoteza skalowania

Istnienie wykładników krytycznych nie wynika z ogólnych zasad termodynamiki. Relację (16) można wyprowadzić z jednego założenia dotyczącego potencjału termodynamicznego. Wyraża się ją w formie hipotezy, że potencjał Gibbsa jest funkcją jednorodną

zredukowanej temperatury ε oraz drugiej zmiennej A — właściwej dla tej funkcji stanu i danego układu fizycznego. Jeżeli $G(\varepsilon, A)$ jest funkcją jednorodną, to zachodzi dla dowolnych wartości λ następująca równość:

$$G(\lambda^a\varepsilon, \lambda^b A) = \lambda G(\varepsilon, A) \qquad (18)$$

(a i b — pewne parametry). Funkcja jednorodna ma tę właściwość, że jeżeli znamy jej wartości wzdłuż dowolnej krzywej otaczającej początek układu, to możemy wyznaczyć jej wartość w dowolnym punkcie przestrzeni drogą przeskalowania układu współrzędnych, gdy znamy wartości parametrów skalujących a i b. Stąd też założenie (18) nazywane jest hipotezą skalowania. Hipoteza skalowania pozwala wyrazić wszystkie wykładniki krytyczne dla układu opisanego funkcją stanu $G(\varepsilon, A)$ przez dwa parametry skalujące a i b. Z faktu tego wynika, że wykładniki krytyczne nie są wzajemnie niezależne. Związki między różnymi wykładnikami krytycznymi otrzymane z hipotezy skalowania są ogólnie nazywane prawami skalowania. Jeden z bardziej znanych związków między wykładnikami krytycznymi opisującymi przejście fazowe w ferromagnetyku ma np. postać:

prawa skalowania

$$\alpha+2\beta+\gamma=2, \qquad (19)$$

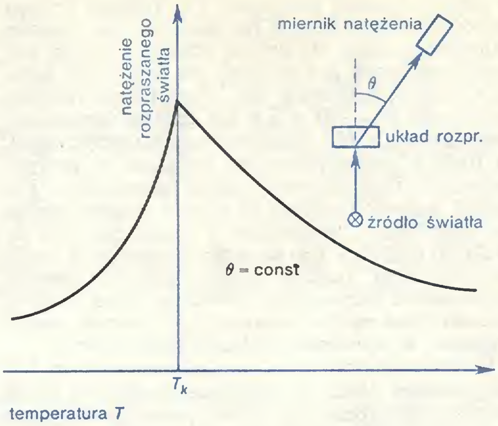
gdzie α, β, γ są wykładnikami krytycznymi opisującymi zgodnie z relacją (16) zachowanie w pobliżu temperatury T_k odpowiednio magnetycznego ciepła właściwego, spontanicznego namagnesowania i podatności magnetycznej. W praktyce, prawa skalowania typu (19) stanowią bardzo użyteczny sprawdzian dokładności przy wyznaczaniu wartości wykładników krytycznych z danych doświadczalnych. Hipoteza skalowania nie daje możliwości wyznaczenia wartości wykładników krytycznych, a jedynie pozwala określić ogólne związki między nimi. Założenie, że potencjały termodynamiczne są funkcjami jednorodnymi, nie zostało udowodnione. Tym niemniej, jednolity i w przybliżeniu zgodny z doświadczeniami opis zjawisk krytycznych, jaki uzyskujemy zakładając potencjał termodynamiczny w postaci (18), pozwala sądzić, że hipoteza skalowania zawiera w sobie pewną ogólną właściwość wyidealizowanych układów fizycznych.

Dla pewnych modeli (np. modelu Isinga, → Termodynamika statystyczna) możliwe jest ścisłe obliczenie funkcji termodynamicznych, jeżeli z układem przechodzimy do granicy termodynamicznej, tzn. dla $N\to\infty$, $V\to\infty$, $N/V=\text{const}$. Przejścia fazowe w takich modelach związane są z punktami osobliwymi funkcji termodynamicznych. Hipoteza skalowania opisuje układy w granicy termodynamicznej.

rozpraszanie krytyczne

Szczególnie ciekawą grupę zjawisk krytycznych tworzą efekty związane z rozpraszaniem promieniowania przez układ znajdujący się w obszarze przejścia fazowego. Ośrodki, które w normalnych warunkach są optycznie przezroczyste, w pobliżu punktu krytycznego stają się mętne. Natężenie promieniowania rozproszonego przez układ pod kątem względem kierunku fali padającej gwałtownie rośnie, gdy $T\to T_k$. Zjawisko to, wyraźne w gazie rzeczywistym lub w cieczach podwójnych, nazywane jest krytyczną opalescencją. Temperaturowy przebieg zjawiska pokazany jest na rys. 9. Zjawisko krytycznej opalescencji zostało wyjaśnione po raz pierwszy przez polskiego fizyka M. Smoluchowskiego, który wykazał, że wzrost natężenia światła rozproszonego pod pew-

krytyczna opalescencja



Rys. 9. Krytyczna opalescencja w gazie

nym kątem spowodowany jest przez fluktuacje gęstości. (Pojęcie fluktuacji gęstości wprowadził do fizyki Smoluchowski; podał on wyrażenie na prawdopodobieństwo ich powstawania). W stanach dalekich od punktu krytycznego fluktuacje są bardzo małe i w opisie właściwości układu można je pominąć. Gdy temperatura zbliża się do punktu krytycznego, lokalne fluktuacje gwałtownie rosną, dzięki czemu układ staje się przestrzennie niejednorodny i wykazuje strukturę „ziarnistą". Fluktuacje w pobliżu punktu krytycznego można traktować jako pojawiające się zarodki nowej fazy. Krytyczna opalescencja jest wynikiem rozpraszania światła na fluktuacjach gęstości. Analogiczne zjawiska występują także w innych układach fizycznych. Na przykład, w magnetykach obserwuje się krytyczne rozpraszanie neutronów na fluktuacjach momentu magnetycznego w pobliżu temperatury Curie; w stopach podwójnych i ferroelektrykach — krytyczne rozpraszanie promieni rentgenowskich.

Teoria krytycznego rozpraszania oparta na modelu Smoluchowskiego nie jest jedyną teorią wyjaśniającą te zjawiska. Opis teoretyczny krytycznego rozpraszania można np. uzyskać na podstawie hipotezy skalowania.

W wynikach doświadczalnych dotyczących przejść fazowych i zjawisk krytycznych jest wiele niejasności, a teoria tych zjawisk zawiera liczne luki i kontrowersje. W tej dziedzinie fizyki prowadzi się intensywne badania, a każdy rok przynosi nowe osiągnięcia. W ostatnich latach główny wysiłek badawczy skierowany był na teoretyczne i doświadczalne wyznaczanie wskaźników krytycznych. Jedną z metod wyznaczania tych wskaźników jest metoda grupy renormalizacyjnej, podana przez K. Wilsona (1971 r.); stanowi ona konsekwentny etap rozwoju teorii skalowania.

metoda grupy renormalizacyjnej

H. E. Stanley *Introduction to Phase Transitions and Critical Phenomena*, Oxford 1971 (tłum. ros., Moskwa 1973).

Elektrodynamika

Jan Mostowski

Trudno jest wyobrazić sobie współczesne życie bez elektryczności. Urządzenia wykorzystujące zjawiska elektryczne spotyka się dosłownie na każdym kroku i nie można podać dziedziny życia, w której urządzenia elektryczne nie odgrywają roli.

Tymczasem historia prac nad poznaniem i wykorzystaniem elektryczności w zasadzie nie jest długa, ma dopiero ok. 200 lat. Wprawdzie najprostsze zjawiska, takie jak elektryzowanie się niektórych ciał, znane były już w starożytności, ale nie było wówczas mowy

o ich głębszym zrozumieniu lub wykorzystaniu. Dopiero w drugiej połowie XVIII w. rozpoczęto ilościowe badania elektryczności. Punktem wyjścia było stwierdzenie Charlesa Coulomba, że punktowe ładunki działają na siebie siłą odwrotnie proporcjonalną do kwadratu odległości między nimi.

Wiek XIX to okres niesłychanie intensywnych prac zmierzających do wyjaśnienia zjawisk elektrycznych i magnetycznych. Ich ukoronowaniem stała się teoria Maxwella, porządkująca i systematyzująca wszystkie zjawiska elektryczne i magnetyczne.

Niemal równocześnie z badaniami podstawowymi pojawiły się urządzenia wykorzystujące zjawiska elektryczne do celów praktycznych.

W XX w. powstała nowa gałąź elektrodynamiki — elektrodynamika kwantowa, która opisuje oddziaływanie pola elektromagnetycznego, m.in. o wielkiej częstotliwości, z mikroskopowymi obiektami, głównie elektronami.

Elektrodynamikę można obecnie uważać za najdoskonalszą teorię fizyczną. Jej prawom podlegają zarówno pola elektromagnetyczne w przestrzeni międzygalaktycznej i międzygwiezdnej, jak i we wnętrzu atomu. Elektrodynamika opisuje olbrzymie bogactwo zjawisk z dziedziny spektroskopii atomów i molekuł, fizyki plazmy i wszystkich innych dziedzin nauki, w których występują ładunki elektryczne. Należy podkreślić, że w żadnym dotychczas przeprowadzonym eksperymencie nie stwierdzono odstępstw od przewidywań elektrodynamiki, mimo iż niektóre doświadczenia były przeprowadzane z dokładnością do 10^{-4}%. Żadna inna teoria fizyczna nie może się szczycić podobnymi sukcesami. Nie oznacza to bynajmniej, że wszelkie efekty elektromagnetyczne są zrozumiałe, wiele pozostało do zrobienia w tej dziedzinie. Można tu wymienić kilka ważnych tematów, nad którymi współcześnie pracują fizycy.

efekty elektromagnetyczne wymagające badań

Pierwsza sprawa wiąże się z optyką i dotyczy silnych pól elektromagnetycznych otrzymywanych z laserów. Duże natężenie pola elektromagnetycznego panujące w wiązce światła laserowego może powodować występowanie tzw. zjawisk nieliniowych. Mimo że podstawowe prawa oddziaływania światła z materią są zrozumiałe, to jednak ich zastosowanie do opisu konkretnych zjawisk jest sprawą trudną i nie jest jeszcze jasne, jaki jest mechanizm powstawania niektórych efektów. Nowo odkryte nieliniowe zjawiska optyczne zachęciły do ogólnych badań nad silnymi wiązkami światła i ich oddziaływaniem z materią (→ Optyka nieliniowa).

zakres stosowalności elektrodynamiki

Drugie bardzo ważne zagadnienie to zakres stosowalności elektrodynamiki. Jak każda teoria fizyczna, tak i elektrodynamika została oparta na doświadczeniach przeprowadzonych z określoną dokładnością. Zwiększanie dokładności pomiarów rozszerza zakres stosowalności teorii lub nakłada na nią ograniczenia. W dalszym ciągu wykonuje się szereg bardzo subtelnych doświadczeń, mających na celu sprawdzenie poprawności elektrodynamiki. Nigdzie nie znaleziono odstępstw od elektrodynamiki, rozszerzył się tylko zakres jej stosowalności.

Trzecia dziedzina współczesnych prac nad elektrodynamiką to kwestia pewnych sprzeczności tkwiących w teorii Maxwella. Chodzi tu o zrozumienie nieskończonej energii oddziaływania punktowego ładunku, np. elektronu, na siebie. Próby modyfikacji teorii Maxwella, które by pozwoliły usunąć te trudności, nie dały dotąd pozytywnych rezultatów, lecz są ciągle na nowo podejmowane.

Co to jest elektrodynamika klasyczna

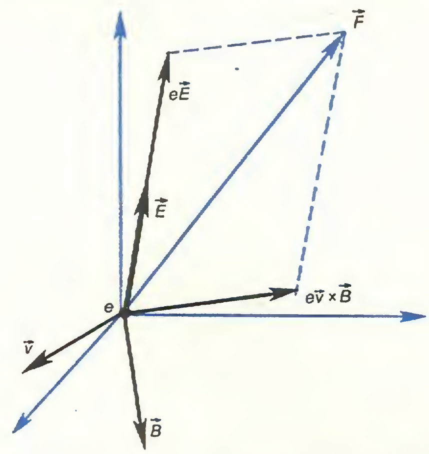
Elektrodynamika jest teorią pola elektromagnetycznego i ładunków elektrycznych. Wiemy, że między ładunkami elektrycznymi działają siły. Wyrażenie opisujące te siły jest dość zawiłe, więc dla wygody wprowadzono pojęcie pola elektromagnetycznego.

siła działająca na ładunek znajdujący się w polu elektromagnetycznym

Wyobraźmy sobie dowolny układ ładunków elektrycznych, w którym ładunki mogą się poruszać. Układ taki wytwarza wokół siebie pole elektryczne i magnetyczne. Pola te charakteryzujemy przez podanie dwóch wektorów: $\vec{E}$ (natężenie pola elektrycznego) i $\vec{B}$ (indukcja pola magnetycznego). Istnienie pola elektromagnetycznego wokół ładunków oznacza, że na dowolny inny ładunek e działa siła:

$$\vec{F} = e\vec{E} + e\vec{v} \times \vec{B},$$

gdzie $\vec{v}$ oznacza prędkość ładunku. Wzór ten stanowi definicję wektorów $\vec{E}$ i $\vec{B}$ (rys. 1).



Rys. 1. Siła działająca na cząstkę o ładunku e poruszającą się z prędkością $\vec{v}$ jest sumą dwóch składowych $e\vec{E}$ i $e\vec{v} \times B$

Widzimy, że siła działająca na ładunek składa się z dwóch członów: pierwszy pochodzi od pola elektrycznego, drugi — od magnetycznego. Jeśli się ładunek e nie porusza, to działa na niego tylko siła ze strony pola elektrycznego, $e\vec{E}$. Jest to dobrze znana siła, za pomocą której definiuje się wektor natężenia pola elektrycznego

Jeśli się ładunek e porusza, działa na niego dodatkowo siła ze strony pola magnetycznego, $e\vec{v} \times \vec{B}$, nazywana siłą Lorentza. Jest ona prostopadła do kierunku prędkości ładunku i definiuje wektor indukcji pola magnetycznego.

równanie Newtona

Znając siłę działającą na cząstkę obdarzoną ładunkiem, można określić jej ruch. Wyznaczony on jest przez równanie Newtona:

$$\frac{d}{dt} m\vec{v} = e\vec{E} + e\vec{v} \times \vec{B},$$

gdzie m jest masą cząstki.

Jeśli przez m rozumiemy zależną od prędkości masę $m = m_0/\sqrt{1-v^2/c^2}$, a przez m_0 — masę spoczynkową cząstki, to równanie powyższe opisuje poprawnie ruch cząstki o prędkości zbliżonej do prędkości światła.

Z tej dość zawiłej postaci równania ruchu wynika, że w zależności od konkretnej postaci pól $\vec{E}$ i $\vec{B}$ ruch ładunków odbywa się po różnych, często bardzo skomplikowanych krzywych. Powyższe równanie ruchu opisuje więc olbrzymie bogactwo zjawisk zachodzących np. w rozrzedzonej plazmie.

pole elektromagnetyczne

Zajmijmy się teraz dokładniej samym polem elektromagnetycznym. Zastanówmy się, czy w ogóle konieczne jest wprowadzanie tego pojęcia. W istocie nie powiedzieliśmy nic ponad to, że ładunki działają na siebie siłami. Przyjmowanie, że jeden ładunek wytwarza pole elektryczne, a drugi doznaje ze strony pola działania siły, może się wydać sztuczną komplikacją. Podobnie jest z siłami magnetycznymi: przewodnik, przez który płynie prąd, działa siłą na poruszające się

ładunki lub na igłę magnetyczną, ale czy koniecznie trzeba tu mówić, że prąd elektryczny powoduje powstanie pola magnetycznego?

Pole elektromagnetyczne, pojęcie wprowadzone przez M. Faradaya (1830 r.) w sposób trochę sztuczny, jako pewna interpretacja oddziaływań elektromagnetycznych ładunków elektrycznych, okazało się jednak niesłychanie głębokie. Pełne zrozumienie konsekwencji płynących z postulatu istnienia pola elektromagnetycznego zawdzięczamy J. C. Maxwellowi (1861 r.). Maxwell uogólnił wcześniej znane wyniki Faradaya, Ampera i in. oraz sformułował słynne równania, zwane równaniami Maxwella, opisujące pole elektromagnetyczne.

równania Maxwella

Pełny układ równań Maxwella jest następujący:

$$\varepsilon_0 \operatorname{div} \vec{E} = \varrho, \tag{1}$$

$$\operatorname{rot} \vec{E} = -\mu_0 \frac{\partial \vec{B}}{\partial t}, \tag{2}$$

$$\operatorname{div} \vec{B} = 0, \tag{3}$$

$$\operatorname{rot} \vec{B} = \varepsilon_0 \frac{\partial \vec{E}}{\partial t} + \vec{j}, \tag{4}$$

gdzie $\vec{E}$ oznacza wektor natężenia pola elektrycznego, $\vec{B}$ — wektor indukcji pola magnetycznego, ε_0 — przenikalność dielektryczną próżni, μ_0 — przenikalność magnetyczną próżni, ϱ — gęstość ładunku elektrycznego, $\vec{j}$ — gęstość prądu elektrycznego.

Treść fizyczna równań Maxwella jest następująca: Z pierwszego równania wynika, że strumień pola elektrycznego przechodzącego przez powierzchnię otaczającą ładunek jest proporcjonalny do wartości tego ładunku. Ponadto strumień ten nie zależy od wielkości powierzchni ani od jej odległości od ładunku. Innymi słowy, liczba linii sił pola elektrycznego przechodzącego przez powierzchnię otaczającą ładunek jest stała. Jeśli będziemy oddalać tę powierzchnię od ładunku nie zmieniając jej kształtu, to jej pole będzie rosnąć proporcjonalnie do kwadratu odległości od ładunku. A zatem pole elektryczne ładunku musi być odwrotnie proporcjonalne do kwadratu odległości. Prawidłowość ta zwana jest prawem Gaussa.

Z drugiego równania Maxwella wynika, że zmienne w czasie pola magnetyczne wytwarza wirowe pole elektryczne. Prawo to znane jest też jako prawo indukcji Faradaya.

Z trzeciego równania Maxwella wynika, że w przyrodzie nie istnieją ładunki magnetyczne. Linie sił pola magnetycznego są zawsze zamknięte — w przeciwieństwie do linii sił pola elektrycznego, które mogą się zaczynać i kończyć na ładunkach elektrycznych. Ten brak symetrii między polem elektrycznym i magnetycznym stał się punktem wyjścia hipotezy istnienia monopoli magnetycznych.

Czwarte równanie Maxwella ma najbardziej złożoną postać. Powiada ono, że źródłami pola magnetycznego są: zmiana w czasie pola elektrycznego i prąd elektryczny.

Nie możemy tu rozważać wszystkich wniosków płynących z tych równań, powiemy jedynie, że równania Maxwella pozwalają znaleźć pole elektromagnetyczne $\vec{E}$ i $\vec{B}$ wówczas, gdy znane jest położenie i prędkość ładunków. I tak np., gdy szukamy pola elektromagnetycznego w otoczeniu pojedynczego spoczywającego punktowego ładunku elektrycznego e, to równanie Maxwella dają:

prawo Coulomba

$$\vec{E}(\vec{r}) = \frac{e}{4\pi\varepsilon_0 r^2} \frac{\vec{r}}{r}, \quad B = 0.$$

Jest to znane prawo Coulomba.

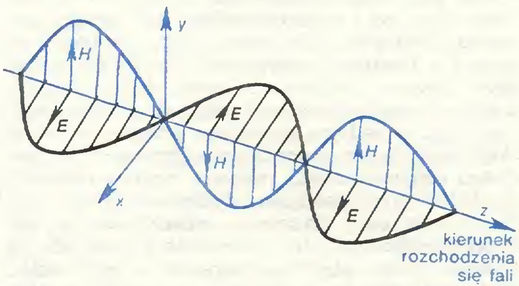
Wokół poruszających się ładunków oprócz pola elektrycznego istnieje również pole magnetyczne. Równania Maxwella pozwalają wyznaczyć to pole i odtworzyć np. prawa magnetycznego działania prądów

Omówiliśmy dwa odrębne zagadnienia: jakie pole elektromagnetyczne powstaje wokół ładunków oraz jak się poruszają ładunki w danym polu elektromagnetycznym. Pełna teoria elektromagnetyzmu powinna opisywać oba te procesy łącznie: poruszające się ładunki wytwarzają pola, które z kolei wpływają na ich ruch. A więc konieczna jest teoria samozgodna, opisująca ruch ładunków w polach, które są wytworzone przez te same ładunki. Warunki te spełniają równania Maxwella oraz równania ruchu ładunków Newtona, traktowane jako jeden układ równań. Pola powodujące ruch ładunków powstają w wyniku istnienia i ruchu ładunków, a ruch ładunków odbywa się pod wpływem sił pola elektromagnetycznego.

Równania Maxwella i równania ruchu ładunku w polu elektromagnetycznym stanowią komplet praw elektrodynamiki w próżni.

fale elektromagnetyczne

Olbrzymim osiągnięciem teorii Maxwella było to, że przewidziała ona istnienie fal elektromagnetycznych (rys. 2). O ich istnieniu może przekonać następujące rozumowanie: Z równań Maxwella wynika, że



Rys. 2. W fali elektromagnetycznej źródłem pola elektrycznego są zmiany pola magnetycznego, a źródłem pola magnetycznego — zmiany pola elektrycznego

w obszarach, w których nie ma ładunków ani prądów elektrycznych, pola elektryczne i magnetyczne spełniają równania falowe:

$$-\frac{1}{c^2}\frac{\partial^2 \vec{E}}{\partial t^2} + \Delta \vec{E} = 0, \quad -\frac{1}{c^2}\frac{\partial^2 \vec{B}}{\partial t^2} + \Delta \vec{B} = 0,$$

gdzie $c = \sqrt{\varepsilon_0 \mu_0}$.

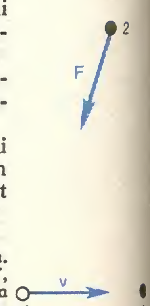
Z tych równań wynika, że w próżni pola elektryczne i magnetyczne mogą istnieć w postaci fal rozchodzących się z prędkością c. Doświadczalne potwierdzenie istnienia fal elektromagnetycznych jest dziełem H. Hertza, który też zbadał ich podstawowe własności. Wkrótce też stwierdzono, że fale świetlne są falami elektromagnetycznymi o dużej częstości. Tak to dokonano syntezy dwóch dziedzin: nauki o elektryczności i magnetyzmie z optyką. Należy podkreślić, że równania Maxwella poprawnie opisują zjawiska optyczne — takie jak dyfrakcja i interferencja fal świetlnych.

prędkość fal elektromagnetycznych

Fale elektromagnetyczne poruszają się z olbrzymią prędkością $(2{,}997924562 \pm 0{,}000000011) \cdot 10^8$ m·s^{-1} w próżni, przy czym prędkość ta nie zależy od długości fali. Z taką prędkością rozchodzą się wszystkie zaburzenia pola elektromagnetycznego. Skończona prędkość rozchodzenia się zaburzeń pola elektromagnetycznego stała się punktem wyjścia szczególnej teorii względności (→ O niektórych podstawowych pojęciach fizycznych).

Podamy kilka prostych, ale fundamentalnych wniosków wynikających ze skończonej prędkości rozchodzenia się zaburzeń pola elektromagnetycznego.

Wyobraźmy sobie dwa poruszające się ładunki (rys. 3). Ładunek *2* znajduje się w polu wytworzonym przez ładunek *1*. Ale wartość pola w punkcie *2* jest



Rys. 3. Siły elektryczne działające między ładunkami w ruchu. Siła działająca na ładunek *2* jest skierowana w stronę punktu *1'*, w którym znajdował się ładunek *1* we wcześniejszej chwili. Zatem siły elektromagnetyczne nie są centralne

taka, jaką wytworzył ładunek *1* będąc we wcześniejszej chwili w punkcie *1'*. W efekcie siła działająca na ładunek *2* nie jest skierowana do ładunku *1*, ale w kierunku *1'*. A więc siły elektryczne działające między ładunkami w ruchu nie są centralne i ich suma algebraiczna nie jest równa zeru. Jak wiemy (→ Zasady zachowania), oznacza to, że pęd układu ładunków nie jest zachowany. I tak doszliśmy do paradoksu, że w izolowanym układzie dwóch ładunków pęd nie jest zachowany.

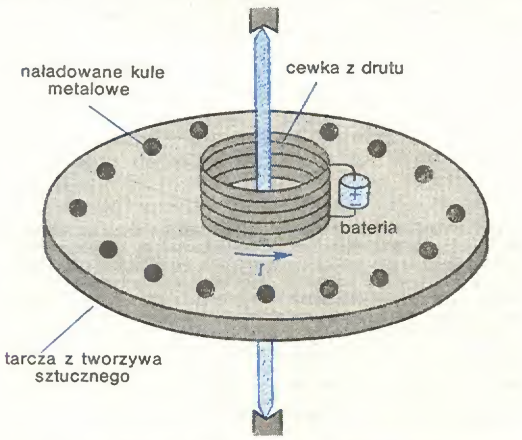
pęd pola elektromagnetycznego

Paradoks można usunąć, ale trzeba przyjąć, że pole elektromagnetyczne wytwarzane przez ładunek ma pewien pęd. Pęd pola elektromagnetycznego dokładnie kompensuje nadmiar pędu ładunków. Tak więc pęd samych ładunków nie jest wielkością stałą (nie jest zachowany), ale pęd układu składającego się z ładunków i pola elektromagnetycznego jest zachowany.

moment pędu pola elektromagnetycznego

Z doświadczenia przedstawionego na rys. 4 wynika, że pole elektromagnetyczne ma pewien moment pędu.

Powyższe rozważania wykazują, że pole elektromagnetyczne nie jest jedynie abstrakcyjnym tworem,



Rys. 4. Na obwodzie tarczy znajdują się naładowane, izolowane kulki. Współosiowo z tarczą umieszczony jest solenoid, przez który płynie prąd z baterii. Gdy przepływ prądu ustanie, tarcza zacznie się obracać. Moment pędu zostanie przekazany tarczy przez pole elektromagnetyczne wytworzone przez solenoid (wg R. P. Feynmana)

służącym do opisu oddziaływań ładunków. Jest ono obiektem materialnym, mającym określone parametry, jak energię, pęd i moment pędu. Materialność pola występuje jeszcze wyraźniej w elektrodynamice kwantowej, gdzie polu elektromagnetycznemu przypisujemy własność cząstek, kwantów γ (fotonów), i obserwujemy przemianę kwantów γ w inne cząstki.

Zauważyliśmy, że spoczywający ładunek wytwarza pole elektryczne, natomiast poruszający się ładunek jest źródłem również pola magnetycznego. Ale prędkość ładunku zależy od wyboru układu odniesienia. Jeśli więc w układzie, w którym ładunek spoczywa, pole magnetyczne nie występuje, to w innym układzie, poruszającym się ruchem prostoliniowym jednostajnym względem pierwszego, wystąpi niezerowe pole magnetyczne. Podobnie jeśli w jednym układzie występuje tylko pole magnetyczne, to w innym wystąpi również pole elektryczne. Tak więc pole elektryczne i magnetyczne nie mają charakteru bezwzględnego, lecz są składowymi jednej wielkości fizycznej — pola elektromagnetycznego; podział tego pola na pole elektryczne i magnetyczne zależy od wyboru układu.

Warto zwrócić uwagę na fakt, że rozważania tego rodzaju stały się punktem wyjścia dla sformułowania szczególnej teorii względności. Aby prawa elektrodynamiki nie zależały od wyboru inercjalnego układu odniesienia, transformacja między układami musi być transformacją Lorentza. Należało więc tak zmodyfikować prawa mechaniki klasycznej, aby teoria ruchu ładunków i pola elektromagnetycznego nie zależała od wyboru układu odniesienia.

Co to jest elektrodynamika kwantowa

Oddziaływanie pola elektromagnetycznego z materią, szczególnie zaś fal o wielkiej częstości z atomami, jest bardziej złożone, niż to wynika z klasycznych równań Maxwella. Nie ma w tym nic dziwnego; atomy nie są obiektami klasycznymi i do analizy ich zachowania się koniecznie trzeba stosować opis kwantowy. Zatem analiza oddziaływania pola elektromagnetycznego z atomami także wymaga spójnego kwantowego opisu. Teoria kwantowa opisująca oddziaływanie pola elektromagnetycznego z atomami nosi nazwę elektrodynamiki kwantowej. Jej początki datują się na rok 1900, kiedy to M. Planck ogłosił teorię promieniowania ciała doskonale czarnego; pełne sformułowanie zasad elektrodynamiki kwantowej nastąpiło na przełomie lat czterdziestych i pięćdziesiątych, w pracach R. Feynmana i J. Schwingera.

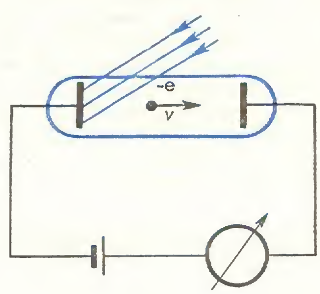
Trudno jest w tym miejscu podać choćby przybliżony zarys formalizmu matematycznego używanego w elektrodynamice kwantowej (→ Teoria pola). Ograniczymy się do podania najważniejszych cech fizycznych odróżniających elektrodynamikę kwantową od elektrodynamiki klasycznej.

elektrodynamika kwantowa a klasyczna

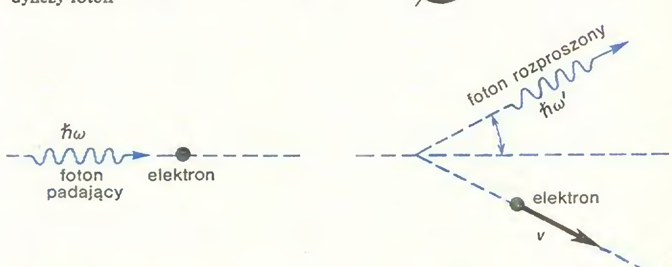
Z elektrodynamiki klasycznej wynika, że przekazywanie energii między układami ładunków i polem elektromagnetycznym ma charakter ciągły. Poruszające się ładunki wysyłają lub pochłaniają promieniowanie elektromagnetyczne; ilość energii pochłoniętej lub wysłanej zależy od ruchu ładunków i może przybierać dowolną wartość. Taki wynik jest zgodny z faktami doświadczalnymi w dziedzinie fal długich — radiowych czy mikrofal. Jednakże w dziedzinie fal krótkich, np. rentgenowskich, wynik ten nie jest zgodny z obserwacjami. Oddziaływanie pola elektromagnetycznego o wielkiej częstości z atomami ma charakter wybitnie nieciągły. Okazało się, że atomy nie mogą przekazać dowolnej ilości energii polu elektromagnetycznemu, lecz tylko określoną jej ilość, równą różnicy energii poziomów atomu, między którymi zachodzi przejście. Energia ta pojawia się jako najmniejsza ilość pola elektromagnetycznego — foton.

fotony

Oddziaływanie pola elektromagnetycznego z ładunkami polega na emisji i absorpcji fotonów, odbywa się więc porcjami, w sposób nieciągły. Jak wiadomo, tylko takie podejście prowadzi do poprawnego, to znaczy zgodnego z doświadczeniem opisu takich procesów jak zjawisko fotoelektryczne, zjawisko Comptona i in. (rys. 5, 6).



Rys. 5. Zjawisko fotoelektryczne polega na wybijaniu elektronów z katody przez wiązkę światła. Poprawną zależność natężenia w obwodzie od natężenia światła i jego częstości otrzymuje się przy założeniu, że elektron zostaje wybity z metalu przez pojedynczy foton



Rys. 6. Zjawisko Comptona — foton o częstości ω „zderza się" z elektronem i w wyniku tego powstaje foton o innej częstości ω'

Elektrodynamika kwantowa przypisuje więc polu elektromagnetycznemu pewne właściwości cząstkowe. Fale elektromagnetyczne są interpretowane jako strumień cząstek — fotonów. Fala o możliwie najmniejszym natężeniu zawiera jeden foton.

Fotony pod niektórymi względami różnią się zasadniczo od innych „zwykłych" cząstek. Fotony mają zerową masę spoczynkową. Mimo to niosą one ze sobą strumień energii, pędu oraz momentu pędu.

Takie właściwości fotonów są zgodne z mechaniką relatywistyczną. Związek między energią E i pędem cząstki o masie spoczynkowej m jest następujący $E=\sqrt{m^2c^4+p^2c^2}$. Jeśli masa spoczynkowa jest równa zeru, to $E=pc$. Związek ten oznacza, że fotonów nie można zatrzymać, innymi słowy — fale elektromagnetyczne nie mogą być nieruchome. Gdyby bowiem pęd fotonu był zerem, to jego całkowita energia byłaby zerowa i foton po prostu nie istniałby.

zerowa masa spoczynkowa fotonu

O zerowej masie spoczynkowej fotonu możemy się też przekonać na podstawie innego rozumowania. Jak wspomnieliśmy, z praw Maxwella wynika, że pole elektryczne i pole magnetyczne w pełni spełniają równania falowe. Równanie falowe jest identyczne z równaniem pola cząstki o zerowej masie (→ Teoria pola). Dlatego też założenie, że fotony, czyli cząstki związane z polem elektromagnetycznym, mają zerową masę, jest konieczne do tego, aby uzyskać zgodność praw elektromagnetyzmu za pomocą teorii pól i z uwzględnieniem istnienia fotonu.

Istotną kwestią jest zrozumienie, jak elektrodynamika kwantowa opisuje zjawiska falowe takie jak dyfrakcja i interferencja fal. Jednoczesny opis zjawisk falowych i cząstkowych jest zgodny z duchem teorii kwantowych; inne cząstki, np. elektrony, również wykazują jednocześnie właściwości cząstkowe i falowe (→ O niektórych podstawowych pojęciach fizycznych).

Właściwości cząstkowe fotonów ujawniają się wyraźnie wtedy, gdy mamy do czynienia z bardzo małą liczbą fotonów — możemy wtedy mówić o reakcji pojedynczego fotonu z pojedynczym elektronem. Tak jest w wiązkach fal o dużej częstości, np. fal ultrafioletowych czy rentgenowskich; w tym zakresie fal najczęściej obserwujemy efekty kwantowe. Kiedy rozważamy wiązki fal elektromagnetycznych o dużym natężeniu, w reakcji bierze udział z reguły wielka liczba fotonów. Właściwości cząstkowe pola elektromagnetycznego zacierają się, ujawniają się natomiast wyraźnie efekty falowe.

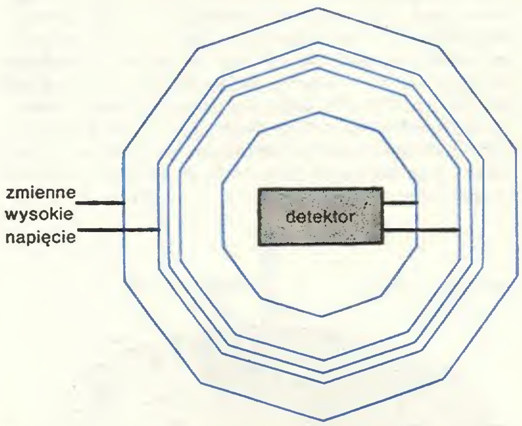
Ciekawa sytuacja występuje w optyce. Wiązki światła o częstości widzialnej uzyskiwane z laserów mogą mieć bardzo duże natężenie, mogą nieść wiele fotonów. Z drugiej jednak strony energia pojedynczego fotonu może być bardzo bliska różnicy dwóch poziomów energetycznych atomu, co sprzyja reakcji atomu z pojedynczym fotonem. Tak więc w oddziaływaniach wiązek laserowych z atomami ujawniają się często zarówno cząstkowe, jak i falowe właściwości pola elektromagnetycznego.

potwierdzenie słuszności prawa Gaussa

Dlaczego uważamy elektrodynamikę za najdoskonalszą teorię fizyczną

Klasyczna i kwantowa teoria elektromagnetyzmu została stworzona na podstawie szeregu doświadczeń i pomiarów. Były to przede wszystkim mało precyzyjne doświadczenia dziewiętnastowieczne. Współcześnie wykonuje się liczne eksperymenty i pomiary mające na celu dokładniejsze zbadanie zakresu stosowalności elektrodynamiki. Opiszemy tu kilka z nich. Zaczniemy od doświadczeń potwierdzających słuszność elektrostatyki i magnetostatyki. Dotyczą one samych podstaw elektrodynamiki.

Według jednego z ważniejszych twierdzeń elektrostatyki — we wnętrzu naładowanej powierzchni pole elektryczne znika. Twierdzenie to wynika bezpośrednio z pierwszego równania Maxwella, nazywanego też prawem Gaussa. Dlatego badanie, czy istnieje pole wewnątrz naładowanej powierzchni, jest bezpośrednim sprawdzianem prawa Gaussa. Dokładność, z jaką można stwierdzić, że natężenie pola jest zerowe, ogranicza dokładność sprawdzenia słuszności prawa Gaussa. Pierwsze doświadczenie tego typu wykonał H. Cavendish w 1773 r. Schemat jego współczesnej wersji, opracowanej w 1971 r. przez E.R. Williamsa, J.E. Fallera i H.A. Hilla w USA, jest przedstawiony na rys. 7. Doświadczenia tego typu pozwalają oszacować słuszność prawa Gaussa, czy też wynikającego



Rys. 7. Współczesna aparatura do sprawdzania prawa Gaussa składa się z pięciu metalowych powierzchni w kształcie dwudziestościanów. Do dwóch zewnętrznych powierzchni przykłada się zmienne wysokie napięcie, które powoduje powstawanie ładunku na tych powierzchniach. Gdyby pole elektryczne istniało we wnętrzu powierzchni metalowych, to powodowałoby ono powstanie napięcia między dwiema wewnętrznymi powierzchniami. Celem doświadczenia jest stwierdzenie, czy istnieje między powierzchniami wewnętrznymi zmienne napięcie o tej samej częstości, co napięcie przyłożone do powierzchni zewnętrznych. Obecność takiego napięcia świadczyłaby o istnieniu pola elektrycznego wewnątrz powierzchni metalowych

z niego prawa Coulomba. Gdyby np. pole ładunku punktowego nie było odwrotnie proporcjonalne do kwadratu odległości od ładunku, lecz było proporcjonalne do e^{-br}/r^2, to wynik cytowanego doświadczenia nakładałby na parametr b ograniczenie $b<2\cdot10^{-8}$ cm^{-1}. A zatem prawo Coulomba jest słuszne przy odległościach r mniejszych od $b^{-1}=5\cdot10^7$ cm, wtedy bowiem czynnik e^{-br} jest z dobrym przybliżeniem równy jedności.

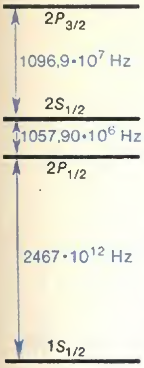
badanie pól magnetycznych

Badanie pól magnetycznych dostarcza też informacji na temat słuszności praw Maxwella. Pola magnetyczne są o tyle łatwe do badania, że w przyrodzie występują one w naturalny sposób w olbrzymiej skali; mowa tu oczywiście o polu magnetycznym Ziemi i innych planet. Pomiary linii sił pola magnetycznego dostarczają informacji, czy pole wokół magnesu, którym jest planeta, spełnia równania Maxwella, czy nie. Bardzo subtelnych pomiarów pola magnetycznego ziemskiego dokonują sztuczne satelity. Pozwoliły one stwierdzić słuszność praw Maxwella w odległościach rzędu 10^9 m.

Doświadczenia te, potwierdzające słuszność elektrodynamiki klasycznej, stanowią też sprawdzenie elektrodynamiki kwantowej. W teorii kwantowej nabierają one nowej interpretacji, nakładają ograniczenia na ewentualną masę fotonów. Wynika z nich, że masa fotonu wynosi mniej niż 10^{-47} kg (dla porównania: masa elektronu, najlżejszej cząstki, wynosi ok. 10^{-30} kg).

pomiary przesunięcia Lamba

Dalsze doświadczenia dotyczą typowo kwantowych efektów. Jednym z bardzo subtelnych efektów przewidzianych przez elektrodynamikę kwantową jest istnienie tak zwanego przesunięcia Lamba. Jest to różnica między poziomami energetycznymi w atomie wodoru oznaczanymi przez $2S_{1/2}$ i $2P_{1/2}$ (rys. 8). Gdyby nie kwantowa natura promieniowania elek-



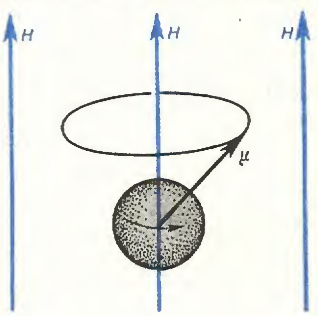
Rys. 8. Struktura stanów energetycznych atomu wodoru. Przesunięcie Lamba to różnica poziomów $2S_{1/2}$ i $2P_{1/2}$

tromagnetycznego, stany te miałyby tę samą energię. Dlatego też pomiary przesunięcia Lamba i porównywanie wielkości eksperymentalnych z teoretycznymi mają podstawowe znaczenie w kwantowej teorii pola elektromagnetycznego. Pierwsze obliczenie tego efektu wykonał w 1947 r. H.A. Bethe, a pomiary zostały w tym samym roku wykonane przez W.E. Lamba i R.C. Retherforda. Przy pomiarach autorzy wykorzystali rezonans magnetyczny między mierzonymi poziomami (→ Spektroskopia atomowa), co na owe czasy było bardzo nowoczesnym pomysłem. Ponieważ częstości radiowe, których tu użyto, dają się mierzyć z bardzo dużą dokładnością, również przesunięcie poziomów można w ten sposób wyznaczyć dokładnie.

Już pierwsze pomiary wykazały bardzo dobrą zgodność wyników doświadczalnych z teoretycznymi. Obecnie zarówno dokładność obliczeń, jak pomiarów niepomiernie wzrosła. Pomiary przesunięcia Lamba wykonali ostatnio R. Robiscoe i T. Shyn z USA (1970); należy podkreślić, że nie stwierdzono tu żadnych odstępstw od wartości teoretycznej, wynoszącej (1057,911±0,12) MHz.

obliczenie momentu magnetycznego elektronu

Jednym z piękniejszych osiągnięć potwierdzających słuszność elektrodynamiki kwantowej jest obliczenie momentu magnetycznego elektronu. W myśl elektrodynamiki kwantowej elektron, mimo że jest cząstką punktową, zachowuje się podobnie jak obracająca się naładowana kulka. Taka obracająca się kulka to jakby igła magnetyczna — oś obrotu kulki wykonuje precesję w polu magnetycznym wokół linii sił pola (rys. 9). Częstość tej precesji zależy od wartości momentu magnetycznego. Tak więc elektron ma moment magnetyczny μ. Teoretyczne obliczenie tego momentu wykonał P.A.M. Dirac, otrzymał on wynik $\mu = 2(e\hbar/mc)$. Dokładniejsze wyniki uwzględniające kwantową strukturę pola elektromagnetycznego uzyskał J. Schwinger (1948), a później inni autorzy.



Rys. 9. Elektron w polu magnetycznym jest w pewnym sensie podobny do obracającej się naładowanej kuleczki, wykonuje precesję wokół pola magnetycznego. Obliczenie częstości precesji, a zatem momentu magnetycznego elektronu to jedno z największych osiągnięć elektrodynamiki kwantowej

pomiar momentu magnetycznego elektronu

Pomiarów momentu magnetycznego elektronu dokonywały liczne grupy badaczy, ale wszystkie doświadczenia opierały się na tej samej zasadzie: spolaryzowane elektrony (tzn. o równoległych momentach magnetycznych) przelatywały przez pole magnetyczne, mierzony był kąt skręcenia momentów magnetycznych.

Jednym z najdokładniejszych wyników było $\mu = (2{,}0011596577 \pm 0{,}0000000035) e\hbar/mc$; jest on zgodny z obliczeniami teoretycznymi, a uzyskali go J.C. Wesley i A. Rich w 1971 r.

Mierzono także momenty magnetyczne innych cząstek — pozytonów i mionów, a wyniki porównywano z wartościami obliczonymi teoretycznie. Odstępstw od nich nie stwierdzono, ale wyniki pomiarów były znacznie mniej dokładne.

Wykazaliśmy na przykładach, że elektrodynamika jest teorią opisującą z wielką precyzją szeroki zakres zjawisk fizycznych. O innych zastosowaniach elektrodynamiki w różnych dziedzinach fizyki można znaleźć informacje w innych hasłach niniejszej Encyklopedii. Przykłady te dobitnie dowodzą, że elektrodynamika jest teorią o niesłychanie szerokim zakresie stosowalności.

Ale nie tylko zgodność z eksperymentem świadczy o doskonałości elektrodynamiki. Ma ona również daleko idące zalety formalne. Na podobieństwo elektrodynamiki, szczególnie kwantowej, próbowano formułować teorie innych oddziaływań elementarnych (→ Cząstki elementarne i ich oddziaływania). Pewne sukcesy odniesiono przy takim podejściu do opisu oddziaływań słabych. Powstała w ten sposób teoria Fermiego rozpadu β, a ostatnio bardzo burzliwie rozwijają się tzw. zunifikowane teorie oddziaływań elektromagnetycznych i słabych. Metody zbliżone do użytych w elektrodynamice kwantowej próbowano też stosować do opisu oddziaływań silnych, w szczególności oddziaływań pionów (→ Oddziaływania silne).

Wewnętrzne sprzeczności elektrodynamiki

Pomimo, że elektrodynamika dobrze wyjaśnia wiele zjawisk, nie można zapominać o zasadniczych jej niekonsekwencjach. W samym bowiem sformułowaniu elektrodynamiki tkwią wewnętrzne sprzeczności, o których istocie tu powiemy.

energia ładunku punktowego

Zastanówmy się, jaka jest energia pola elektrycznego w otoczeniu ładunku e w kształcie kuli o promieniu a. Ponieważ pole elektryczne w punkcie odległym o r od środka kuli wynosi $E = e/4\pi\varepsilon_0 r^2$, to energia pola dana jest przez wzór $W = {}^1/_2 \int E^2 dV$, przy czym po wykonaniu łatwego rachunku otrzymujemy $W = e^2/\varepsilon_0^2 a$. Energia kuli jest odwrotnie proporcjonalna do jej promienia. Na pierwszy rzut oka nic nas tu nie razi. Trudność zaczyna się wtedy, gdy rozważany ładunek jest punktowy, to znaczy, gdy jego promień jest zerem. Wtedy bowiem energia pola elektrycznego otaczającego ładunek staje się nieskończenie wielka. Ta „nieskończoność" jest podstawową trudnością elektrodynamiki. Nie można się jej pozbyć zakładając, że w przyrodzie nie występują ładunki punktowe, mamy bowiem powody uważać elementarne ładunki — elektrony, za cząstki punktowe. Gdyby elektron nie był punktowy, lecz był jakąś kulką o skończonym promieniu, trudno byłoby go uznać za elementarną cząstkę. Wszystko świadczy o tym, że elektron jest cząstką punktową.

nadanie elektronowi przyspieszenia

Nieskończoność, o której mówiliśmy, przejawia się i w innych, bardziej jeszcze drastycznych miejscach. Jeśli np. chcemy nadać elektronowi przyśpieszenie, to oprócz pokonania zwykłej bezwładności elektronu, pochodzącej od jego masy, musimy pokonać bezwładność pola elektrycznego otaczającego elektron. Niestety — siła, jakiej trzeba użyć do pokonania bezwładności pola, jest nieskończenie wielka. Wygląda to tak, jakby masa elektronu była nieskończenie wielka, nie można bowiem oddzielić elektronu od pola, które wytwarza. Wiemy jednak, że masa elektronu wynosi ok. $9 \cdot 10^{-31}$ kg i na pewno nie jest nieskończona.

Wydawać by się mogło, że trudności te powinny zniknąć w elektrodynamice kwantowej, która konsekwentnie uwzględnia wszelkie efekty kwantowe w małych odległościach od ładunku. Do pewnego stopnia istotnie tak jest. W elektrodynamice kwantowej można się nie posługiwać wielkościami nieskończonymi. Jednakże w takim sformułowaniu masa i ładunek elektronu nie mogą być obliczone w ramach teorii, lecz muszą być wzięte z doświadczenia.

Tak więc, mimo że potrafimy z wielką precyzją obliczać przebieg różnych procesów elektromagnetycznych, a wyniki naszych obliczeń zgodne są z doświadczeniami, to jednak nie bardzo rozumiemy, dlaczego się tak dzieje, nie rozumiemy bowiem do końca podstaw elektromagnetyzmu.

L.N. Cooper *Istota i struktura fizyki*, Warszawa 1975; R.P. Feynman i in. *Feynmana wykłady z fizyki*, Warszawa 1974; R.H. March *Fizyka dla poetów*, Warszawa 1974.

Teoria pola

Marian Kupczyński

Teoria pola jest jednym z najtrudniejszych i najciekawszych działów fizyki. Stosuje się w niej bardzo zaawansowane metody matematyczne, a przeprowadzane obliczenia są długie i żmudne. Należą do niej najwspanialsze osiągnięcia fizyki teoretycznej — ogólna teoria względności oraz elektrodynamika klasyczna i kwantowa (→ Elektrodynamika).

Na gruncie kwantowej teorii pola przewidziano istnienie antymaterii, a ostatnio cząstek powabnych (→ Cząstki elementarne i ich oddziaływania). Cała teoria cząstek elementarnych oparta jest na ideach kwantowej teorii pola. Metody teorii pola są ostatnio szeroko stosowane w fizyce statystycznej.

Kwantowa teoria pola nie jest teorią zamkniętą. Problemów nie rozwiązanych i trudnych jest bardzo wiele. Po ostatnich sukcesach nowej teorii oddziaływań słabych (→ Oddziaływania słabe) teoria pola jest działem fizyki, w którym są prowadzone najintensywniejsze badania teoretyczne.

Fizyka jest nauką, która szuka wyjaśnienia obserwowanych zjawisk (→ Czym jest fizyka). Podstawowe idee teorii fizycznych można sformułować dość prosto, podobnie jak prosto można sporządzić plan najtrudniejszej wspinaczki w Himalajach.

modele w fizyce

Typowe dla fizyki jest tworzenie modeli dwojakiego rodzaju. Pierwszy rodzaj to modele jakościowe zjawisk (obrazy) odwołujące się do intuicji i wyobraźni. Stają się one elementem języka, którym fizycy mówią potocznie o rzeczywistości. Drugi — to abstrakcyjne modele matematyczne, które umożliwiają obliczanie porównywalnych z doświadczeniem liczb. Zgodność przewidywań modelu z doświadczeniem decyduje o sukcesie lub niepowodzeniu modelu. Ten drugi rodzaj modeli to właściwe modele tworzone przez fizykę. Związek obrazu jakościowego z modelem ścisłym jest dość luźny. Często niewłaściwy model jakościowy bardzo utrudnia sformułowanie właściwego modelu abstrakcyjnego. Rozwój metod stosowanych w fizyce zmierza wyraźnie w kierunku coraz bardziej abstrakcyjnego opisu zjawisk fizycznych. Mimo wszystko tworzenie obrazów (które nie są rozumiane jak wierne odbicie procesów rzeczywistych) jest często pomocne w konstruowaniu czy modyfikowaniu schematu matematycznego.

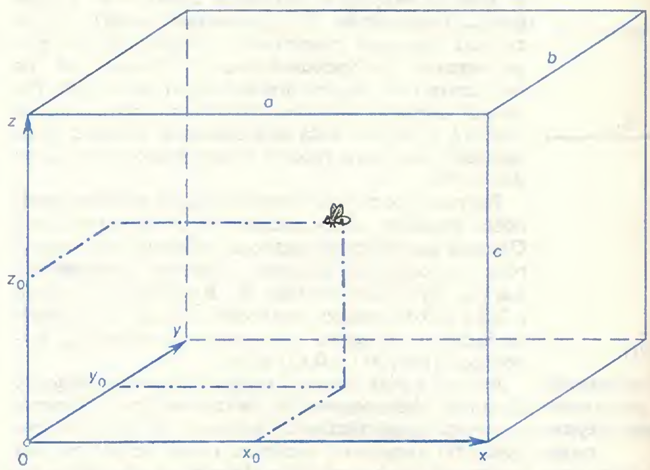
Tak więc omawiając teorię pola będziemy przedstawiać zarówno obrazy zjawisk jak i zarys schematu matematycznego teorii. Bez tego schematu byłaby to nie opowieść o teorii, lecz tylko bajka o niej.

O czasoprzestrzeni, polu temperatury i pochodnych cząstkowych

Podstawowym pojęciem w fizyce jest czasoprzestrzeń i od niej zaczniemy nasze rozważania o polu (→ O niektórych podstawowych pojęciach fizycznych). Czasoprzestrzeń jest to zbiór wszystkich możliwych zdarzeń. Zdarzeniem nazywamy takie wydarzenie, które można scharakteryzować przez podanie czterech liczb t, x, y, z, gdzie x, y, z są współrzędnymi wektora położenia $\vec{r}$, a t jest czasem, w którym to zdarzenie zachodzi. Na przykład (rys. 1) jeśli o godz. 15^{00} mała muszka siada w kącie prostopadłościennego pokoju o wymiarach $a \times b \times c$ i jeśli umówimy się, że określimy położenia przedmiotów podając ich odległość od trzech krawędzi ścian wychodzących z kąta wybranego przez naszą muszkę, to omawiane zdarzenie z jej życia w wybranym przez nas układzie odniesienia ma

czasoprzestrzeń i punkty materialne

współrzędne: 15^{00}, 0, 0, 0. Można by się dziwić, że wybraliśmy muszkę, a nie stół. Ale widać od razu, że położenia stołu nie można określić przez podanie trzech liczb. Stół w pokoju o godz. 15^{00} to zbiór bar-



Rys. 1. W pokoju o wymiarach $a \times b \times c$ znajduje się mucha. Ponieważ mucha nie jest punktowa, zaznaczamy na jej tułowiu punkt, którego położenie określa położenie muchy o godz. 15^{10}. Zdarzenie przedstawione na rysunku ma współrzędne (15^{10}, x_0, y_0, z_0)

dzo wielu zdarzeń w czasoprzestrzeni, z których każde jest położeniem odpowiedniego punktu stołu o tej godzinie. Małą muszkę możemy traktować jako punkt materialny, najprostszy obiekt fizyczny poruszający się w czasoprzestrzeni.

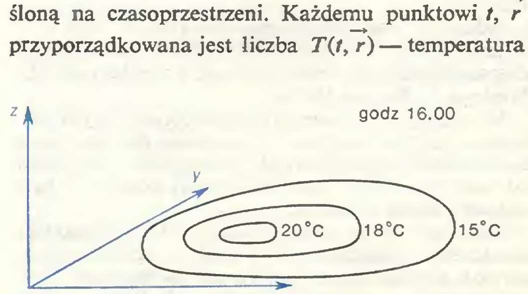
Reprezentowanie obiektu fizycznego przez punkt materialny okazuje się b. dobrą idealizacją w opisie wielu zjawisk. Ruch planet wokół Słońca i naładowanych cząstek elementarnych w zewnętrznym polu elektromagnetycznym można przedstawić jako ruch punktów materialnych. Wróćmy do ruchu punktu materialnego w czasoprzestrzeni. Ruch ten jest w pełni określony, jeśli znamy trzy funkcje zależne od czasu $x(t)$, $y(t)$, $z(t)$, tworzące razem funkcję o wartościach wektorowych $\vec{r}(t)$. Minimalna liczba zmiennych w czasie parametrów koniecznych do opisu układu fizycznego nazywa się liczbą stopni swobody układu. Tak więc punkt materialny (gdy nie ma więzów) ma trzy stopnie swobody.

stopnie swobody

pole temperatury

Wyobraźmy sobie teraz nieskończoną liczbę punktowych obserwatorów wyposażonych w punktowe i niezwykle czułe termometry i zsynchronizowane zegary, którzy umówili się notować mierzoną temperaturę w każdym punkcie przestrzeni w każdej chwili (czyli w każdym punkcie czasoprzestrzeni). Zebrane wyniki pomiarów można zapisać jako funkcję określoną na czasoprzestrzeni. Każdemu punktowi t, $\vec{r}$ przyporządkowana jest liczba $T(t, \vec{r})$ — temperatura





Rys. 2. Wybrane wartości pola temperatury $T(16^{00}, x, y, 0)$ przedstawione w postaci izoterm — krzywych, na których funkcja $T(t = \mathrm{const}, x, y, 0)$ jest stała

w punkcie $\vec{r}$ w chwili t (rys. 2). Mówimy, że $T(t, \vec{r})$ określa pole temperatury w przestrzeni. Załóżmy teraz, że pomiary są wykonywane na przykład w pokoju, w którym rozmieszczono odpowiednie grzejniki i termostaty utrzymujące na wszystkich ścianach, podłodze i suficie tę samą stałą temperaturę T_0. Zastanówmy się, jakie pytanie, czy też jakie zadanie może postawić sobie fizyk w tej sytuacji. Podobnie jak w mechanice punktów materialnych na podstawie znajomości początkowych położeń i prędkości chcemy przy pomocy równań Newtona wydedukować położenia punktów materialnych w przyszłości, tak w naszym problemie chcielibyśmy — na podstawie znajomości pola w chwili $t = 0$ i znajomości warunków doświadczalnych — umieć przewidywać pole temperatury w przyszłości. Aby to zadanie zrealizować, musimy dojść, jakie jest równanie pola, i rozwiązać je przy danych warunkach początkowych:

poszukiwanie równań pola

$$T(0, \vec{r}) = f(\vec{r}), \tag{1}$$

gdzie f jest znaną funkcją, oraz warunkach brzegowych:

$$T(t,0,y,z) = T(t,x,0,z) = T(t,x,y,0) = = T(t,a,y,z) = T(t,x,b,z) = T(t,x,y,c) = T_0. \tag{2}$$

Ponieważ pokój ma wymiary $a \times b \times c$, szukamy rozwiązania dla $\vec{r}$ takich, że:

$$0 \leqslant x \leqslant a, \quad 0 \leqslant y \leqslant b \quad \text{i} \quad 0 \leqslant z \leqslant c.$$

Poszukiwanie równań pola poprzedzimy przypomnieniem równań ruchu punktu materialnego (równań Newtona). Mają one dobrze znaną postać $\vec{F}(\vec{r},\vec{v},t) = m\vec{a}$. Prędkość $\vec{v}$ i przyspieszenie $\vec{a}$ wyrażają się przez pochodne względem czasu wektora położenia $\vec{r}(t)$, a mianowicie:

$$\vec{v} = \lim_{\Delta t \to 0} \frac{\vec{r}(t+\Delta t) - \vec{r}(t)}{\Delta t} \overset{\text{df}}{=} \frac{d\vec{r}}{dt}$$

oraz

$$\vec{a} = \lim_{\Delta t \to 0} \frac{\vec{v}(t+\Delta t) - \vec{v}(t)}{\Delta t} \overset{\text{df}}{=} \frac{d\vec{v}}{dt} = \frac{d^2\vec{r}}{dt^2}.$$

Po skorzystaniu z tych związków równanie Newtona można zapisać:

$$m\frac{d^2\vec{r}}{dt^2} = F\left(\vec{r}, \frac{d\vec{r}}{dt}, t\right). \tag{3}$$

Równanie to jest równaniem różniczkowym zwyczajnym drugiego rzędu. Jednoznaczne rozwiązanie tego równania otrzymuje się przy warunkach początkowych

$$\vec{r}\,(t=0) = \vec{r}_0 \quad \text{oraz} \quad \vec{v}\,(t=0) = \vec{v}_0.$$

Ponieważ pole $T(t,\vec{r})$ zależy od czterech zmiennych, w równaniach pola mogą się pojawić pochodne względem wszystkich tych zmiennych. Są to tzw. pochodne cząstkowe. Na przykład pochodna cząstkowa względem czasu t, oznaczana $\partial T/\partial t$, jest określona wzorem:

$$\frac{\partial T}{\partial t} = \lim_{\Delta t \to 0} \frac{T(t+\Delta t, x, y, z) - T(t,x,y,z)}{\Delta t}. \tag{4}$$

pochodne cząstkowe

Podobnie definiuje się pochodne cząstkowe względem innych zmiennych: $\partial T/\partial x$, $\partial T/\partial y$, $\partial T/\partial z$ oraz drugie pochodne cząstkowe (jest ich 10):

$$\frac{\partial^2 T}{\partial t^2}, \frac{\partial^2 T}{\partial x^2}, \frac{\partial^2 T}{\partial y^2}, \frac{\partial^2 T}{\partial z^2}, \frac{\partial^2 T}{\partial x \partial y}, \frac{\partial^2 T}{\partial y \partial z},$$
$$\frac{\partial^2 T}{\partial t \partial x}, \frac{\partial^2 T}{\partial t \partial y}, \frac{\partial^2 T}{\partial t \partial z}, \frac{\partial^2 T}{\partial x \partial z}.$$

Pochodne te określone są wzorem analogicznym do wzoru (4), w którym zamiast funkcji T występują funkcje (czterech zmiennych) $\partial T/\partial t$, $\partial T/\partial x$, $\partial T/\partial y$, $\partial T/\partial z$. Najogólniejsze równanie pola $T(t, \vec{r})$ mogłoby mieć postać:

$$F\left(\frac{\partial T}{\partial t}, \frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}, \frac{\partial T}{\partial z}, \text{wyższe pochodne cząstkowe}\right) = 0. \tag{5}$$

równanie pola

Równanie typu (5) jest równaniem różniczkowym cząstkowym. Aby opisać w pełni zmienność w czasie pola dla każdego ustalonego wektora $\vec{r}$, musimy znać funkcję $T(t, \vec{r})$. Ponieważ wszystkich wektorów $\vec{r}$ jest tyle, ile punktów w przestrzeni, czyli nieprzeliczalnie wiele (pole jest układem fizycznym o nieprzeliczalnej liczbie stopni swobody), więc warunek (1) jest pojęciowo równoważny daniu położenia początkowego dla punktu materialnego. Ponieważ chcemy, aby warunek początkowy (1) wystarczał do wyznaczenia rozwiązania (w przyszłości), równanie (5) może zawierać pochodne cząstkowe względem czasu tylko pierwszego rzędu. Dokładna analiza problemu prowadzi do równania na pole temperatury:

$$\frac{1}{\varkappa}\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2},$$

gdzie $\varkappa(t, x, y, z)$ jest pewną funkcją dodatnią charakteryzującą ośrodek. Jest to równanie przewodnictwa cieplnego.

Innym rodzajem pola jest pole, które otrzymaliby punktowi obserwatorzy mierzący zamiast temperatury prędkość wiatru w punkcie $\vec{r}$. Prędkość wiatru $\vec{V}$ w punkcie $\vec{r}$ jest wektorem o trzech współrzędnych

$$\vec{V}(t, \vec{r}) = \left(V_x(t,\vec{r}), V_y(t,\vec{r}), V_z(t,\vec{r})\right).$$

pole prędkości wiatru

Informacja zebrana z całej przestrzeni z różnych chwil daje się zapisać jako pole prędkości $\vec{V}(t, \vec{r})$ w czasoprzestrzeni (w każdym punkcie zaczepiony jest odpowiedni wektor).

Przykład ten, podobnie jak poprzedni, nie da się zrealizować praktycznie. Nie jesteśmy w stanie nigdy zmierzyć temperatury i prędkości w punkcie. Mierzymy zawsze tylko pewne wartości średnie w odpowiednio małych obszarach przestrzeni. Pole jest wygodną idealizacją, która te małe obszary ściąga do pojedynczych punktów.

pole jako idealizacja

W ten sposób wprowadziliśmy pole jako pewien obiekt matematyczny, który służy do wygodnego zapisu obserwacji dokonywanych w czasoprzestrzeni. Teraz przejdziemy do pól, które są obiektami fizycznymi (mówiąc inaczej, są one formą istnienia materii).

O polach, które niosą energię, obiektach geometrycznych i zasadzie wariacyjnej

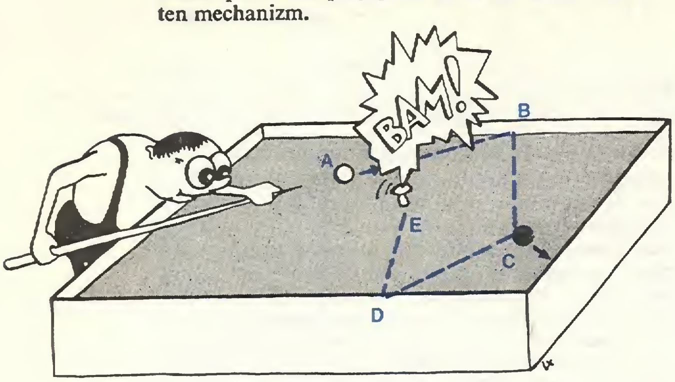
Spójrzmy na równanie (3); jeśli siła $\vec{F}$ działająca na obiekt materialny w punkcie $\vec{r}$ nie zależy od prędkości tego obiektu, to w każdym punkcie przestrzeni $\vec{r}$, w którym znajdzie się ten obiekt, mamy jednoznacznie określony wektor siły $\vec{F}(t, \vec{r})$, czyli pole sił. Pole $\vec{F}(t, \vec{r})$ zawiera informację o przyspieszeniu, jakie zostanie nadane cząstce o masie m w punkcie $\vec{r}$. W tym ujęciu pole $\vec{F}(t, \vec{r})$ nie jest traktowane jako obiekt fizyczny, lecz jako twór podobny do $T(t, \vec{r})$ i $\vec{V}(t, \vec{r})$.

pole sił

Istotną cechą równań Newtona jest ich nielokalność. Na przykład w problemie dwóch ciał przy opisie ruchu planety wokół Słońca przyspieszenie planety w chwili t zależy od położenia Słońca w tej samej chwili. Zmiana położenia Słońca w chwili $t+\Delta t$ zmienia siłę działającą na planetę w chwili $t+\Delta t$. Zdumiewające jest, że przyczyna i skutek są tak od-

nielokalność równań Newtona

dalone. Teoria Newtona zajmuje się tylko wyznaczaniem ruchu ciał bez wnikania w mechanizm uzyskiwania przez nie przyspieszeń. Spróbujemy wyjaśnić ten mechanizm.

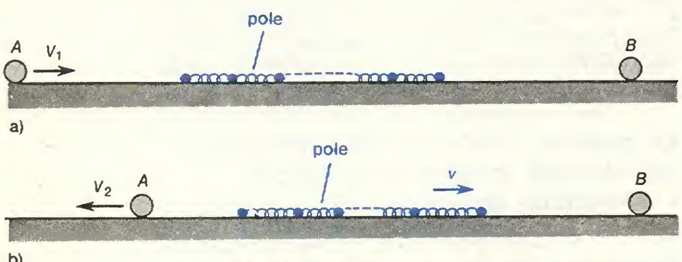


Rys. 3. Niefortunny przypadek na stole bilardowym. Litery A B, C, D i E oznaczają punkty, w których lokalnie następuje zmiana prędkości bili. Odcinki ruchu swobodnego zaznaczone są linią przerywaną

Doświadczenie dnia codziennego uczy, że najprościej jest zmienić prędkość ciała oddziałując na nie bezpośrednio. Wyobraźmy sobie np. kulę poruszającą się po stole bilardowym (rys. 3). Prędkość jej ulegnie zmianie wtedy i tylko wtedy, gdy uderzymy ją kijem, trafimy ją drugą kulą lub też ona sama wpadnie na przeszkodę, którą może być ściana stołu, druga kula lub grzybek. Nadanie przyspieszenia następuje jedynie w momencie odpowiedniego zderzenia i uzyskiwane jest w określonym punkcie stołu.

pole jako obiekt fizyczny

Naturalną tendencją jest próba redukcji nieznanego do czegoś prostego i dobrze rozumianego. Taką próbą jest wprowadzenie pola jako obiektu fizycznego, istniejącego niezależnie od cząstek, które odgrywa rolę niewidzialnego sprawcy uzyskiwanych przez te cząstki przyspieszeń. Ponieważ energia i pęd pojedynczych cząstek ulegają zmianie (a wiadomo, że całkowita energia i pęd układu powinny być zachowane), należy ze zmianą energii i pędu cząstki w punkcie $\vec{r}$ wiązać zmianę energii i pędu hipotetycznego pola. Ilustruje to jeszcze jeden mechaniczny model dokładniej opisany pod rys. 4.



Rys. 4. Mechaniczny model pola — zbiór bardzo małych mas połączonych sprężynkami: a) kula A z prędkością V_1 zbliża się do obszaru, w którym jest pole, b) widzimy kulę A ze zmienioną energią i pędem oraz pole, któremu energia i pęd zostały przekazane, poruszające się z prędkością $\vec{v}$ i drgające przy tym zawzięcie. Widać, że po pewnym czasie nastąpi oddziaływanie tego pola z kulą B

Program opisu oddziaływań cząstek za pośrednictwem pola został uwieńczony sukcesem — wspaniałym rozwojem elektrodynamiki (→ Elektrodynamika, Oddziaływania elektromagnetyczne), a następnie sformułowaniem einsteinowskiej teorii grawitacji. Przypomnimy tu kilka etapów tego rozwoju.

pole elektromagnetyczne

Podstawowe wielkości charakteryzujące pole elektromagnetyczne to $\vec{E}$ i $\vec{B}$, występujące we wzorze Lorentza na siłę działającą na ładunek q:

$$\vec{F} = q(\vec{E} + \vec{v} \times \vec{B}).$$

Natężenie pola elektrycznego $\vec{E}(\vec{r})$ oraz wektor indukcji magnetycznej $\vec{B}(\vec{r})$ są polami bezpośrednio mierzalnymi. W elektrostatyce w wypadku stacjonarnego pola $\vec{B}$ wprowadza się bezpośrednio niemierzalne pola $\varphi(\vec{r})$ i $\vec{A}(\vec{r})$, zw. potencjałem skalarnym i wektorowym. Związki między $\vec{E}$ i φ oraz między $\vec{B}$ i $\vec{A}$ są następujące:

$$\vec{E} = -\operatorname{grad}\varphi \stackrel{\mathrm{df}}{=} -\left(\frac{\partial\varphi}{\partial x}, \frac{\partial\varphi}{\partial y}, \frac{\partial\varphi}{\partial z}\right),$$

$$\vec{B} = \operatorname{rot}\vec{A} \stackrel{\mathrm{df}}{=} \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}, \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}, \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)$$

(dla pól zmiennych w czasie związki te mają bardziej skomplikowaną postać).

Okazuje się, że znajomość pola elektrostatycznego w całej przestrzeni dają pełną informację o rodzaju i rozmieszczeniu ładunków, a mianowicie gęstość rozkładu ładunku $\varrho(\vec{r})$ jest określona przez potencjał φ za pośrednictwem równania Poissona:

$$\Delta\varphi(\vec{r}) = -\frac{\varrho(\vec{r})}{\varepsilon} \qquad (6)$$

(ε — przenikalność elektryczna ośrodka). Gdy ładunek punktowy znajduje się w punkcie $\vec{R}$, φ ma osobliwość: $\varphi(\vec{r} = \vec{R}) = \infty$. Tak więc miejsce, w którym pole ma osobliwość, odpowiada miejscu, w którym znajduje się ładunek punktowy. Operacja Δ we wzorze (6) nazywa się laplasjanem i zdefiniowana jest jak następuje:

ładunki jako osobliwości pola

$$\Delta\varphi(\vec{r}) \stackrel{\mathrm{df}}{=} \frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} + \frac{\partial^2\varphi}{\partial z^2}$$

Jeszcze jedną operacją występującą w elektrodynamice jest dywergencja pola wektorowego:

$$\operatorname{div}\vec{E} \stackrel{\mathrm{df}}{=} \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}$$

fale elektromagnetyczne

Maxwell doszedł do swoich równan wyobrażając sobie pole elektromagnetyczne jako odkształcenie idealnego niewidzialnego sprężystego ośrodka zwanego eterem. Odkrycie przewidzianych przez jego teorię fal elektromagnetycznych i interpretacja światła jako fali elektromagetycznej to jedno z największych odkryć w fizyce. Subtelne doświadczenia doprowadziły do odrzucenia koncepcji eteru. Po odrzuceniu tej koncepcji w zasadzie nie mamy na gruncie elektrodynamiki klasycznej żadnego intuicyjnego obrazu rozchodzenia się fali elektromagnetycznej. Mimo to, potrafimy przewidywać wyniki eksperymentów optycznych ze światłem, zachowanie ładunków próbnych pod wpływem fali elektromagnetycznej i wiele innych efektów. Otrzymujemy to wszystko z abstrakcyjnego modelu matematycznego, którym są równania Maxwella. Spotyka się często sformułowanie „fala elektromagnetyczna rozchodzi się w próżni". Wydaje się wskazanym wyjaśnić, jak to należy rozumieć, by uniknąć zbędnych metafizycznych dyskusji. Próżnia w tym sformułowaniu jest rozumiana jako obszar przestrzeni, z którego usunięte zostały wszystkie cząstki materialne mające masę spoczynkową. Tak właśnie, rozrzedzając gaz do granic technicznych możliwości, otrzymuje się w technice stany wysokiej próżni. Usuwając cząstki nie usuwamy jednak pola, które może zmieniać swój stan w czasie. Jednym z rodzajów tych zmian jest fala elektromagnetyczna. Przeto zamiast mówić „fala elektromagnetyczna może rozchodzić się w próżni", lepiej jest powiedzieć: „stan pola zwany falą elektromagnetyczną może ewoluować w przestrzeni pod nieobecność cząstek materialnych". Termin „próżnia" pojawia

się jeszcze w innym znaczeniu w kwantowej teorii pola.

Tak jak sygnalizowaliśmy, polu elektromagnetycznemu możemy przypisać energię i pęd. Wzór na energię pola elektromagnetycznego w próżni ma postać:

$$W = \frac{1}{2}\int_{R^3}\left(\varepsilon_0\vec{E}^2 + \frac{1}{\mu_0}\vec{B}^2\right)d^3\vec{r}, \tag{7}$$

gdzie ε_0 i μ_0 są stałymi, zw. przenikalnością elektryczną i magnetyczną próżni.

fotony

Z wielu eksperymentów wynikło, że fala elektromagnetyczna może przekazywać swą energię w ściśle określonych porcjach, tzw. kwantach. Najmniejsza porcja, czy kwant pola elektromagnetycznego, nazywa się fotonem. Odkrycie to doprowadziło do sformułowania elektrodynamiki kwantowej, a następnie kwantowej teorii pola. Kwantowa teoria pola z każdą obserwowaną cząstką wiąże pole. Skoro cząstki mają atrybuty takie jak masa, ładunek, spin itp., to i polom wiązanym z nimi nadaje się te same atrybuty. Tak więc wyprzedzając rozdział o kwantowej teorii pola będziemy mówić o klasycznych polach masowych naładowanych itp. Ponieważ nie udało się dotychczas stworzyć w pełni zadowalającej teorii cząstek elementarnych, prowadzi się wiele badań modelowych wprowadzając różne hipotetyczne pola i próbując przy ich pomocy wyjaśnić wyniki eksperymentów. Badanie różnorodnych pól klasycznych ma więc duże znaczenie, choć często sens fizyczny zostaje im nadany dopiero po ich skwantowaniu. Omówimy teraz krótko klasyczne pola swobodne i oddziałujące. Pola klasyczne nazywamy oddziałującymi, jeśli w równaniu każdego z nich występują inne pola. Oddziaływanie pól występuje np. przy rozchodzeniu się fali elektromagnetycznej i fali akustycznej w ośrodku sprężystym, którego stan wpływa na pole elektromagnetyczne. Wówczas równania pola elektromagnetycznego będą oczywiście zawierały parametry fali akustycznej.

Przed ogólnym omówieniem metod otrzymywania równań pól klasycznych oddziałujących między sobą zrobimy dygresję o obiektach geometrycznych stowarzyszonych z czasoprzestrzenią.

Wiemy, że w różnych układach odniesienia otrzymujemy różne współrzędne zdarzeń. Struktura czasoprzestrzeni jest dana przez określenie klasy równoważnych układów odniesienia i podanie związków między współrzędnymi dowolnego zdarzenia dla dowolnej pary układów O i O' z klasy równoważnej. Oznaczmy współrzędne zdarzenia p w układzie O: $p_O = (x^0, x^1, x^2, x^3)$, gdzie $x^0 = ct$. Oznaczmy współrzędne tego samego zdarzenia w układzie odniesienia O': $p_{O'} = (x^{0'}, x^{1'}, x^{2'}, x^{3'})$. Związek między tymi współrzędnymi ma postać:

$$x^{\mu'} = B_0^{\mu'}x^0 + B_1^{\mu'}x^1 + B_2^{\mu'}x^2 + B_3^{\mu'}x^3 + a^{\mu'}.$$

Przyjmując umowę, że powtarzające się te same indeksy — jeden „na górze", a drugi „na dole" — oznaczają sumowanie, możemy powyższy wzór zapisać w prostszej postaci:

$$x^{\mu'} = B_\mu^{\mu'}x^\mu + a^{\mu'}$$

(zbiór 16 liczb $B_\mu^{\mu'}$ nazywa się macierzą transformacji). Wyróżnioną klasę układów odniesienia w teorii Newtona i w szczególnej teorii względności tworzą układy inercjalne. Obie te teorie różnią się wyborem macierzy $B_\mu^{\mu'}$. W obu teoriach występują macierze obrotów w trzech wymiarach, które oznaczymy $A_i^{i'}$. W teorii Einsteina w miejsce transformacji Galileusza występują transformacje Lorentza, które oznaczymy $\Lambda_\mu^{\mu'}$ (→ O niektórych podstawowych pojęciach fizycznych).

obiekty geometryczne stowarzyszone z czasoprzestrzenią

Obiektem geometrycznym stowarzyszonym z czasoprzestrzenią nazywamy obiekt określony w każdym układzie współrzędnych przez tabelę liczb, które przy zmianie układu współrzędnych zmieniają się w ściśle określony sposób. Wyróżniamy następujące obiekty geometryczne:

skalary — $T_0' = T_0$;

wektory (kartezjańskie) — a^i (przy $B_i^{i'} = A_i^{i'}$), $a^{i'} = A^{i'} = A_i^{i'}a^i$;

czterowektory — a^μ (przy $B_\mu^{\mu'} = \Lambda_\mu^{\mu'}$), $a^{\mu'} = \Lambda_\mu^{\mu'}a^\mu$;

tensory kartezjańskie drugiego rzędu — $T^{i'j'} = A_i^{i'}A_j^{j'}T^{ij}$;

tensory lorentzowskie drugiego rzędu — $T^{\mu'\nu'} = \Lambda_\mu^{\mu'}\Lambda_\nu^{\nu'}T^{\mu\nu}$.

Analogicznie wprowadza się tensory wyższego rzędu. Przykładem skalara kartezjańskiego jest iloczyn skalarny wektorów $\vec{a}\cdot\vec{b}$, przykładem zaś wektora kartezjańskiego są wektory takie, jak wektor siły, prędkości, przyspieszenia. W przyjętej notacji sumacyjnej

$$\vec{a}\cdot\vec{b} = a^ib_i; \quad i = 1, 2, 3.$$

Przykładem skalara lorentzowskiego jest

$$a^\mu a_\mu = a^0a^0 - a^1a^1 - a^2a^2 - a^3a^3,$$

gdzie $a_0 \overset{\mathrm{df}}{=} a^0, a_1 = -a^1, a_2 = -a^2$ i $a_3 = -a^3$.

Własności geometryczne określają wskaźniki, po których nie ma sumowania. Wskaźniki tensorów lorentzowskich będziemy oznaczać literami greckimi, kartezjańskich — łacińskimi. Zgodnie z tymi ustaleniami wektorami są wielkości

$$f^{\mu\nu}g_\nu, \quad t^{mns}r_{ns}, \quad \gamma^\mu,$$

a skalarami — wielkości

$$f^{\mu\nu}f_{\mu\nu}, \quad \gamma^\mu\gamma_\mu.$$

Skalary, które przy transformacji odbicia przestrzennego $\vec{r} \to -\vec{r}$ zmieniają znak, nazywa się pseudoskalarami, a wektory nie zmieniające znaku — pseudowektorami (np. moment pędu). Wprowadza się też bispinory — kolumienki ψ_A czterech liczb zespolonych, o specyficznym prawie transformacji. Z bispinorów zbudować można skalar lorentzowski $\psi_A\ \psi^A$.

postulat Einsteina

Einstein wysunął postulat, że w każdym z równoważnych układów odniesienia równania fizyki powinny mieć tę samą postać. Winny być one związkami między obiektami geometrycznymi 4-wektorami, tensorami itp. Taką postać mają równania Newtona (3), w których zarówno $d^2\vec{r}/dt^2$ jak i $\vec{F}$ są wektorami kartezjańskimi. Czysta (bez obrotów) transformacja Galileusza nie zmienia współrzędnych tych wektorów. Zbiór transformacji współrzędnych zdarzeń w szczególnej teorii względności (określonych macierzami $B_\mu^{\mu'}$) nazywamy grupą Lorentza, a po dołączeniu przesunięć czasoprzestrzennych — grupą Poincarégo. Ponieważ czasoprzestrzeń ma strukturę wyznaczoną przez tę grupę (nazywa się ją grupą symetrii czasoprzestrzeni), przeto wszystkie równania powinny mieć formę współzmienniczą (tensorową) ze względu na transformacje tej grupy. Taki postulat doprowadził Einsteina do zmodyfikowanej postaci równań Newtona (3):

$$\frac{dp^\mu}{d\tau} = F^\mu.$$

W równaniach tych $p^\mu = (E/c, \vec{p})$ i

$$F^\mu = \left(\frac{\vec{F}\cdot\vec{v}}{\sqrt{1-v^2/c^2}}, \frac{\vec{F}}{\sqrt{1-v^2/c^2}}\right),$$

gdzie wielkości E, $\vec{p}$, $\vec{F}$ i $\vec{v}$ są odpowiednio energią, pędem, siłą i prędkością, zaś τ jest skalarem lorentzowskim (czasem własnym, mierzonym na zegarze poruszającym się z punktem materialnym). Postulat Einsteina ułatwia znalezienie właściwych równań pola.

Efektywną metodą otrzymywania równań pola o dobrze określonych własnościach transformacyjnych jest metoda wariacyjna. Wywodzi się ona z siedemnastowiecznej zasady Fermata, która brzmi: między dowolnymi dwoma punktami A i B światło wybiera drogę, dla której czas przebiegu jest najkrótszy lub — rzadziej — najdłuższy. Ponieważ prędkość światła w punkcie $\vec{r}$ w ośrodku o współczynniku załamania $n(\vec{r})$ wynosi $c/n(\vec{r})$, to światło przebywa odcinek łuku krzywej dl w czasie $dt = \dfrac{dl}{c/n} = \dfrac{n\,dl}{c}$.

wariacyjna postać równań ruchu

Czas $T_{AB}[K]$, w którym światło przebiega krzywą K, dany jest wzorem:

$$T_{AB}[K] = \int_K^{AB}\frac{n(\vec{r})dl}{c}.$$

Metoda matematyczna zwana rachunkiem wariacyjnym umożliwia otrzymanie równania różniczkowego krzywej K, dla której wielkość $T_{AB}[K]$ ma wartość najmniejszą lub największą. Równanie to zapisuje się w postaci

$$\delta T = 0; \tag{8}$$

odczytujemy je: wariacja T równa jest zeru. Korzystając z rachunku wariacyjnego i znając wyrażenie na T można otrzymać z równania (8) równanie szukanej krzywej.

Wykazano, że równanie ruchu układu punktów materialnych również można zapisać w formie wa-

zasada najmniejszego działania

riacyjnej $\delta I = 0$, gdzie I, zwane działaniem, ma postać:

$$I = \int L dt,$$

a L jest funkcją Lagrange'a (lagrangianem):

L = energia kinetyczna – energia potencjalna.

Znajdowanie równań ruchu metodą wariacyjną ma kilka zalet. Otrzymane równania mają dobrze określone właściwości tensorowe, o ile działanie jest skalarem ze względu na odpowiednią dla problemu grupę symetrii. Z każdą transformacją zmiennych występujących w funkcji L nie zmieniającą działania związane jest istnienie odpowiedniej całki równań ruchu. Twierdzenie Noether daje efektywną metodę znajdywania tych całek równań ruchu. Na przykład z niezmienniczości ze względu na przesunięcia w czasie wynika całka energii, ze względu na przesunięcia w przestrzeni — całka pędu, ze względu na obroty — całka momentu pędu (→ Zasady zachowania).

Te wszystkie zalety metody wariacyjnej spowodowały, że się ją najczęściej stosuje do wyprowadzania równań pola. Ponieważ właściwą grupą symetrii czasoprzestrzeni jest grupa Poincarégo, działanie I zapisuje się w relatywistycznej teorii pola w trochę innej postaci:

$$I = \int_{R^4} \mathcal{L}\left(x, \varphi^\alpha, \frac{\partial \varphi^\alpha}{\partial x^\mu}, \ldots\right) d^4x,$$

gdzie $\mathcal{L}(x)$ jest gęstością lagrangianu, a α oznacza wszystkie indeksy pola φ, łącznie z indeksami tensorowymi ze względu na grupę Lorentza. Ponieważ element objętości czasoprzestrzeni d^4x jest skalarem lorentzowskim, to aby działanie było skalarem, gęstość lagrangianu musi też być skalarem.

Dla układu pól oddziałujących gęstość lagrangianu składa się z sumy gęstości lagrangianów odpowiadających polom swobodnym i gęstości lagrangianu oddziaływania $\mathcal{L}_I$. W $\mathcal{L}_I$ występują przeważnie najprostsze skalary lorentzowskie skonstruowane z pól swobodnych występujących w rozważanym problemie.

Zanim podamy najważniejsze gęstości lagrangianów i równania pola, wprowadzimy pewne upraszczające oznaczenia:

$$\frac{\partial}{\partial x^\mu} A^\nu = \partial_\mu A^\nu = A^\nu{}_{,\mu},$$

$$\frac{\partial}{\partial x_\mu} A^\nu = \partial^\mu A^\nu,$$

$$\square = \Delta - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} = -\partial_\mu \partial^\mu.$$

Operację $\square$ nazywa się operatorem d'Alemberta. Pola można reprezentować również przez funkcje o wartościach zespolonych; sprzężenie zespolone oznaczamy gwiazdką, np. φ^* jest funkcją sprzężoną zespoloną do φ. Pola takie służą do opisu cząstek naładowanych.

gęstości lagrangianów i równania pól swobodnych

Podstawowe gęstości lagrangianów i równania pól swobodnych są następujące:

Naładowane pola skalarne U_μ o masie m i spinie 0

$$\mathcal{L} = (\partial_\mu \varphi)^*(\partial^\mu \varphi) - m^2\varphi\varphi^*, \tag{9}$$

$$(\square - m^2)\varphi = 0. \tag{10}$$

Naładowane pole wektorowe U_μ o masie m i spinie 1

$$\mathcal{L} = -\frac{1}{4} F^{*\mu\nu} F_{\mu\nu} + \frac{m^2}{2} U_\mu U^\mu$$

(gdzie $F^{\mu\nu} = \partial^\mu U^\nu - \partial^\nu U^\mu$),

$$(\square - m^2) U^\mu = 0.$$

Swobodne pole elektromagnetyczne

$$\mathcal{L} = -{}^1/_4 f_{\mu\nu} f^{\mu\nu}, \tag{11}$$

(gdzie $f^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$),

$$\square A^\mu = 0;$$

pole A^μ wyraża się przez potencjał elektrostatyczny φ i potencjał wektorowy $\vec{A}$ w następujący sposób: $A^\mu = (c\varphi, \vec{A})$.

Pole o spinie $^1/_4$ i masie m

$$\mathcal{L} = -{}^1/_4(\overline{\psi}\gamma^\mu \psi_{,\mu} - \overline{\psi}_{,\mu}\gamma^\mu \psi) - m\overline{\psi}\psi, \quad (i\gamma^\mu \partial_\mu - m)\psi(x) = 0; \tag{12}$$

ψ i $\overline{\psi}$ oznaczają pola bispinorowe, a γ^μ ($\mu = 0, 1, 2, 3$) — cztery macierze 4×4 transformujące się jak składowe czterowektora przy zmianie układu odniesienia (dla uproszczenia opuszczamy we wzorze 12 i następnych wysumowane wskaźniki bispinorowe).

Podamy teraz kilka podstawowych gęstości lagrangianów $\mathcal{L}_I$ opisujących oddziaływania pól. W elektrodynamice

$$\mathcal{L}_I = -eA_\mu j^\mu,$$

gdzie j^μ nazywa się prądem elektromagnetycznym. Prąd j^μ może być klasyczny lub kwantowy; jeśli jest kwantowy, to j^μ wyraża się przez bispinorowe pole elektronowe:

$$j^\mu = \overline{\psi}\gamma^\mu \psi,$$

czyli j^μ to najprostszy wektor lorentzowski zbudowany z $\overline{\psi}$, ψ i γ^μ. Podobnie najprostszym oddziaływaniem pola bispinorowego ψ z polem skalarnym φ jest

$$\mathcal{L}_I = g\overline{\psi}\psi\varphi,$$

a z polem pseudoskalarnym φ':

$$\mathcal{L}_I = g\overline{\psi}\gamma_5\psi\varphi'$$

gdzie γ_5 transformuje się jak pseudoskalar. To ostatnie oddziaływanie zwie się oddziaływaniem Yukawy i zaproponowane było do opisu oddziaływania np. nukleonów z pionami. Stałą g nazywa się stałą sprzężenia.

Oprócz indeksów lorentzowskich dodaje się polom indeksy określające ich własności ze względu na inne grupy symetrii, np. określające izospin, spin unitarny, kolor, zapach itp. (→ Zasady zachowania). Wtedy, kiedy oddziaływanie łamie symetrię, $\mathcal{L}_I$ nie może być skalarem ze względu na grupę określającą tę symetrię. Jest to pewna wskazówka przy poszukiwaniu odpowiedniej gęstości lagrangianu. Dodatkowe ograniczenie daje żądanie renormalizowalności teorii po skwantowaniu, do czego wrócimy później.

Bardzo wazną symetrią, która odegrała kluczową rolę w konstruowaniu najnowszej teorii oddziaływań słabych i elektromagnetycznych, jest symetria cechowania. Zilustrujemy ją na przykładzie skalarnego pola naładowanego (9). Pole to, jako najprostsze najczęściej jest stosowane w rozważaniach modelowych. Widać, że gęstość lagrangianu w tym wzorze jest niezmiennicza ze względu na transformację zw. globalną transformacją cechowania:

$$\varphi' = e^{iQ\lambda}\varphi, \tag{13}$$

globalna transformacja cechowania

gdzie Q to ustalony, a λ — dowolny rzeczywisty parametr niezależny od x. Dowód jest oczywisty, bo ze wzoru (13) wynika, że $\varphi'^* = e^{-iQ\lambda}\varphi^*$ i odpowiednie czynniki fazowe się znoszą. Pole φ jest polem o wartościach zespolonych, czyli $\varphi = \mathrm{Re}\varphi + i\mathrm{Im}\varphi$. Transformacja (13) ma interpretację obrotu na wykresie, na którym na osi x odkłada się $\mathrm{Re}\varphi$, a na osi y odkłada się $\mathrm{Im}\varphi$.

lokalna transformacja cechowania

Lokalną transformację cechowania wprowadza się wzorem

$$\varphi' = e^{iQ\lambda(x)}\varphi. \tag{14}$$

Okazuje się, że stała Q we wzorze (14) ma interpretację elektrycznego ładunku pola. Widać, że gęstość (9) nie jest niezmiennicza ze względu na tę transformację, ponieważ:

$$\partial^\mu \varphi' = e^{iQ\lambda(x)}\left(\partial^\mu \varphi + iQ\partial^\mu \lambda(x)\right).$$

Zastanówmy się, jak należy zmienić gęstość (9), żeby nowa gęstość lagrangianu była niezmiennicza ze względu na transformację (14). Jedno z rozwiązań jest następujące: dodajemy do operatora ∂_μ pole wektorowe A_μ pisząc:

$$D_\mu = \partial_\mu - iQA_\mu; \tag{15}$$

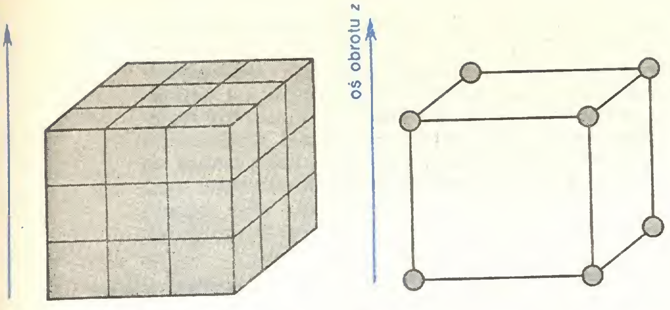
wtedy w miejsce wzoru (9) otrzymujemy

$$\tilde{\mathcal{L}} = (D_\mu\varphi)^* D^\mu\varphi - m^2\varphi\varphi^*. \tag{16}$$

Aby $\tilde{\mathcal{L}}$ było niezmiennicze ze względu na (14), musimy określić, jak pole A_μ transformuje się przy transformacji cechowania. Jeśli:

$$A'_\mu = A_\mu - \partial_\mu \Lambda(x), \tag{17}$$

to łatwo wykazać, że $D'_\mu D'^\mu = D_\mu D^\mu$, czyli $\tilde{\mathcal{L}}' = \tilde{\mathcal{L}}$. (Elementarną ilustracją lokalnych i globalnych transformacji cechowania jest rys. 5). Z tego wynika wniosek, że z niezmienniczością ze względu na lokalną transformację cechowania naładowanych pól (transformacje (14) tworzą grupę $U(1)$), związane jest w naturalny sposób wektorowe pole cechowania A_μ sprzę



b)

Rys. 5. Ilustracja lokalnej grupy cechowania 0(2) obejmującej obroty wokół osi z. a) Sześcian ułożony z identycznych klocków, b) sześcian złożony z prętów i odpowiednio naciętych kul, tak aby wygląd kuli nie zależał od obrotu wokół osi z, oraz aby każda kula mogła być obracana niezależnie bez psucia konstrukcji. Obie bryły są niezmiennicze ze względu na globalną grupę cechowania — czyli na obroty bryły jako całości wokół osi z. Przy przejściu do lokalnej grupy cechowania — niezależne obroty klocków i kul — bryła pierwsza ulegnie deformacji, a druga zachowa swój kształt

pole cechowania

gające się z polem naładowanym tak, jak we wzorze (16). Do $\tilde{\mathcal{L}}$ musimy dodać gęstość lagrangianu swobodnego pola cechowania A_μ. Widać, że pole A_μ musi być bezmasowe, ponieważ tylko gęstość lagrangianu swobodnego pola wektorowego bezmasowego (11) jest niezmiennicza ze względu na transformację (17). Po dodaniu tej swobodnej gęstości i rozpisaniu wzoru (16) przy pomocy wzoru (15), otrzymujemy gęstość lagrangianu, którą przez analogię do elektrodynamiki wypisalibyśmy dla naładowanego pola skalarnego oddziałującego z polem elektromagnetycznym. Pozwala to utożsamić w tym przypadku pole cechowania A_μ z polem elektromagnetycznym (nieprzypadkowa zbieżność oznaczeń). Podobnie startując z naładowanego pola bispinorowego (którego kwanty odpowiadają elektronom) i szukając najprostszej gęstości lagrangianu niezmienniczej ze względu na lokalną grupę cechowania, otrzymujemy gęstość lagrangianu elektrodynamiki. Stąd bierze się idea poszukiwania nowych interesujących fizycznie pól jako pól cechowania.

teorie z wyższymi grupami lokalnymi cechowania

Wyposażając pola w dodatkowe własności można budować gęstości lagrangianu niezmiennicze ze względu na grupy bardziej złożone, takie jak $SU(2)$, $SU(3)$ (→ Cząstki elementarne i ich oddziaływania). Transformacje np. z grupy $SU(2)$ można przedstawić w postaci:

$$e^{I_i\lambda_i} \overset{\text{df}}{=} 1+(\lambda_i I^i)+{}^1/_2(\lambda_i I^i)^2+\ldots,$$

gdzie I_1, I_2, I_3 są pewnymi macierzami zwanymi generatorami grupy $SU(2)$. Dla grupy $SU(3)$ mamy 8 generatorów. Żądając, aby gęstość lagrangianu była niezmiennicza ze względu na lokalną grupę $SU(2)$, czyli ze względu na transformacje:

$$e^{I_i\lambda_i(x)},$$

otrzymuje się w naturalny sposób (analogiczny do dyskutowanego wyżej) 3 bezmasowe wektorowe pola cechowania. Ponieważ w przyrodzie nie są znane żadne cząstki wektorowe bezmasowe poza fotonem, wydawało się przez wiele lat, że teorie z wyższymi grupami lokalnymi cechowania nie znajdą zastosowania w fizyce. Subtelne rozważania na gruncie kwantowej teorii pola wykazały, że jest inaczej.

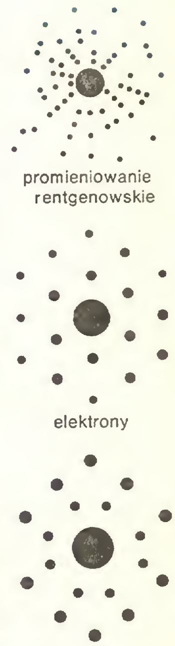
Czym zajmuje się mechanika kwantowa

Podstawowym przesłankami powstania mechaniki kwantowej były następujące obserwacje. Zauważono, że w mikroświecie pewne wielkości fizyczne, takie jak energia E i moment pędu J, mogą zmieniać się tylko skokowo. Stwierdzono poza tym, że pewne wielkości fizyczne, zwane komplementarnymi, mają tę własność, że ich jednoczesny dokładny pomiar jest niemożliwy (np. im bardziej precyzyjny będzie pomiar położenia mikrocząstki, tym większą niekontrolowaną zmianę jej pędu pomiar ten spowoduje). Mimo silnego wpływu pomiarów na stan mikrocząstki obserwuje się wiele regularności w eksperymentach przeprowadzanych z bardzo dużą liczbą identycznie przygotowanych mikrocząstek. Na przykład, nie jesteśmy w stanie przewidzieć, kiedy rozpadnie się obserwowana mikrocząstka nietrwała; natomiast okazuje się, że czas, w jakim rozpada się połowa cząstek z próbki materiału radioaktywnego, nie zależy od wielkości próbki, tylko od wybranego pierwiastka i wszystkie eksperymenty dają bardzo zbliżoną jego wartość. Podobnie promienie rentgenowskie, elektrony lub neutrony po przejściu przez kryształ tworzą piękne obrazy dyfrakcyjne (rys. 6), natomiast zachowania się wybranej mikrocząstki nie jesteśmy w stanie przewidzieć. Tak więc regularności obserwowane w mikroświecie mają charakter statystyczny, to znaczy, jeśli w pewnym eksperymencie mamy dużą liczbę identycznie przygotowanych mikroukładów fizycznych (w stanie początkowym i), to z reguły w wyniku eksperymentu otrzymamy grupy układów różniące się od siebie (w różnych stanach końcowych f). Regularność przyrody polega na tym, że powtarzając ten sam eksperyment wiele razy zawsze otrzymujemy te samą czestość występowania mikroukładu w stanie końcowym f, jeśli był on w stanie początkowym i; prawdopodobieństwa P_{if} występowania stanów f spełniają związek

statystyczny charakter mechaniki kwantowej

$$\sum_{f=1}^{N} P_{if} = 1$$

(N oznacza liczbę stanów końcowych). Celem mechaniki kwantowej jest wyznaczanie wartości prawdopodobieństw P_{if} odnoszących się do różnych eksperymentów, jak również znajdowania dopuszczalnych wartości wielkości fizycznych charakteryzujących stan mikroukładów (np. poziomy energetyczne atomów).



Rys. 6. Obrazy dyfrakcyjne Lauego wytworzone przez: promieniowanie rentgenowskie, elektrony i neutrony

Jeśli podzielimy przestrzeń na bardzo wiele (N bliskie nieskończoności) sześcianików, możemy za stan f mikrocząstki uznać fakt obserwacji jej w f-tym sześcianiku. Wyobraźmy sobie, że mamy urządzenie wysyłające mikrocząstki (źródło promieniotwórcze, wyjście z akceleratora itp.) i interesuje nas położenie mikrocząstki po czasie T od chwili wyjścia z urządzenia. Stanem i jest teraz wyjście cząstki z określonego urządzenia, a $P_{if}(T)$ — prawdopodobieństwem rejestracji jej po czasie T w f-tym sześcianiku. Aby obliczyć prawdopodobieństwa $P_{if}(T)$ wprowadza się w mechanice kwantowej funkcję falową $\Psi(\vec{r}, t)$. Prawdopodobieństwa $P_{if}(T)$ otrzymuje się przy pomocy wzoru

$$P_{if}(T) = \int_{V_f} d^3\vec{r}\,\Psi^*(\vec{r}, T)\Psi(\vec{r}, T) \sim$$

$$\sim \Psi^*(\vec{r_f}, T)\Psi(\vec{r_f}, T)\,|V_f|,$$

gdzie V_f oznacza sześcianik, $\vec{r_f}$ — jego środek, a $|V_f|$ — jego objętość. Funkcja falowa $\Psi(\vec{r}, t)$ jest obiektem abstrakcyjnym. Jest to funkcja o wartościach zespolonych, a ponieważ służy do obliczania prawdopodobieństw, nazywa się ją amplitudą prawdopodobieństwa; oczywiście musi mieć własność

funkcja falowa

$$\int_{R^3} \Psi^*(\vec{r}, t)\Psi(\vec{r}, t)\,d^3\vec{r} = 1.$$

Zmienność w czasie funkcji falowej cząstki o masie m jest dana przez równanie Schrödingera:

równanie Schrödingera

$$\hbar i \frac{\partial}{\partial t}\Psi = -\frac{\hbar^2}{2m}\Delta\Psi(\vec{r}, t) + V(\vec{r})\Psi(\vec{r}, t). \qquad (18)$$

Równanie to zapisuje się również w postaci:

$$\hbar i \frac{\partial}{\partial t} \Psi = \hat{H}\Psi,$$

zasada korespondencji Bohra

gdzie operator różniczkowy $\hat{H}$ nazywa się hamiltonianem. Ważną rolę przy wypisywaniu równania (18) dla konkretnego problemu odgrywa zasada korespondencji Bohra, zgodnie z którą, jeżeli dokonujemy pomiary z dokładnością makroskopową (np. obserwujemy tor cząstki elementarnej w komorze pęcherzykowej, iskrowej czy w emulsji), to wartość mierzona wielkości fizycznej musi się zmieniać zgodnie z prawami mechaniki klasycznej (położenie „makroskopowe" musi zmieniać się w czasie zgodnie z równaniami Newtona dla danego problemu). Panuje przekonanie, że w funkcji falowej danego układu fizycznego jest zawarta najpełniejsza możliwa informacja o własnościach mikrocząstek. W ostatnim rozdziale wspomnimy o poglądach odmiennych. W istocie przy pomocy funkcji ψ można przewidzieć wyniki bardzo różnych eksperymentów. Na przykład jeśli zamiast położenia mierzylibyśmy w omawianym powyżej eksperymencie średni pęd mikrocząstek w kierunku osi x w chwili T, $\langle P_x(T)\rangle$, to dany on byłby wzorem

$$\langle P_x(T)\rangle = \frac{\hbar}{i} \int\limits_{R^3} \Psi^*(\vec{r}, T) \frac{\partial}{\partial x} \Psi(\vec{r}, T) d^3\vec{r}.$$

Podobnie z każdą inną wielkością fizyczną O można skojarzyć operator $\hat{O}$ tak, żeby średnia mierzona wartość $\langle O(T)\rangle$ była dana wzorem:

$$\langle O(T)\rangle = \int\limits_{R^3} \Psi^*(\vec{r}, T)\hat{O}\Psi(\vec{r}, T) d^3\vec{r}. \qquad (19)$$

Ponieważ funkcję falową można zapisywać jako funkcję innych zmiennych niż $\vec{r}$, na przykład pędu $\vec{p}$, używa się ściślejszego i bardziej abstrakcyjnego języka mówiąc, że każdemu stanowi mikrocząstki odpowiada wektor stanu Ψ należący do przestrzeni Hilberta, a każdej wielkości fizycznej O — operator samosprzężony $\hat{O}$ w tej przestrzeni taki, że średnia wartość $\langle O\rangle$ wynosi:

$$\langle O\rangle = \langle \Psi, \hat{O}\Psi\rangle,$$

gdzie $\langle \Psi, \hat{O}\Psi\rangle$ oznacza iloczyn skalarny; dla $\Psi = \Psi(\vec{r}, t)$ ma on postać (19). Intuicje klasyczne są istotne przy wyborze $\hat{H}$. Na przykład dla atomu wodoru (który klasycznie wyobrażamy sobie jako oddziałujące punktowe cząstki: lekki elektron i ciężki proton, tworzące stan związany) w równaniu Schrödingera (18) w pierwszym przybliżeniu w miejsce $V(\vec{r})$ pojawi się klasyczny potencjał elektrostatyczny. Równanie to umożliwi wyznaczenie poziomów energetycznych atomu wodoru i wielu zjawisk dotyczących tego atomu, nie da nam jednak żadnego wyobrażenia o budowie atomu. Nie wiemy, co się dzieje z elektronem w atomie. Wszelkie rysunkowe wyobrażenia są nieprawdziwe.

prawdopodobieństwo przejścia

Jak już wspominaliśmy, mechanika kwantowa umożliwia również obliczenie prawdopodobieństw przejścia P_{if}:

$$P_{if} = |\langle \Psi_f, \hat{S}\Psi_i\rangle|^2, \qquad (20)$$

gdzie Ψ_i i Ψ_f oznaczają początkową i końcową funkcję falową, $\hat{S}$ — pewien operator dający się wyznaczyć z równania Schrödingera. Jeśli stan końcowy można otrzymać w wyniku N różnych procesów, a eksperyment nie pozwala na rozstrzygnięcie, który proces zaszedł w danym wypadku, to amplituda $A_{fi} = \langle \Psi_f, \hat{S}\Psi_i\rangle$ daje się przedstawić jako suma N amplitud A^p_{fi} (A^p_{fi} jest amplitudą odpowiadającą p-temu procesowi):

$$A_{fi} = \sum_{p=1}^{N} A^p_{fi}. \qquad (21)$$

Ten postulat odegrał podstawową rolę w sformułowaniu mechaniki kwantowej przez Feynmana. Opierają się na nim rozważania w artykule „Oddziaływania elektromagnetyczne". Jeśli procesów jest nieprzeliczalnie wiele, suma we wzorze (21) musi być zastąpiona całką. Taka struktura amplitud przejścia jest właściwa dla kwantowej teorii pola, którą buduje się w podobny sposób jak mechanikę kwantową.

O kwantowaniu pól swobodnych

Jak już wspominaliśmy, fale elektromagnetyczne w wielu eksperymentach (efekt fotoelektryczny, zjawisko Comptona) zachowują się jak strumienie cząstek — fotonów o dobrze określonej energii, pędzie i o zerowej masie. Cząstki te biorą udział w wielu oddziaływaniach z elektronami i pozytonami (anihilacja par e^+e^-, kreacja par, produkcja dodatkowych fotonów przy zderzeniach e^+e^- i e^-e^- itp.). Dla opisania tych oddziaływań powstała elektrodynamika kwantowa. Służy ona do znajdowania amplitud prawdopodobieństwa wyżej wymienionych procesów. Dla opisania procesów oddziaływania cząstek elementarnych powstała kwantowa teoria pola. Według kwantowej teorii pola z każdą obserwowaną w przyrodzie cząstką elementarną jest stowarzyszone pole. Oddziaływania cząstek to oddziaływania odpowiednich pól. Pole będąc obiektem fizycznym może znajdować się w różnych stanach, np. 1 foton to stan jednofotonowy, 2 fotony to stan dwufotonowy. W kwantowej teorii pola, podobnie jak w mechanice kwantowej, stanom układu przyporządkowuje się wektory stanu w przestrzeni Hilberta, a wielkościom mierzalnym, tzw. obserwablom — operatory. Pomocą przy znajdowaniu takiej reprezentacji matematycznej jest znajomość teorii klasycznej, wyników doświadczeń oraz zasada korespondencji. Podstawowymi obserwablami w mikroświecie są energia E i pęd $\vec{p}$. Energia pola i pęd są w teorii klasycznej funkcjami samych pól (np. wzór 13). Dla pól bez analogii klasycznej wzory na E i $\vec{p}$ otrzymuje się z twierdzenia Noether. Ponieważ E i $\vec{p}$ w teorii kwantowej mają być operatorami, więc pola kwantowe, z których zgodnie z zasadą korespondencji są one zbudowane, muszą być operatorami. Mówiąc ściślej, operatorami są pola po wycałkowaniu ich z funkcją odpowiednio regularną. W matematyce obiekty takie nazywa się dystrybucjami o wartościach operatorowych.

kwantowa teoria pola

Prześledzimy strukturę matematyczną teorii na przykładzie swobodnego bezmasowego pola skalarnego. Stan pola, w którym jest n kwantów o energii $\hbar\omega$ i pędzie $\hbar\vec{k}$ jest z definicji stanem o określonej energii i pędzie. Oznaczmy wektor z przestrzeni Hilberta reprezentujący ten stan przez $\Phi_n = |n, \vec{k}\rangle$. Wartości oczekiwane energii i pędu są równe:

wartości oczekiwane energii i pędu

$$E = n\hbar\omega \quad \text{i} \quad \langle \vec{p}\rangle = n\hbar\vec{k}, \qquad (22)$$

uwzględniając (20) i wprowadzając oznaczenia

$$\langle n, \vec{k}|\hat{E}|n, \vec{k}\rangle \overset{\text{df}}{=} \langle \Phi_n, \hat{E}\Phi_n\rangle,$$

$$\langle n, \vec{k}|\hat{\vec{p}}|n, \vec{k}\rangle \overset{\text{df}}{=} \langle \Phi_n, \hat{\vec{p}}\Phi_n\rangle,$$

można to zapisać w postaci

$$\langle n, \vec{k}|\hat{E}|n, \vec{k}\rangle = n\hbar\omega, \qquad (23)$$

$$\langle n, \vec{k}|\hat{\vec{p}}|n, \vec{k}\rangle = n\hbar\vec{k}, \qquad (24)$$

gdzie wielkości ω i $k = |\vec{k}|$ związane są równością $\omega^2/c^2 - k^2 = 0$. Wielkości E i $\vec{p}$ tworzą razem cztero-

wektor, zatem (22) można zapisać krótko

$$\langle p^\mu \rangle = n\hbar k^\mu, \qquad \mu = 0, 1, 2, 3,$$

gdzie $k^\mu = (\omega/c, \vec{k})$. Najprostsze rozwiązanie równania (10) dla $m = 0$ ma postać:

$$\varphi_k = c'(\vec{k}) e^{i(\omega t - \vec{k}\vec{x})} = c'(\vec{k}) e^{ik_\mu x^\mu}, \qquad (25)$$

gdzie c' to dowolna stała, która może zależeć od $\vec{k}$. Energia pola klasycznego odpowiadająca (na podstawie twierdzenia Noether) temu rozwiązaniu wynosi:

$$E(\varphi_k) = c'^*(\vec{k}) c'(\vec{k}) \omega(k) = c^*(\vec{k}) c(\vec{k}) \hbar\omega(\vec{k}).$$

Najprostszym sposobem przyporządkowania energii E — operatora — (w zgodzie ze wzorem 23) jest zastąpienie współczynnika $c^*(\vec{k})\, c(\vec{k})$ operatorem liczby cząstek o pędzie $\vec{k}$, mającym własność:

$$\hat{n}(\vec{k})|n, \vec{k}\rangle = n|n, \vec{k}\rangle.$$

Operator $\hat{n}(k)$ ma żądane własności, jeśli współczynnikom c^* i c przyporządkujemy operatory $\hat{c}$ i $\hat{c}^+$ takie, że

$$\hat{c}(\vec{k})|n, \vec{k}\rangle = \sqrt{n}\,|n-1, \vec{k}\rangle, \qquad (26)$$

$$c^+(\vec{k})|n, \vec{k}\rangle = \sqrt{n+1}\,|n+1, \vec{k}\rangle. \qquad (27)$$

operatory kreacji i anihilacji

Operatory (26) i (27) nazywamy odpowiednio operatorem anihilacji i kreacji cząstek o pędzie $\vec{k}$. Najogólniejsze rozwiązanie równania (10) dla $m = 0$ jest kombinacją liniową rozwiązań (25):

$$\varphi(x) = \int_{R^3} \frac{d^3\vec{k}}{\sqrt{2\omega}} c(\vec{k}) e^{-i\omega t + i\vec{k}\cdot\vec{x}} + c^*(\vec{k}) e^{i\omega t - i\vec{k}\cdot\vec{x}}, \qquad (28)$$

a energia

$$E(\varphi) = \hbar \int_{R^3} c^*(\vec{k}) c(\vec{k}) \omega d^3\vec{k}. \qquad (29)$$

Zastępując we wzorach (28) i (29) stałe $c(\vec{k})$ i $c^*(\vec{k})$ przez operatory (26) i (27) otrzymujemy operator pola $\hat{\varphi}$ i operator energii $\hat{E}$:

$$\hat{\varphi}(x) = \int_{R^3} \frac{d^3\vec{k}}{\sqrt{2\omega}} \hbar(\hat{c}(\vec{k}) e^{-ik_\mu x^\mu} + \hat{c}^+(\vec{k}) e^{ik_\mu x^\mu}), \qquad (30)$$

$$\hat{E} = \hbar \int_{R^3} \hat{c}^+(\vec{k}) \hat{c}(\vec{k}) \omega(\vec{k}) d^3\vec{k}.$$

kwantowanie pola

Podobnie wszystkie inne obserwable wyrażające się przez pole $\varphi(x)$ stają się po uwzględnieniu wzoru (30) operatorami

$$O(\varphi) \to :O(\hat{\varphi}):.$$

Procedurę przyporządkowywania polom klasycznym pól kwantowych nazywa się kwantowaniem. Symbol $: \ :$ oznacza tzw. normalne uporządkowanie iloczynów operatorów kreacji i anihilacji występujących w $O(\hat{\varphi})$; uporządkowanie musi być takie, aby wszystkie operatory anihilacji stały na prawo od operatorów kreacji, np.:

$$:\hat{c}(\vec{k}_2)\hat{c}^+(\vec{k}_1)\hat{c}(\vec{k}_1)\hat{c}^+(\vec{k}_2): \stackrel{\text{df}}{=} \hat{c}^+(\vec{k}_1)c^+(\vec{k}_2)c(\vec{k}_2)c(\vec{k}_1).$$

Szczególne znaczenie ma stan pola Ω zwany próżnią, zdefiniowany równościami

$$\hat{c}(\vec{k})\Omega = 0 \quad \text{dla każdego } \vec{k}.$$

stan próżni

Za pomocą Ω i operatorów kreacji można przedstawić każdy stan pola jako stan otrzymany z próżni przez działanie operatorów kreacji:

$$|1, \vec{k}\rangle = c^+(\vec{k})\Omega,$$

$$|2, \vec{k}\rangle = {}^1/_2 c^+(\vec{k}) c^+(\vec{k})\Omega,$$

$$\cdots$$

$$|n, \vec{k}\rangle = \frac{(\hat{c}^+(\vec{k}))^n}{n!}\Omega.$$

Przy takim opisie okazuje się, że stany pola klasycznego mogą być reprezentowane przez wektory stanu z nieskończoną liczbą cząstek, które nazywa się stanami koherentnymi. Wartość oczekiwana pola kwantowego $\hat{\varphi}$ w tych stanach równa jest polu klasycznemu. Z doświadczenia wiadomo, że w stanach o niewielkiej liczbie kwantów pole przejawia własności cząstkowe, natomiast w stanach o bardzo wielkiej liczbie kwantów (np. klasyczna makroskopowa fala elektromagnetyczna) pole przejawia własności falowe. Na tym polega dualizm falowo-korpuskularny.

Aby wartość oczekiwana pola kwantowego $\varphi(x)$ (30) dla stanu koherentnego spełniała równanie pola klasycznego, przyjmuje się taką samą postać równania pola kwantowego $\hat{\varphi}$, jak pola klasycznego φ. Przy takim założeniu wektory stanu pola nie zależą od czasu. Tego rodzaju opis ewolucji układu w czasie nazywamy obrazem Heisenberga. Inny możliwy opis to obraz Schrödingera, w którym operatory pola nie zależą od czasu, a wektory stanu Φ spełniają równanie:

$$i\hbar \frac{d\Phi(t)}{dt} = \hat{H}\Phi(t),$$

gdzie przeważnie za operator $\hat{H}$ można przyjąć operator energii $\hat{E}$. Dygresja ta będzie rozszerzona po wprowadzeniu oddziaływań pól kwantowych.

Kwantowanie innych swobodnych pól klasycznych przebiega podobnie, z tym że dopuszczalne stany pól są bogatsze, kwanty ich charakteryzuje ładunek, niezerowy spin (całkowity dla pól wektorowych i dla pól bispinorowych) oraz dodatkowe własności takie jak izospin, spin unitarny itp. Tak więc operatory $\hat{c}(\vec{k})$ przechodzą w $\hat{c}_\alpha(\vec{k})$, a wskaźnik α oznacza dodatkowe własności.

istnienie antycząstek

Okazuje się, że dla każdego kwantu pola, który charakteryzuje masa m, czteropęd $\hbar k^\mu$, ładunki q_i (ładunek, hiperładunek, liczba barionowa), rzut spinu S_z na oś z, istnieje kwant pola (zwany antycząstką) o masie m, czteropędzie $\hbar k^\mu$, ładunkach $-q_i$ i rzucie spinu $-S_z$. Często spotyka się w literaturze popularnonaukowej zdanie, od którego zwykłemu czytelnikowi jeżą się włosy: „Antycząstka to cząstka o energii ujemnej poruszająca się wstecz w czasie". W przyrodzie nic nie porusza się wstecz w czasie. Zdanie powyższe należy rozumieć w następujący sposób: jeśli w równaniu opisującym cząstkę zmienimy występujący tam parametr energii E na $-E$ oraz t na $-t$, to otrzymamy równanie opisujące antycząstkę o energii $E\,(E > 0)$, poruszającą się w przód w czasie. Bez tego wyjaśnienia zdanie przytoczone (które jest właściwie rozumiane przez fizyków) zmienia się w zdanie nonsensowne. Należy go więc unikać.

Konieczność porządkowania normalnego wymaga w pewnych sytuacjach znajomości własności operatorów $\hat{c}(\vec{k})$ i $\hat{c}^+(\vec{k})$ przy ich przestawieniu, czyli znajomości tzw. związków przemienności. Aby wyrażenie mające interpretację energii było dodatnie, muszą zachodzić następujące związki:

$$[\hat{c}_\alpha(\vec{k}), c^+_\beta(\vec{k}')]_\pm \stackrel{\text{df}}{=} \hat{c}_\alpha(\vec{k}) c^+_\beta(\vec{k}') \pm c^+_\beta(\vec{k}') c_\alpha(\vec{k}) = \delta_{\alpha\beta}\,\delta(\vec{k} - \vec{k}'), \qquad (31)$$

związki przemienności

gdzie $\delta_{\alpha\beta}\delta(\vec{k}-\vec{k}')$ znika dla $\alpha \neq \beta$ i $\vec{k} \neq \vec{k}'$. Znak plus odpowiada polom o spinie połówkowym, a minus — polom o spinie całkowitym. Mówimy, że kwanty tych pól podlegają odpowiednio statystyce Fermiego–Diraca lub Bosego–Einsteina (→ Termodynamika statystyczna).

Za pomocą operatorów $\hat{c}^+_\alpha$ budujemy z próżni różne stany m-cząstkowe o ustalonych pędach, np.

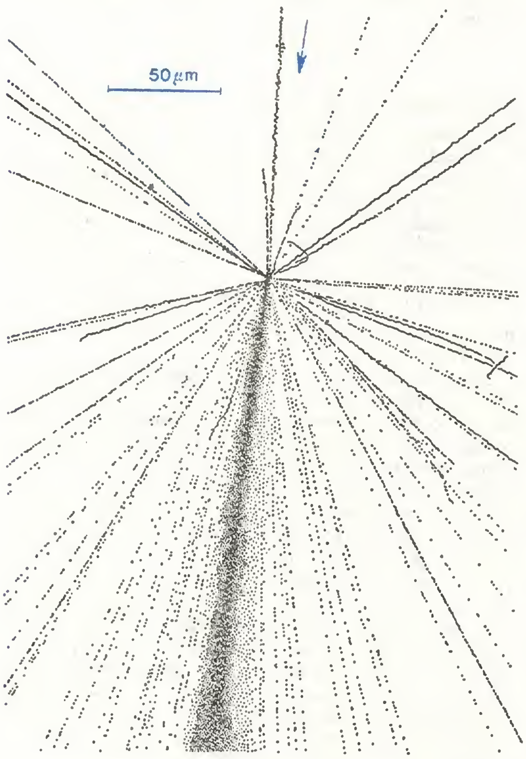
$$|m, \vec{k}\rangle = A(\hat{c}^+_{\alpha_1}(\vec{k}))^{r_1} \ldots (\hat{c}^+_{\alpha_p}(\vec{k}))^{r_p}\Omega,$$

gdzie $r_1 + r_2 + \ldots + r_p = m$, a $\alpha_1, \ldots \alpha_p$ oznaczają różniące się zbiory parametrów. Dla cząstek o spinach połówkowych $r_i = 0$ lub 1 — wynika to ze związków przemienności i nazywa się zakazem Pauliego.

O diagramach Feynmana, renormalizacji i kłopotach z silnymi oddziaływaniami

tworzenie i znikanie cząstek

Doświadczenie sugeruje, że oddziaływania pól wiążą się z tworzeniem i znikaniem cząstek. Na przykład zderzenie neutronu o wielkiej energii z jądrem srebra zarejestrowane zostało w emulsji jądrowej (rys. 7) w postaci „gwiazdy" — tor protonu zakończony jest pękiem torów wychodzących z jednego punktu. To, co wydarzyło się w czasie właściwego oddziaływania na odległościach rzędu 10^{-13} cm i w czasie rzędu 10^{-23} s, jest dla nas tylko źródłem domysłów. Jedno jest pewne: szybko zmieniające się w obrazie Schrödingera stany pola oddziałującego nie mogą być stanami o ustalonej liczbie cząstek. Nie jest w pełni jasne, jak należy konstruować wektory odpowiadające tym stanom, ani jak



Rys. 7. Neutron z promieniowania kosmicznego zderza się z jądrem srebra w emulsji jądrowej. Można rozróżnić 169 naładowanych cząstek, z których co najmniej 130 jest cząstkami wyprodukowanymi. Strzałka oznacza kierunek padającego neutronu, który jako cząstka nienaładowana nie pozostawił śladu w emulsji

przedstawiać operatory pola. Najogólniej postawiony problem to zapisanie w obrazie Heisenberga równań klasycznych oddziałujących pól i szukanie rozwiązań tych równań w postaci funkcji (dystrybucji), których wartościami są operatory działające na przestrzeni Hilberta stanów pól oddziałujących. Podkreślmy, że nie znamy zarówno przestrzeni Hilberta jak i operatorów. Równania oddziałujących pól są skomplikowanymi operatorowymi równaniami nieliniowymi. Dotychczas nikomu nie udało się ich rozwiązać dla realistycznej teorii.

rachunek zaburzeń

Sukcesy i to duże osiągnięto jedynie w wypadku, gdy oddziaływanie między polami jest słabe (jak np. w elektrodynamice kwantowej), dzięki czemu można stosować metodę rachunkową zw. rachunkiem zaburzeń. Metodę tę można stosować wtedy, kiedy gęstość lagrangianu oddziaływania $\mathfrak{L}_I$ zawiera małą stałą sprzężenia oraz kiedy procesy fizyczne obserwowane w przyrodzie wyglądają tak, jakby oddziaływanie się włączało i wyłączało. Innymi słowy — kiedy można wyróżnić trzy fazy: pierwszą i ostatnią, w których występują kwanty pól swobodnych, oraz środkową, w której odbywa się oddziaływanie. Wygodne jest wówczas stosowanie opisu zw. obrazem oddziaływania. W obrazie tym zarówno operatory pola jak i wektory stanu zależą od czasu. Operatory pola spełniają równania pola swobodnego, zaś wektory stanu równanie następujące:

obraz oddziaływania

$$i\hbar\frac{d}{dt}\Phi = \hat{H}_I\Phi, \tag{32}$$

gdzie $\hat{H}_I$, zwany hamiltonianem oddziaływania, wyraża się w prosty sposób przez gęstość lagrangianu oddziaływania $\hat{\mathfrak{L}}_I$, jeśli $\hat{\mathfrak{L}}_I$ nie zależy od pochodnych pól, to

$$\hat{H}_I = -\int_{R^3} :\hat{\mathfrak{L}}_I: d^3x.$$

Podstawowymi wielkościami, które chcemy obliczać, są amplitudy prawdopodobieństwa przejścia A_{fi} ze stanów początkowych $|i\rangle$ do stanów końcowych $|f\rangle$. Amplitudy te można otrzymać w postaci

$$A_{fi} = \langle f|\hat{S}|i\rangle. \tag{33}$$

macierz S

Unitarny operator $\hat{S}$ zw. macierzą rozpraszania (lub macierzą S) można przedstawić w postaci rozwinięcia perturbacyjnego, czyli nieskończonego szeregu zw. perturbacyjnym:

$$\hat{S} = \hat{I} + \sum_{n=1}^{\infty}\left(\frac{i}{\hbar c}\right)^n \frac{1}{n!}\int d^4x_1 \ldots d^4x_n T[:\hat{\mathfrak{L}}_I(x_1): \ldots :\hat{\mathfrak{L}}_I(x_n):], \tag{34}$$

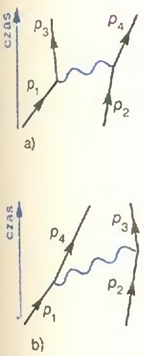
gdzie $\hat{I}$ — macierz jednostkowa, zaś $T[\ldots]$, zw. iloczynem chronologicznym, oznacza czasowe uporządkowanie wyrażeń w nawiasie kwadratowym od największej współrzędnej czasowej (na lewym końcu) do najmniejszej (na prawym). Wyrażenie (34) jest formalnym rozwiązaniem równania (32). Ponieważ w każdym $\hat{\mathfrak{L}}_I(x_i)$ zawarta jest mała stała sprzężenia g, tak więc n-ty wyraz szeregu $\hat{S}_n$ zawiera czynnik $g^n \ll 1$. To jednak nie wystarcza do rozstrzygnięcia, czy formalny szereg (34) jest zbieżny, tzn. czy jego suma jest dobrze określonym operatorem. Aby się o tym przekonać, należałoby oszacować wszystkie wkłady do amplitud $\langle f|\hat{S}_n|i\rangle$, co nie jest możliwe. Dotychczas nie ma dowodu zbieżności szeregu (34), nawet w elektrodynamice. Nie zmienia to faktu, że już kilka pierwszych wyrazów tego szeregu (po renormalizacji, o której za chwilę będzie mowa) daje w elektrodynamice wyniki wspaniale zgodne z doświadczeniem (→ Elektrodynamika, Oddziaływania elektromagnetyczne).

Operator $\hat{S}_n$ po wyrażeniu przez operatory pola staje się kombinacją operatorów kreacji i anihilacji, a po przeprowadzeniu uporządkowania normalnego staje się sumą normalnie uporządkowanych składników. Otrzymujemy sumę, bo na mocy związków przemienności (31) każde przestawienie operatorów pola powoduje konieczność dodania pewnej funkcji, np. dla pola bispinorowego:

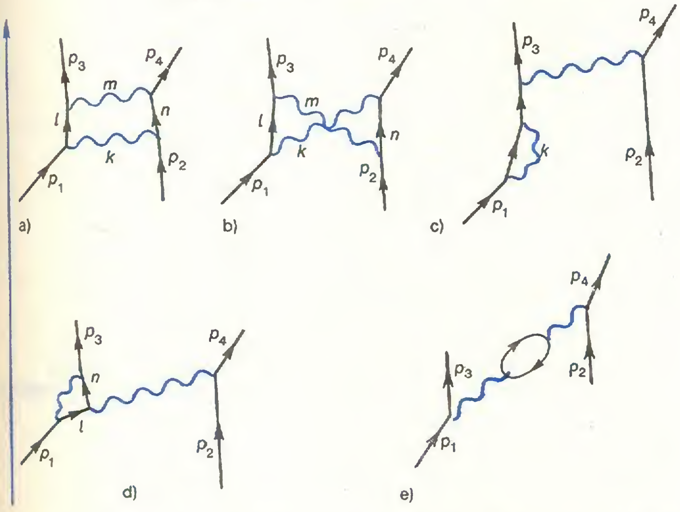
$$\overline{\hat{\psi}}(x_1)\hat{\psi}(x_2) = \hat{\psi}(x_2)\overline{\hat{\psi}}(x_1) + S(x_1 - x_2),$$

gdzie $S(x_1 - x_2)$ jest pewną dobrze określoną funkcją. Tak więc wyrażenie, w którym było n operatorów pola, przechodzi po jednokrotnym przestawieniu dwóch odpowiednich operatorów na sumę dwóch wyrażeń — jednego z n operatorami pola i drugiego z $n-2$ operatorami. Wykonując wszystkie wymagane przez operację : : przestawienia otrzymujemy $\hat{S}_n$ w postaci złożonej sumy. Każdy wyraz tej sumy jest graficznie przedstawiony przez diagram zw. diagramem Feynmana. Ograniczymy się tylko do omówienia diagra-

diagramy Feynmana mów Feynmana dla amplitud A_{fi} (33) w reprezentacji pędowej. Diagramy te są reprezentacją rysunkową składników sumy, którą staje się A_{fi} po uporządkowaniu normalnym każdego $\hat{S}_n$. Diagramy składają się z wierzchołków, linii wewnętrznych i linii zewnętrznych. Rodzaj wierzchołków zależy od gęstości lagrangianu oddziaływania $\hat{\mathcal{L}}_I$. Jeśli $\hat{\mathcal{L}}_I$ jest sumą n gęstości $\hat{\mathcal{L}}^i_I$, $i = 1, ..., n$, takich, że $\hat{\mathcal{L}}^i_I$ jest iloczynem normalnym p_i operatorów pola, to w diagramach występuje n typów wierzchołków takich, że z wierzchołka i-tego typu wychodzi p_i linii. Każdej linii i każdemu wierzchołkowi odpowiada w sposób wzajemnie jednoznaczny wyrażenie analityczne. Rysunki 8 i 9 przedstawiają przykładowo kilka diagramów dla amplitudy opisującej elastyczne rozpraszanie dwóch elektronów. Jako stan początkowy przyjmujemy dwa elektrony o czteropędach p_1 i p_2, a jako stan końcowy — te same elektrony o czteropędach p_3 i p_4 takich, że $p_1+p_2 = p_3+p_4$. Przypominamy, że A_{fi} jest amplitudą prawdopodobieństwa zajścia wyżej omawianego procesu. Wkłady do amplitudy A_{fi} dają wszystkie operatory S_n. Wkłady pochodzące od S_n nazywamy n-tego rzędu względem stałej sprzężenia. Na rys. 8 przedstawione są wszystkie wkłady do omawianego procesu drugiego rzędu, a na rys. 9 — kilka wkładów czwartego rzędu względem stałej sprzężenia e. Diagramy c), d) i e) z rys. 9 nazywa się poprawkami promienistymi do diagramu a) z rys. 8. Całkowanie po czteropędzie c, konieczne przy obliczaniu wyrażeń odpowiadających tym diagramom, prowadzi do wartości nieskończonych. Procedura, która pozwala z tych nieskończoności wydobyć w sposób jednoznaczny wartości skończone, nazywa się renormalizacją.



Rys. 8. Diagramy Feynmana 2. rzędu względem stałej sprzężenia e dla procesu $p_1+p_2 \to p_3+p_4$. Różnym rodzajom linii przyporządkowane są odpowiednie wyrażenia analityczne. Zewnętrznym liniom elektronowym (czarne ze strzałkami) odpowiadają funkcje zależne od czteropędów p_i; wewnętrznej linii fotonowej zw. propagatorem (wężyk niebieski) jest przyporządkowane wyrażenie analityczne zależne od czteropędu tak dobranego, aby w każdym wierzchołku czteropęd był zachowany. Tak więc dla rysunku (a) $k = p_3-p_1$, a dla rysunku (b) $k = p_4-p_1$. Wyrażenie przypisane wierzchołkowi zawiera czynnik e (ładunek elektronu). Po wykonaniu wszystkich obliczeń obu diagramom odpowiadają wyrażenia postaci $\alpha f(p_1, ... p_4)$, gdzie $\alpha = e^2/4\pi\varepsilon_0\hbar c \approx$ $\approx {}^1/_{137}$. Odwrócenie strzałki czasu pozwala interpretować oba grafy jako wkłady do amplitudy elastycznego rozpraszania pozytonów



Rys. 9. Kilka diagramów 4. rzędu względem stałej sprzężenia e (drugiego rzędu względem stałej sprzężenia α). Na diagramach tych pojawia się nowy element — wewnętrzna linia elektronowa, propagator elektronu (czarna linia prosta ze strzałką). Po czteropędach związanych z liniami wewnętrznymi zawsze musi być wykonane całkowanie z uwzględnieniem ograniczeń wynikających z zachowania czteropędu w wierzchołkach. Ograniczenia te umożliwiają w grafach a) i b) wykonanie całek po wszystkich czteropędach k, l, m, n, zaś w grafach c), d) i e) pozostaje jedna całka do wykonania, np. po czteropędzie l (bez żadnych ograniczeń na wartości l). Całki te są rozbieżne. W diagramie e) występuje również wewnętrzna linia pozytonowa — jest to linia, na której strzałka ma przeciwny kierunek niż kierunek biegu czasu

renormalizacja Renormalizację przeprowadza się poczynając od modyfikacji rozbieżnych całek przez wprowadzenie odpowiednich parametrów, tak aby całki te były zbieżne, a rozbieżności można było otrzymać przez odpowiednie przejście graniczne z wprowadzonymi parametrami — jest to tzw. regularyzacja. Ostatnio najczęściej używa się regularyzacji wymiarowej, która polega na potraktowaniu czteropędu, po którym występuje całkowanie w rozbieżnej całce, jako wektora n-wymiarowego, $n > 4$ (dla $n = 4$ otrzymuje się rozbieżność). Ze zregularyzowanej całki staramy się wydobyć część, która w granicy, gdy $n \to 4$, pozostaje skończona. W elektrodynamice w jednoznacznym wydobyciu tej części pomaga żądanie, aby teoria była niezmiennicza ze względu na transformacje cechowania pól. Aby usunąć część osobliwą przy przejściu $n \to 4$ dodajemy do wyjściowej gęstości lagrangianu odpowiedni wyraz $\mathcal{L}_c$, zw. kontrczłonem, tak aby generował on diagram, którego osobliwa część znosiłaby się z osobliwą częścią diagramu renormalizowanego. Procedurę tę można przeprowadzić przy dowolnej gęstości $\mathcal{L}_I$. W celu usuwania coraz to nowych osobliwości pojawiających się w wyższych rzędach rachunku zaburzeń trzeba w ogólności wprowadzać nieskończenie wiele kontrczłonów. Wprowadzenie każdego kontrczłonu prowadzi do pojawienia się w teorii stałej dowolnej. Jeśli wystarczy wprowadzić skończoną liczbę typów kontrczłonów, teorię nazywamy renormalizowalną.

W elektrodynamice wyrażenia operatorowe w kontrczłonach są tego samego typu, co w wyjściowej gęstości lagrangianu $\mathcal{L}_0+\mathcal{L}_I$. W rezultacie dodanie kontrczłonów zmienia parametry stojące w wyjściowej teorii przy wyrażeniach operatorowych, np. zamiast masy m stojącej w elektrodynamice przy wielkości $\bar{\psi}\psi$ we wzorze (12) pojawi się $m' = m-\sum_i \delta_i(m)$. Stałe $\delta_i(m) > 0$ dążą do nieskończoności, gdy $n \to 4$. Widzimy, że wyjściowy parametr m zw. gołą masą musi być nieskończony, aby m' zw. masą zrenormalizowaną mógł być skończony. Ponieważ masę elektronu znamy z doświadczenia, przyjmujemy, że m' to ta dobrze znana masa. Podobnie dzieje się z ładunkiem. Teorię renormalizowalną można tak wprowadzić, aby renormalizacja sprowadzała się tylko do zmiany mas i stałych sprzężenia. Warunek renormalizowalności narzucony na kwantową teorię pola jest bardzo ograniczający. Wyróżnia on kilka typów teorii wśród nieskończonej liczby teorii pola, jakie by można sformułować. Nieprzypadkowo okazuje się, że teorie renormalizowalne odgrywają dużą rolę w fizyce.

Nie będziemy omawiać sukcesów elektrodynamiki kwantowej, gdyż są one omówione w innych artykułach. Przejdziemy teraz do porównania otrzymanego formalizmu obliczania amplitud A_{fi} z obrazem fizycznym oddziaływania elektronów z fotonami.

poglądowy obraz oddziaływania W klasycznym polowym obrazie oddziaływania elektronów oddziałują one za pośrednictwem pola, tak więc nie widać powodów, dla których całkowita energia elektronów miałaby być zachowana. Podstawowym procesem jest proces, w którym energia ta maleje, ponieważ w wyniku oddziaływania jest ona przekazywana polu, inaczej mówiąc — wypromieniowana. To właśnie jest przyczyną, że klasyczny atom wodoru nie może być trwały. Z drugiej strony wiemy, że obserwuje się elastyczne rozpraszanie elektronów i innych ładunków. Elastyczne rozpraszanie elektronów jest możliwe wtedy i tylko wtedy, gdy całkowita energia i pęd przekazane polu przez jeden elektron są pobrane z pola przez drugi elektron. Wiedząc już, że przekazy energii i pędu są skwantowane, można powiedzieć, że elektron zmienia pęd i energię emitując odpowiedni kwant pola, czyli foton, który zostaje pochłonięty przez drugi elektron. Foton ten jest przeważnie fotonem wirtualnym, tzn. jego czteropęd k^μ nie spełnia warunku $k_\mu k^\mu = 0$, który to warunek określa **foton wirtualny**

fotony swobodnej fali elektromagnetycznej. W istocie $k^\mu = (p_f)^\mu - (p_i)^\mu$, gdzie $(p_i)^\mu$ i $(p_f)^\mu$ określają początkowy i końcowy czteropęd elektronu, czyli

$$k_\mu k^\mu = (p_i)_\mu (p_i)^\mu + (p_f)_\mu (p_f)^\mu - (2p_i)_\mu (p_f)^\mu$$

i zwykle $k_\mu k^\mu \neq 0$. Tak więc widzimy, że podstawowymi aktami oddziaływania są emisje i absorpcje fotonów. Elektron może również absorbować fotony, które wcześniej sam wyemitował. Elastyczne rozpraszanie elektronów: $p_1+p_2 \to p_3+p_4$, może wystąpić w wyniku wielu z tych procesów. Każdemu z nich przyporządkowana jest amplituda prawdopodobieństwa. Jeśli dany stan końcowy f może być osiągnięty ze stanu początkowego i w wyniku różnych, nierozróżnialnych w eksperymencie procesów fizycznych, to amplitudę prawdopodobieństwa A_{fi} należy przedstawić jako sumę amplitud (21) odpowiadających każdemu z tych procesów. Widzimy, że otrzymana w elektrodynamice kwantowej amplituda reakcji w postaci sumy ma żądaną strukturę. Diagramy odpowiadające wyrazom sumy umożliwiają stworzenie prostego obrazu fizycznego, który nazwiemy procesem wirtualnym. Musimy tu podkreślić z całą mocą, że proces wirtualny nie daje żadnej informacji dlaczego i w jaki sposób należy obliczać odpowiadającą mu amplitudę. Informacja ta zawarta jest tylko w formalizmie matematycznym. Proces wirtualny nie jest realnym procesem, gdyż przy założeniu, że każdy obserwowany przypadek jest konkretną realizacją jednego z tych procesów, całkowite prawdopodobieństwo reakcji powinno być sumą prawdopodobieństw, z których każde byłoby prawdopodobieństwem otrzymania stanu końcowego w wyniku odpowiedniego procesu wirtualnego. Dodawanie prawdopodobieństw prowadzi często do innych wyników niż dodawanie amplitud prawdopodobieństwa. Tak więc zwrotu: elektrony oddziałują ze sobą wymieniając fotony — nie należy rozumieć dosłownie.

Obraz fizyczny okazuje się jednak pomocny nie tylko wówczas, gdy mówimy o zjawiskach fizycznych, lecz również przy formułowaniu nowych ściślejszych modeli. Po tym, jak stwierdzono istnienie krótkozasięgowego (rzędu 10^{-13} cm) oddziaływania wiążącego silnie protony i neutrony w jądrze atomowym, H. Yukawa zaproponował opisywanie tego oddziaływania przez wymianę hipotetycznych kwantów pola, które nazwano mezonami.

oszacowanie masy mezonu

Proste rozumowanie, prowadzące do oszacowania masy mezonu, jest na tyle ciekawe, że warto je tu przytoczyć. Wyobraźmy sobie spoczywające dwa nukleony w odległości od siebie R rzędu 10^{-12} cm (z dobrym przybliżeniem nukleony w jądrze spoczywają). Jeśli ich oddziaływanie polega na wymianie mezonu, to w stanie pośrednim powinny istnieć dwa nukleony i dodatkowo jeszcze mezon. Proces tworzenia tego stanu nie jest możliwy w fizyce klasycznej, naruszone byłoby bowiem prawo zachowania energii. Energia musi się zmienić o wielkość ΔE, która co najmniej jest równa energii spoczynkowej mezonu mc^2 ($\Delta E \geqslant mc^2$). Natomiast w mechanice kwantowej udowadnia się, że dokładność określenia energii ΔE w czasie Δt jest ograniczona od dołu $\Delta E \geqslant \hbar/\Delta t$. Ponieważ prędkość wirtualnego mezonu jest rzędu prędkości światła c, to $\Delta t \approx R/c$. Stąd $mc^2 \approx \hbar c/R$, co daje masę m równą ok. 250 masom elektronu. Ponieważ możliwe są procesy $p+n \to n+p$, $n+p \to p+n$ i $p+p \to p+p$, mezony muszą być trzech rodzajów i nieść ładunek 0, $+e$ i $-e$. Dopiero w 12 lat po tym, jak Yukawa wysunął swoją hipotezę, odkryto cząstki o zbliżonych własnościach — mezony π; masa ich wynosi 273 masy elektronu.

kłopoty z modelem Yukawy

Ścisły opis polowy oddziaływań silnych napotkał poważne trudności. Oddziaływania te są na tyle silne, że zrenormalizowana stała sprzężenia występująca w gęstości lagrangianu oddziaływania Yukawy $\mathcal{L}_I$ wynosi ok. 10, czyli nie jest mała. Stosowanie rachunku zaburzeń wydaje się kompletnie nieuzasadnione, a w dodatku praktycznie nieprzydatne. Poza tym liczba zaobserwowanych cząstek elementarnych, a szczególnie hadronów jest tak wielka, że opis, w którym każdej z tych cząstek odpowiadałoby swoiste pole kwantowe, a w $\mathcal{L}_I$ występowałyby wszystkie możliwe sprzężenia typu Yukawy tych pól, wydaje się niewłaściwy. Nie wiadomo też, co zrobić z bardzo dużą liczbą cząstek zw. rezonansami. Te trudności spowodowały sceptycyzm i odejście od metod teorii pola w teorii oddziaływań silnych w latach 50-ych i 60-ych (→ Oddziaływania silne). Mimo to obraz Yukawy, według którego nukleon otoczony jest chmurą wirtualnych mezonów, został przyjęty przez ogół fizyków i do dziś większość danych na temat rozpraszania mezonów π na sobie otrzymuje się interpretując pewne oddziaływania mezonów π z nukleonami jako oddziaływania tych mezonów z mezonami chmury wirtualnej.

Bardzo płodną okazała się koncepcja M. Gell-Manna, zgodnie z którą wszystkie cząstki elementarne oddziałujące silnie są zbudowane z 3 podstawowych cząstek o ładunkach ułamkowych i spinach połówkowych, zw. kwarkami. Obecnie najczęściej przyjmuje się, że obiektów tych jest 18 (→ Cząstki elementarne i ich oddziaływania). Kwarki różnią się kolorami (trzy kolory) i zapachami (których jest co najmniej cztery, a ostatnio są dane, aby sądzić, że jest ich sześć).

uwięzienie kwarków

Zadziwiające jest to, że w wielu procesach hadrony zachowują się tak, jakby były złożone ze swobodnych i lekkich kwarków, natomiast mimo usilnych poszukiwań swobodnych kwarków, nie udało się ich odkryć. Zjawisko to nazywa się uwięzieniem kwarków. Istnieje prosty model klasyczny mezonu — punkty materialne połączone elastyczną struną o długości spoczynkowej l_0. Kiedy struna nie jest napięta, punkty zachowują się tak, jak swobodne. Przy próbie oddalenia punktów od siebie na odległość większą od l_0 struna wytwarza siłę o potencjale wprost proporcjonalnym do kwadratu jej wydłużenia, a przy stałym napięciu — o potencjale wprost proporcjonalnym do jej wydłużenia. Jeśli założymy, że odpowiednio napięta struna może pęknąć, przy tym na nowych końcach pojawią się natychmiast odpowiednie punkty materialne, to obraz ten będzie bliski „obrazowi polowemu" mezonu. W obrazie tym rolę punktów odgrywają kolorowe kwarki i antykwarki, a oddziaływanie typu struny jest generowane w niecałkiem jeszcze zrozumiany sposób przez wymiany kolorowych bezmasowych wektorowych gluonów, będących kwantami pól cechowania typu Yanga–Millsa, związanymi z grupą cechowania $SU(3)$. Nie musimy już powtarzać, że poglądowe obrazy są dość złudne. Poza tym prawidłowość tego przedstawienia nie została wykazana za pomocą żadnego ścisłego matematycznego modelu. Istnieje jednak nadzieja, że zjawisko uwięzienia kwarków będzie w ten sposób wyjaśnione. Nadzieja ta jest oparta na obserwacji, że efektywny parametr występujący w perturbacyjnym rozwinięciu zależy od energii. W renormalizowalnych teoriach Yanga–Millsa parametr ten dąży do zera, jeśli energia dąży do nieskończoności, natomiast rośnie, gdy energia maleje. Procesy wysokoenergetyczne wydają się być związane z oddziaływaniem na bardzo małych odległościach. Procesy niskoenergetyczne zaś — z dużymi odległościami.

niewola podczerwona

Obszar niskich energii nazywa się często obszarem podczerwonym (przez analogie z optyką i z uwagi na związek między energią i częstością $E = h\nu$). Wzrost parametru rozwinięcia interpretuje się jako wzrost siły oddziaływań, tak więc wydaje się, że efekt ten wiąże się z uwięzieniem kwarków i został nazwany przez fizyka amerykańskiego S. Glashowa niewolą podczerwoną (ang. *infrared slavery*). Droga od tych przypuszczeń do ścisłego udowodnienia, że kwarki muszą być uwięzione na dużych, a swobodne — na małych odległościach, wydaje się jeszcze daleka. Pojawia się przy tym jeszcze jedna poważna trudność techniczna.

problem ze stanami związanymi

Istnieją w teorii pola efektywne algorytmy obliczania amplitud rozpraszania procesów, w których zmienia się liczba swobodnych kwantów pola, oraz oblicza-

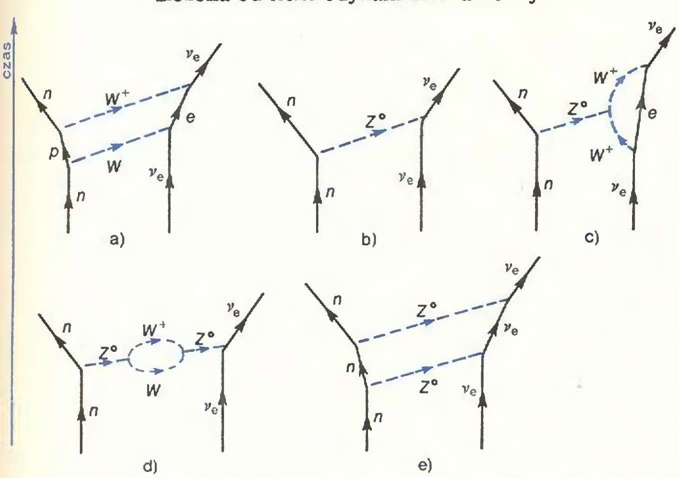
nia czasów życia cząstek nietrwałych. Nie ma natomiast, jak dotąd, algorytmów obliczeń dotyczących procesów rozpraszania stanów związanych, a w modelu kwarkowym wszystkie hadrony są uważane za takie stany.

renormalizowalna teoria oddziaływań słabych

Nowy zwrot w kierunku metod polowych w teorii oddziaływań silnych jest wynikiem znalezienia renormalizowalnej teorii oddziaływań słabych. Teoria ta omawiana jest w artykule „Oddziaływania słabe", tu podkreślimy kilka jej aspektów. Jak już wspominaliśmy, pole elektromagnetyczne ma interpretację pola cechowania stowarzyszonego z grupą cechowania $U(1)$. Niezmienniczość ze względu na tę grupę jest cechą bardzo istotną w przeprowadzaniu renormalizacji. Poszukując renormalizowalnej teorii oddziaływań słabych natrafiono na renormalizowalną teorię pól cechowania Yanga–Millsa, opartą na grupie cechowania $SU(2)$. Występują w niej trzy bezmasowe wektorowe pola cechowania (dwa naładowane, jedno neutralne). Można sobie łatwo wyobrazić, że oddziaływania słabe polegają na wymianie cząstek wektorowych naładowanych. Okazuje się jednak, że odpowiednie cząstki powinny mieć masę i to masę większą od masy protonu ok. 30 razy. Wprowadzenie członu masowego do teorii Yanga–Millsa łamie symetrię cechowania i psuje renormalizowalność teorii. Po wielu trudach okazało się, że jeśli się wprowadzi do teorii dodatkowe pole skalarne (pole Higgsa) i założy, że stan próżni nie jest niezmienniczy ze względu na grupę cechowania, to pola cechowania mogą zyskać masę i teoria pozostanie renormalizowalna. Mechanizm wprowadzania w ten sposób masy nazywa się mechanizmem Higgsa.

model Weinberga–Salama

Wyniki te leżą u podstaw teorii Weinberga–Salama, w której zakłada się istnienie trzech nowych wektorowych cząstek: W^+ i W^- o masach większych od 39,8 mas protonu oraz Z^0 o masie większej od 79,6 mas protonu. Wprowadzenie cząstki Z^0 implikuje istnienie nowych reakcji, np. rozpraszania elektronów przez neutrina mionowe. Reakcje te zostały zaobserwowane, co było pierwszym wielkim sukcesem modelu. Pełnym sukcesem teorii byłoby odkrycie cząstek $W^\pm$ i Z^0. Na rys. 10 przedstawione są diagramy Feynmana, dające po renormalizacji wkład najniższego rzędu do procesu $n+\nu_e \rightarrow n+\nu_e$ w teorii Weinberga–Salama. Pozwala to zorientować się, o ile teoria ta jest bardziej złożona od elektrodynamiki kwantowej.



Rys. 10. Diagramy, które muszą być uwzględnione, aby po renormalizacji otrzymać skończony wkład najniższego rzędu do amplitudy rozpraszania elastycznego neutrina elektronowego na neutronie. W odróżnieniu od rys. 8 i 9 zaznaczony jest tylko rodzaj wymienianych cząstek

Łamanie symetrii cechowania przez stan próżni (zw. spontanicznym łamaniem symetrii) wskazuje na to, że próżnia — stan pola o najniższej energii — ma ciekawe własności fizyczne. Nie powinien nas wprowadzać w błąd fakt, że energia i pęd tego stanu wynoszą 0. Po prostu wprowadzając w definicji operatorów energii i pędu uporządkowanie normalne przyjęliśmy energię próżni za punkt zerowy na skali wartości energii.

własności próżni

Możliwość zmiany masy cząstki dzięki wprowadzeniu odpowiednich własności próżni skłania do przyjęcia obrazu próżni jako pewnego skomplikowanego ośrodka. Wiadomo, że pole elektromagnetyczne może wnikać do nadprzewodnika tylko bardzo płytko. Interpretuje się to często jako zjawisko zyskiwania masy przez fotony w ośrodku nadprzewodzącym, a stąd już krok do analogii z mechanizmem Higgsa. Z kolei to, że wirtualny foton emitowany przez ładunek może wytworzyć wirtualną parę elektron-pozyton (co efektywnie prowadzi do renormalizacji wyjściowego ładunku), przypomina zjawisko polaryzacji dielektryka; nazywa się je dlatego polaryzacją próżni. Tak więc stan próżni jest skomplikowanym stanem pola, którego zrozumienie jest pomocne przy formułowaniu nowych ścisłych modeli.

O optymizmie, wątpliwościach i głównych kierunkach badań

Optymiści sądzą, że kwantowa teoria pola łącznie z procedurą renormalizacji jest tym ogólnym schematem, w ramach którego zostanie znaleziony jednolity opis zjawisk we Wszechświecie.

Tak więc cele, które sobie stawiają, to: znalezienie podstawowych pól i odpowiadających im gęstości lagrangianu, rozwiązanie problemu stanów związanych i uwięzienia kwarków oraz stworzenie efektywnych algorytmów obliczania wszystkich fizycznie interesujących wielkości. Niektórzy natomiast sądzą, że podstawowym pojęciem przyszłej teorii będą pola cechowania Yanga–Millsa. Teoria tych pól jest skomplikowana, dlatego bada się często modelowe teorie pól w czasoprzestrzeni o zmniejszonej liczbie wymiarów lub na nieskończonych dyskretnych sieciach czasoprzestrzennych. Badania te stanowią obecnie główny nurt w teorii pola. Wynika z nich m.in. konieczność modyfikacji ogólnej teorii względności dla małych odległości, ponieważ obecna teoria pola grawitacyjnego po skwantowaniu daje teorię nierenormalizowalną.

Drugą grupę badań, którą charakteryzuje wysoka ścisłość i zaawansowanie metod matematycznych, stanowią prace nad aksjomatyczną teorią pola. Celem ich jest takie sformułowanie elektrodynamiki i innych znanych teorii pola, aby od początku nie występowały w nich żadne rozbieżności. Poszukuje się wyników, które można uzyskać w sposób ścisły, a w wypadku stosowania rachunku zaburzeń bada się zbieżność szeregu perturbacyjnego. Ten kierunek badań rozwinął się w wyniku niezadowolenia ze zwykłej procedury renormalizacji, która bardziej przypomina przepis rachunkowy niż element podstawowej teorii (podkreślał to wielokrotnie P.A.M. Dirac, jeden z twórców kwantowej teorii pola). Można zaliczyć do tej grupy pojawiające się ostatnio prace poświęcone kwantowej teorii solitonów. Ponieważ solitony cieszą się dużym zainteresowaniem, zamieścimy o nich krótką dygresję.

solitony

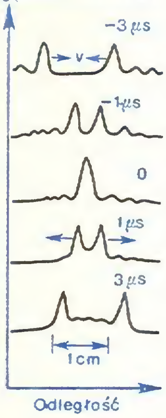
Zacznijmy od historii. Pierwszy soliton został zaobserwowany w 1834 r. przez J. Russela. Przejeżdżając konno wzdłuż wąskiego kanału zauważył on, że gwałtowne zatrzymanie się barki spowodowało powstanie fali o wysokości ok. 30 cm i długości ok. 9 m, która bez zmiany kształtu przesuwała się z jednakową prędkością kilkunastu kilometrów na godzinę wzdłuż kanału. Russel towarzyszył tej fali na koniu na odcinku ok. 2 km. Liczne doświadczenia potwierdziły tę obserwację. Stwierdzono, że im wyższa fala, tym szybciej się porusza. Zjawisko to, zw. falą samot-

ną, zostało analitycznie wyjaśnione przez otrzymanie w 1895 r. rozwiązania (o żądanych własnościach) nieliniowego równania różniczkowego cząstkowego:

równanie Kortewega–de Vriesa

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^3 u(x,t)}{\partial x^3} = 0,$$

które nazwano od nazwisk twórców równaniem Kortewega–de Vriesa (K–dV). W roku 1965 M. Kruskal i N. Zabusky wykazali, że dwa rozwiązania równania K–dV odpowiadające dwóm samotnym falom skierowanym ku sobie, przechodzą przez siebie bez deformacji, zachowując stabilność (rys. 11). Fale samotne o tej właściwości nazwano solitonami, a rozwiązanie równania nieliniowego, rozpadające się dla $t \to -\infty$ i $t \to +\infty$ na N rozwiązań solitonowych — rozwiązaniem N-solitonowym. Znaleziono również inne równania dopuszczające rozwiązania solitonowe. Bardzo ważne z nich jest skalarne równanie falowe w czasoprzestrzeni (1+1)-wymiarowej (1 zmienna przestrzenna i czas) z członem nieliniowym sin Φ zw. równaniem sG (sinusowym Gordona). Równanie to opisuje ruch defektów w ośrodkach uporządkowanych, takich jak np. ruch zlokalizowanych dyslokacji w kryształach. Równanie K–dV spotyka się np. przy opisie płytkich fal wodnych, jonowych fal akustycznych, fal Alfvéna w zimnej plazmie itp. Równanie sG można traktować jako nieliniowe równanie pola w 1+1 wymiarach, podejmuje się więc próby kwantowania tego pola. Odkrycie i badanie solitonów to postęp w teorii nieliniowych równań cząstkowych oraz postęp w rozumieniu różnych nieliniowych zjawisk fizycznych. Soliton w 1+1 wymiarach jest pod wieloma względami podobny do cząstki elementarnej: gęstość jego energii jest zlokalizowana, całkowita energia jest skończona, zderza się zachowując swą tożsamość, istnieją rozwiązania równania sG odpowiadające stanom związanym soliton-antysoliton (ang. *breather* 'dychacz') itp. Niektórzy chcieliby tę analogię rozwinąć i wierzą, że w 3+1 wymiarach zaobserwuje się w pewnej klasie równań produkcję solitonów przy zderzeniach lub ich rozpraszanie, co uwoźliwi stworzenie obrazu cząstki elementarnej jako solitonu. Program ten napotyka na razie trudności zarówno natury matematycznej, jak i fizycznej.



Rys. 11. Zderzenie akustycznych solitonów w plazmie. Przemieszczanie się zagęszczeń jonów w plazmie opisuje równanie K–dV. Rysunek przedstawia ewolucję w czasie rozwiązania dwusolitonowego

Trzecia grupa badań to próby wyjścia poza lokalną teorię pola. Motywacja tych badań może być różnorodna. Już na gruncie klasycznym widzieliśmy, że pomiar pola w punkcie jest idealizacją (zbiory obserwowanych zdarzeń są zawsze dyskretne, a odrzucenie założenia ciągłości czasoprzestrzeni i konstrukcja czasoprzestrzeni dyskretnych nie daje, jak dotychczas, zadowalających wyników). Idealizacją jest również wprowadzenie punktowych ładunków. Pole punktowego ładunku jest osobliwe w punkcie, w którym ładunek ten się znajduje: klasyczna energia samooddziaływania jest nieskończona. Osobliwość ta znajduje odbicie w nieskończonościach pojawiających się w kwantowej teorii pola. Wydaje się, że najprostszym sposobem zniknięcia tych nieskończoności byłaby odpowiednia modyfikacja teorii klasycznej na małych odległościach, a następnie skwantowanie zmodyfikowanej teorii. Elegancką i niesprzeczną modyfikacją elektrodynamiki klasycznej jest nieliniowa teoria Borna–Infelda, jednak trudności pojawiają się przy jej kwantowaniu, podobnie jak przy kwantowaniu każdej teorii, w której nieliniowość odgrywa istotną rolę. Wraz z rezygnacją z punktowości pojawia się możliwość wystąpienia tzw. oddziaływań nielokalnych. Wiąże się to z możliwością równoczesnego oddziaływania z różnymi punktami obiektu rozciągłego. Istnieją metody kwantowania nielokalnych teorii pola i próbowano już nawet zastosować takie teorie do opisu hadronów.

Założenie punktowości elektronów w elektrodynamice, stosowanej do opisu procesów atomowych przy odległościach między ładunkami rzędu 10^{-8} cm, wydaje się przybliżeniem rozsądnym. Natomiast bogata struktura hadronów i bardzo krótki, ale dobrze określony zasięg oddziaływań silnych rzędu 10^{-13} cm uniemożliwiają założenie punktowości hadronów. Dlatego podejmuje się próby kwantowania rozciągłych klasycznych układów, takich jak struny, membrany i worki. Przy próbach porównania otrzymanych modeli z doświadczeniem pojawia się konieczność wprowadzenia dodatkowych upraszczających założeń, co stawia pod znakiem zapytania większość otrzymywanych tym sposobem wyników.

struktura hadronów

Zwolennicy lokalnej teorii pola utrzymują, że obraz punktowej cząstki otoczonej „chmurą" wirtualnych punktowych kwantów oraz obraz stanu związanego punktowych kwarków, jakie ta teoria daje, dobrze uwzględnia strukturę cząstek.

Można sobie wyobrazić bardziej drastyczne odejście od założenia punktowości hadronów przy odległościach 10^{-13} cm niż to ma miejsce w teoriach nielokalnych czy też w obrazie chmury wirtualnych kwantów. Klasycznym przykładem drastycznego załamania założenia punktowości jest przejście od opisu ruchu meteorytu (z dobrym przybliżeniem punktowego) do opisu jego zderzenia z powierzchnią Ziemi. Do opisu procesu tworzenia się potężnego leja (np. słynny krater w Arizonie) trzeba używać metod zupełnie różnych matematycznie i fizycznie od równań Newtona dla punktów materialnych. W przykładzie tym charakterystyczny jest brak ciągłego przejścia pomiędzy tymi dwoma opisami fizycznymi. Wątpliwość co do stosowalności teorii kwantowej do opisu zjawisk fizycznych na odległościach 10^{-13} cm wysunął pierwszy W.C. Heisenberg już w latach 30-ych. Zaproponował on wprowadzenie nowej uniwersalnej stałej fizycznej (o wymiarze długości) — tzw. długości elementarnej — takiej, że przy odległościach mniejszych od tej stałej powinna nastąpić zmiana sposobu opisu zjawisk. Wobec wspomnianych wątpliwości i trudności dotyczących polowego opisu oddziaływań silnych zaczęto konstruować modele innego rodzaju (jak np. różne modele statystyczne i hydrodynamiczne stosowane z dużym niekiedy powodzeniem do opisu produkcji wielorodnej przy wysokoenergetycznych zderzeniach hadronów).

długość elementarna

Tak więc pesymiści nie wierzący w ostateczny sukces badań nurtu głównego mogą konstruować modele niezależne i porównywać je z doświadczeniem, mogą też sprawdzać prawdziwość najbardziej ogólnych twierdzeń dowodzonych na gruncie standardowego podejścia do lokalnej kwantowej teorii pola oraz poszukiwać nowych subtelnych efektów doświadczalnych, wynikających z obrazu inspirowanego istnieniem długości elementarnej. Badania te wiążą się w pewien sposób ze współczesnymi badaniami podstaw mechaniki kwantowej w duchu wątpliwości Einsteina, Schrödingera, de Broglie'a i Bohma, którzy kwestionowali traktowanie mechaniki kwantowej jako najbardziej kompletnej teorii pojedynczych mikroukładów fizycznych (→ O niektórych podstawowych pojęciach fizycznych).

Cały program tych pesymistycznych badań nie cieszy się wielkim uznaniem i sympatią optymistów, choć na tej drodze mogą być znalezione dodatkowe argumenty zarówno teoretyczne jak i doświadczalne przemawiające za słusznością czy też koniecznością opisu wszystkich zjawisk w obrębie konwencjonalnie rozumianej teorii kwantowej.

I. Białynicki-Birula *Wstęp do teorii pól kwantowych*, Warszawa 1971; I. Białynicki-Birula, Z. Białynicka-Birula *Elektrodynamika kwantowa*, Warszawa 1974; W.B. Bierestecki i in. *Relatywistyczna teoria kwantów* cz. 1 Warszawa, 1972; L.N. Cooper *Istota i struktura fizyki*, Warszawa 1975; R.P. Feynman i in. *Feynmana wykłady z fizyki*, t. 3 Warszawa 1974; M. Kupczyński *Odkrycie powabu*, Post. Fiz. 28, 275 (1977); E.M. Lifszyc i L.P. Pitajewski *Relatywistyczna teoria kwantów* cz. 2, Warszawa 1973; L.I. Schiff *Mechanika kwantowa*, Warszawa 1977; A. Szymacha *Zunifikowane teorie oddziaływań słabych i elektromagnetycznych* Post. Fiz. 27, 117 (1976); E.H. Wichmann *Fizyka kwantowa* Warszawa 1975.

CZĄSTKI ELEMENTARNE I FIZYKA WIELKICH ENERGII

Cząstki elementarne i ich oddziaływania · Struktura cząstek elementarnych · Atomy egzotyczne · Detekcja cząstek · Akceleratory Oddziaływania silne · Oddziaływania elektromagnetyczne · Oddziaływania słabe

Cząstki elementarne i ich oddziaływania

Grzegorz Białkowski

pojęcie cząstki elementarnej

Cząstki elementarne są to obiekty fizyczne, z których, według obecnego stanu wiedzy, składają się wszystkie ciała materialne oraz rozmaite rodzaje promieniowania. Ujmując rzecz historycznie, za jedną z podstawowych cech cząstek elementarnych uważano ich niepodzielność, co jest wyrażone w greckim terminie atom — *atomos* — niepodzielny. Niepodzielność cząstki elementarnej można rozumieć dwojako: albo jako nierozbijalność (a więc trwałość), albo też jako brak struktury wewnętrznej (a więc niezłożoność). Oba możliwe kryteria elementarności cząstek, oparte na tych dwu cechach, są jednak zawodne. Wiadomo dziś, że obiekty uważane za cząstki elementarne mają strukturę wewnętrzną, są więc tworami złożonymi. Poza tym, w przyrodzie istnieje tylko bardzo nieliczna grupa cząstek trwałych (lub też mających tak długie czasy życia, że obecnie uważa się je za trwałe).

Trwałymi cząstkami elementarnymi są: proton, elektron, foton oraz neutrina elektronowe i mionowe.

Trwałości cząstki nie można uważać za kryterium elementarności z następujących powodów. Po pierwsze, liczne obiekty, których nieelementarność jest pewna (np. jądra atomowe lub atomy w stanie podstawowym), są trwałe. Po drugie, trwałe cząstki elementarne mają zwykle „kuzynów" nietrwałych; pokrewieństwo cząstek trwałych i nietrwałych jest tak wyraźne, że logicznie jest jedne i drugie uznać za elementarne. Na przykład proton i neutron tworzą parę cząstek elementarnych, mimo iż proton jest trwały, a neutron — nie. Po trzecie wreszcie, zdarza się często, iż w wyniku niektórych oddziaływań cząstki elementarne trwałe przemieniają się w inne cząstki, nietrwałe. W takich wypadkach trudno jest uznać, że cząstki trwałe stają się składnikami obiektów nietrwałych. Przykładem takiego procesu może być np. anihilacja pary proton-antyproton w kilka mezonów π, czyli w cząstki nietrwałe.

W tej sytuacji nie jest łatwo podać proste kryterium elementarności cząstki, odwołujące się wprost do faktów eksperymentalnych. Nasuwa się więc myśl, aby za elementarne uznać te i tylko te cząstki, których istnienie jest z punktu widzenia teorii warunkiem koniecznym występowania wszystkich innych obiektów fizycznych. Praktyczne stosowanie tego kryterium byłoby jednak możliwe tylko wtedy, gdyby istniała pełna, sprawna rachunkowo i przynosząca jednoznaczne odpowiedzi teoria cząstek. Mimo, że opracowaniem takiej teorii zajmuje się wielu najwybitniejszych fizyków doby dzisiejszej, teoria ta jeszcze nie została sformułowana. W najlepszym razie można uznać, że stworzono pewne teorie cząstkowe, które zapewne wejdą w skład teorii kompletnej. Ostatecznie istnieje możliwość zdefiniowania cząstek elementarnych przez ich wyliczenie. Najlepiej zbadane cząstki są wymienione w załączonej tabeli.

Już pierwszy rzut oka na tę tabelę pozwala stwierdzić, że sytuacja nie jest najlepsza, gdyż obiektów elementarnych jest bardzo dużo, a liczba ich przy tym wzrasta z roku na rok. Ponieważ zaś, jak stwierdziliśmy poprzednio, żadna ze znanych cząstek nie wydaje się być bardziej elementarna od innych, pozostają jeszcze dwie inne możliwości poradzenia sobie z tym chaosem. Po pierwsze, można przyjąć, że rzeczywiście wszystkie obiekty uznawane obecnie za cząstki elementarne są naprawdę równie elementarne w tym sensie, że każda z nich wymaga do swojego istnienia wszystkich innych cząstek, z których jest w jakiś sposób zbudowana. Ta hipoteza zwana hipotezą demokracji cząstek nie zyskała sobie powszechnego uznania, m.in. w związku z trudnościami rachunkowymi występującymi przy próbach jej sprawdzenia. Po drugie, nasuwa się myśl, że może żadna z cząstek znanych teraz jako elementarne, naprawdę elementarna nie jest, i że taką prawdziwie elementarną cząstkę czy cząstki należy dopiero wykryć. Za „kandydatów" na takie cząstki prawdziwie elementarne uważa się m.in. kwarki.

hipoteza demokracji cząstek

kwarki

Hipoteza kwarków jest obecnie niemal powszechnie uważana za słuszną. Powstaje więc pytanie, czy same kwarki są obiektami elementarnymi, i czy dalsze badania nie doprowadzą do wniosku, że należy szukać cząstek jeszcze bardziej elementarnych. W tym wypadku to, co uważamy za elementarne byłoby funkcją niewiedzy; elementarne jest to, o czym jeszcze nie wiemy, że naprawdę elementarne nie jest. Doświadczenie skłania jednak do przypuszczenia, że pogląd ten może nie jest słuszny. Okazuje się mianowicie, że kwarki dotychczas nie zostały wykryte jako cząstki swobodne i znane są jedynie jako składniki hadronów. Być może kwarki swobodne zostaną w przyszłości wykryte, ale wielu fizyków sądzi, że nie nastąpi to nigdy, gdyż kwarki — według nich — mogą istnieć jedynie wewnątrz hadronu, a więc jako cząstki związane. Ta hipoteza kwarków uwięzionych rozstrzygałaby więc problem elementarności w sposób nieoczekiwany: cząstki elementarne byłyby wprawdzie złożone, ale nie mogłyby być nigdy rozbite na swoje składniki. Poznawanie coraz głębszych poziomów struktury materii zostałoby wtedy w zasadniczy sposób zmodyfikowane. Można by nadal poszukiwać składników cząstek takich jak kwarki, ale sens tego pojęcia w istotny sposób różniłby się od pojęcia składnika np. atomu.

hipoteza uwięzionych kwarków

Tabela cząstek elementarnych

Nazwa	Symbol	J^P	I^G	P_C	S, C	Masa MeV	Czas życia, s lub Γ, MeV	Główne kanały rozpadu	Względne prawdopodobieństwo
FOTON	γ	1^-		-1	0,0	0 $< 6\cdot10^{-22}$	trwały		
LEPTONY									
Neutrino elektronowe	ν_e	1/2				0 $< 6\cdot10^{-5}$	trwałe		
Neutrino mionowe	ν_μ	1/2				0 $< 0,57$	trwałe		
Neutrino taonowe	ν_τ	1/2				< 250	trwałe		
Elektron	e	1/2				0,511	trwały		
Mion	μ	1/2				105,66	$2,197\cdot10^{-6}$	$e\nu_\mu\bar{\nu}_e$	100%
Taon	τ	1/2				1784	$< 2,3\cdot10^{-12}$		
HADRONY									
MEZONY									
Pion	$\pi^\pm$	0^-	1^-		0,0	139,57	$2,603\cdot10^{-8}$	$\mu\bar{\nu}_\mu$	100%
								$e\bar{\nu}_e$	$1,27\cdot10^{-4}$
	π^0	0^-	1^-	$+1$	0,0	134,96	$8,3\cdot10^{-17}$	$\gamma\gamma$	98,85%
Kaon	K	0^-	1/2		$\pm1,0$	493,7	$1,24\cdot10^{-8}$	$\mu\nu_\mu$	63,5%
								$\pi^0\pi^\pm$	21,2%
								$\pi^\pm\pi^-\pi^+$	5,6%
								$\pi^\pm\pi^0\pi^0$	1,7%
								$\mu^\pm\pi^0\nu_\mu$	3,2%
								$e^\pm\pi^0\nu_e$	4,8%
	K_S^0	0^-	1/2			497,7	$8,92\cdot10^{-11}$	$\pi^+\pi^-$	68,67%
								$\pi^0\pi^0$	31,4%
	K_L^0	0^-	1/2			497,7	$5,18\cdot10^{-8}$	$\pi^0\pi^0\pi^0$	21,5%
								$\pi^+\pi^-\pi^0$	12,4%
								$\pi^\pm\mu^\mp\nu_\mu$	27,0%
								$\pi^\pm e^\mp\nu_e$	38,8%
								$\pi^+\pi^-$	0,203%
								$\pi^0\pi^0$	0,094%
	η	0^-	0^+	$+1$	0,0	548,8	0,85 keV	$\gamma\gamma$	38%
								$3\pi^0$	29,9%
								$\pi^+\pi^-\pi^0$	23,6%
								$\pi^+\pi^-\gamma$	4,9%
								$\mu^+\mu^-$	$2,2\cdot10^{-5}$
	ρ	1^-	1^+	-1	0,0	776	158 MeV	$\pi\pi$	~100%
								$\pi\gamma$	0,024%
								e^+e^-	0,0043%
								$\mu^+\mu^-$	0,0067%
	ω	1^-	0^-	-1	0,0	782,4	10,1 MeV	$\pi^+\pi^-\pi^0$	89,8%
								$\pi^+\pi^-$	1,4%
								$\pi^0\gamma$	8,8%
								e^+e^-	0,0076%
	η'	0^-	0^+	$+1$	0,0	957,6	0,92 MeV	$\eta\pi\pi$	65,6%
								$\rho^0\gamma$	29,8%
								$\omega\gamma$	2,7%
								$\gamma\gamma$	1,9%
	φ	1^-	0^-	-1	0,0	1019,6	4,1 MeV	K^+K^-	48,6%
								K_LK_S	35,2%
								$\pi^+\pi^-\pi^0$	14,7%
								$\eta\gamma$	1,5%
								e^+e^-	0,032%
								$\mu^+\mu^-$	0,025%
	A_1	1^+	1^-	$+1$	0,0	~1100	~300 MeV	$\rho\pi$	interpretacja niepewna
	B	1^+	1^+	-1	0,0	1231	129 MeV	$\omega\pi$	jedyny zaobserwowany
	f	2^+	0^+	$+1$	0,0	1273	178 MeV	$\pi\pi$	83%
								$2\pi^+2\pi^-$	2,9%
								$K\bar{K}$	2,8%
	A_2	2^+	1^-	$+1$	0,0	1317	102 MeV	$\rho\pi$	70,0%
								$\eta\pi$	14,6%
								$\omega\pi\pi$	10,6%
								$K\bar{K}$	4,8%
	f'	2^+	0^+	$+1$	0,0	1516	67 MeV	$K\bar{K}$	główny kanał
	ρ'	1^-	1^+	-1	0,0	~1600	~300 MeV	4π	główny kanał
	A_3	2^-	1^-	$+1$	0,0	~1660	~200 MeV	$f\pi$	interpretacja niepewna
	g	3^-	1^+	-1	0,0	1690	180 MeV	2π	24%
								4π	72,1%
	h	4^+	0^+	$+1$	0,0	~2040	~150 MeV	2π, $K\bar{K}$	jedyne zaobserwowane
	K^*	1^-	1/2		$\pm1,0$	891,8	50,3 MeV	$K\pi$	~100%
								$K\gamma$	0,15%
						$m(K^{*0}) =$ $= 898,5$			
	K^*	2^+	1/2		$\pm1,0$	1434	100 MeV	$K\pi$	49,1%
								$K^*\pi$	27,0%
								$K\rho$	6,6%
								$K\omega$	3,7%
	K^*	3^-	1/2		$+1,0$	1785	126 MeV		
	J/ψ	1^-	0^-	-1	0,0	3097	0,063 MeV	e^+e^-, $\mu^+\mu^-$	7%
								hadrony	86%
	ψ'	1^-	0^-	-1	0,0	3685	0,215 MeV	e^+e^-, $\mu^+\mu^-$	0,9%
								hadrony, np.:	98,1%
								$\psi\pi^+\pi^-$	33%
								$\psi\pi^0\pi^0$	17%
								$\psi\eta$	4,2%
	ψ''	1^-	?	-1	0,0	4414	43 MeV	e^+e^-	0,0013%
								hadrony	głównie

Nazwa	Symbol	J^P	I^G	Pc	S, C	Masa MeV	Czas życia s lub Γ, MeV	Główne kanały rozpadu	Względne prawdopodobieństwo
Ypsilon	Υ	1^-	0^-	-1	0,0	9458	0,060	e^+e^-, $\mu^+\mu^-$ hadrony	
	Υ'	1^-	0^-	-1	0,0	10016	< 12		
	$D^\pm$	0^-	1/2		0,1	1868,3±0,9	$2{,}5^{+3,5}_{-1,5}\cdot 10^{-13}$ s	$D^+ \to$ $\to K^-\pi^+\pi^+$ $\to K^0\pi^+$	3,9% 1,5%
	D^0	0^-	1/2		0,1	1863,1±0,9	$3{,}5^{+3,5}_{-1,7}\cdot 10^{-13}$ s	$D^0 \to K^-\pi^+ \to$ $\to K^-\pi^+\pi^0$ $\to K^-\pi^+\pi^+\pi^-$ $\to \eta\pi^+$ (widziane)	1,8% 12% 3,5%
	D^{*+}	1^-	1/2		0,1	2008,6	< 2	$D^0\pi^+$ $D^+\pi^0$	64% 28%
	D^{*0}	1^-	1/2		0,1	2006,0	< 5	$D^0\pi^0$ $D^0\gamma$	55% 45%
	F^+	0^-	0		$-1{,}1$	2030±60	$2{,}24^{+2,78}_{-1,05}\cdot 10^{-13}$ s		
BARIONY Proton	p	1/2+	1/2		0,0	938,28	trwały ($\tau > 2\cdot 10^{30}$ lat)		
Neutron	n	1/2+	1/2		0,0	939,57	917	$pe^-\bar{\nu}$	100%
	Λ	1/2+	0		$-1{,}0$	1115,6	$2{,}63\cdot 10^{-10}$	$p\pi^-$ $n\pi^0$ $pe^-\nu$ $p\mu^-\nu$ $p\pi^-\gamma$	64,2% 35,8% $8{,}07\cdot 10^{-4}$ $1{,}57\cdot 10^{-4}$ $0{,}85\cdot 10^{-3}$
	Σ^+	1/2+	1		$-1{,}0$	1189,4	$0{,}80\cdot 10^{-10}$	$p\pi^0$ $n\pi^+$ $p\gamma$ $n\pi^+\gamma$ $\Lambda e^+\nu$	51,6% 48,4% $1{,}24\cdot 10^{-3}$ $0{,}93\cdot 10^{-3}$ $2{,}02\cdot 10^{-5}$
	Σ^0	1/2+	1		$-1{,}0$	1192,5	$0{,}58\cdot 10^{-19}$	$\Lambda\gamma$ Λe^+e^-	~100% $5{,}45\cdot 10^{-3}$
	Σ^-	1/2+	1		$-1{,}0$	1197,3	$1{,}48\cdot 10^{-10}$	$n\pi^-$ $ne^-\nu$ $n\mu^-\nu$ $\Lambda e^-\nu$ $n\pi^-\gamma$	~100% $1{,}1\cdot 10^{-3}$ $0{,}45\cdot 10^{-3}$ $0{,}6\cdot 10^{-4}$ $4{,}6\cdot 10^{-4}$
	Ξ^0	1/2+	1/2		$-2{,}0$	1314,9	$2{,}90\cdot 10^{-10}$	$\Lambda\pi^0$	100%
	Ξ^-	1/2+	1/2		$-2{,}0$	1321,3	$1{,}64\cdot 10^{-10}$	$\Lambda\pi^-$ $\Lambda e^-\nu$ $\Lambda\mu^-\nu$	100% $2{,}8\cdot 10^{-4}$ $3{,}1\cdot 10^{-4}$
	Δ	3/2+	3/2		0,0	1232	1,10 MeV	$N\pi$ $p\gamma$	99,4% 0,6%
	$\Sigma(1385)$	3/2+	1		$-1{,}0$	+:1383 −:1386	35,0 MeV 42 MeV }		88% 12%
	$\Xi(1530)$	3/2+	1/2		$-2{,}0$	0:1532 −:1535	9,1 MeV 10,1 MeV		100%
	Ω^-	3/2+	0		$-3{,}0$	1672	$0{,}82\cdot 10^{-10}$	$\Xi\pi^-$, $\Xi^-\pi^0$, ΛK^-	31,4% 68,6%
	$N(1470)$	1/2+	1/2		0,0	pomiędzy 1390 a 1470	120–350 MeV	$N\pi$ $N\eta$ $N\pi\pi$	~60% ~18% ~25%
	$N(1520)$	3/2−	1/2		0,0	pomiędzy 1510 a 1530	110–150 MeV	$N\pi$ $N\pi\pi$	~55% ~45%
	$N(1688)$	5/2+	1/2		0,0	pomiędzy 1670 a 1690	120–145 MeV	$N\pi$ $N\pi\pi$	60% 40%
	$\Delta(1950)$	7/2+	3/2		0,0	pomiędzy 1910 a 1940	200–240 MeV	$N\pi$ $N\pi\pi$	40% ponad 25%
	$\Lambda(1405)$	1/2−	0		$-1{,}0$	1405	40 MeV	$\Sigma\pi$	100%
	$\Lambda(1520)$	3/1−	0		$-1{,}0$	1519	15 MeV	$N\bar{K}$ $\Sigma\pi$ $\Lambda\pi\pi$	46% 42% 10%
	$\Lambda(1815)$	5/2+	0		$-1{,}0$	1820	70–100 MeV	$N\bar{K}$ $\Sigma\pi$ $\Sigma(1385)\pi$	60% 12% 15–20%
	$\Sigma(1670)$	3/2−	1		$-1{,}0$	1670	35–70 MeV	$N\bar{K}$ $\Sigma\pi$	10–25% 20–60%
	$\Sigma(1915)$	5/2+	1		$-1{,}0$	od 1905 do 1930	70–140 MeV	$N\bar{K}$ $\Lambda\pi$	5–15% 20%
	$\Sigma(2030)$	7/2+	1		$-1{,}0$	od 2020 do 2040	120–200 MeV	$N\bar{K}$ $\Lambda\pi$ $\Sigma\pi$	~20% ~20% 5–10%
	Λ_c	1/2	0		0, +1	2273	$1{,}14^{+0,90}_{-0,44}\cdot 10^{-13}$ s		

Oddziaływania cząstek elementarnych

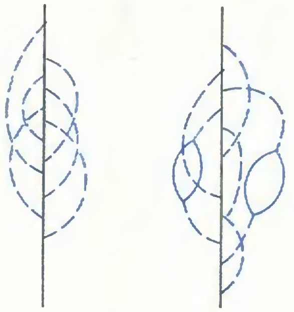
Jedną z podstawowych cech cząstek elementarnych jest ich zdolność do oddziaływania wzajemnego, która umożliwia m.in. ich obserwację i badanie. Zdolność tę mają nie tylko te cząstki, które rzeczywiście w danym miejscu i w danej chwili oddziałują z innymi, ale i te, które znajdują się tak daleko od innych cząstek, że możemy je uważać za swobodne. W tym drugim wypadku mamy do czynienia z tzw. samooddziaływaniem. Polega ono na stałym emitowaniu i pochłanianiu przez daną cząstkę innych cząstek, tych mianowicie, z którymi dana cząstka może oddziaływać. W wyniku tych nieustannych aktów emisji i absorpcji rozważana cząstka otacza się jakby chmurą cząstek, które nazywa się „wirtualnymi". Jeżeli bowiem rozważana cząstka jest trwała, to emitowane i pochłaniane przez nią cząstki nie mogą być prawdziwymi cząstkami fizycznymi, które można by wprost zaobserwować doświadczalnie. Proces rzeczywistej emisji bądź też absorpcji jest wówczas wzbroniony przez prawo zachowania czteropędu. W polu własnym danej cząstki pojawiają się więc wyłącznie obiekty przypominające cząstki rzeczywiste wszystkimi swoimi własnościami, poza masą, której wartość nie jest równa masie cząstek fizycznych i może przybierać m.in. także wartości urojone. Są to tzw. cząstki wirtualne.

proces samooddziaływania cząstek

cząstki wirtualne

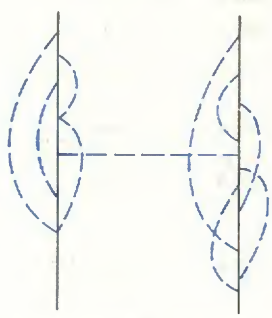
Tak więc proces samooddziaływania cząstek elementarnych nie jest dostępny bezpośredniej obserwacji i o jego istnieniu można dowiedzieć się *a posteriori* z analizy teoretycznej danych eksperymentalnych. Analizę tę przeprowadza się w ramach kwantowej teorii pola (rys. 1).

Wyobraźmy sobie dwie cząstki elementarne, które znajdują się bardzo blisko siebie. Może się wówczas zdarzyć, że cząstka wirtualna wyemitowana przez jedną z nich zostanie pochłonięta nie przez tę samą cząstkę, lecz przez inną. Dojdzie wtedy nie do samooddziaływania, lecz do oddziaływania dwu cząstek



a) b)

Rys. 1. Cząstka fermionowa (rzeczywista) otoczona chmurą wirtualnych bozonów (linie przerywane). Na rys. 1b zostały narysowane także pętle wirtualnych par fermion-antyfermion. Zakładamy, że badamy cząstkę w jej układzie spoczynkowym, a więc oś czasu jest równoległa do linii świata tej cząstki



Rys. 2. Dwie cząstki rzeczywiste fermionowe otoczone chmurami cząstek wirtualnych oddziałują przez wymianę wirtualnego bozonu

przez wymianę między nimi jednej — lub wielu — cząstek wirtualnych (rys. 2). Można więc powiedzieć obrazowo, że akt emisji cząstki wirtualnej jest demonstracją zdolności do oddziaływania cząstki emitującej. Jeżeli natomiast cząstka wirtualna zostanie reabsorbowana przez cząstkę macierzystą, to właśnie zachodzi samooddziaływanie.

oddziaływanie przez wymianę

Typy oddziaływań

W przyrodzie istnieje kilka typów oddziaływań wyraźnie różniących się od siebie. Różnice te przejawiają się w doświadczeniu w ten sposób, że procesy wywoływane przez poszczególne oddziaływania przebiegają różnie. Wnikliwsze zbadanie tych procesów prowadzi do wniosku, iż oddziaływania te różnią się od siebie zasięgiem, intensywnością oraz symetrią. Rozważania te wskazują na istnienie w przyrodzie co najmniej czterech podstawowych typów oddziaływań, a mianowicie oddziaływania grawitacyjnego, słabego, elektromagnetycznego i silnego, przy czym oddziaływania silne kwarków wewnątrz hadronu mogą się zasadniczo różnić od silnych oddziaływań hadronów. Zarazem dąży się do ujednolicenia wszystkich typów oddziaływań według pewnych określonych kryteriów. W przeszłości np. oddziaływania elektryczne i magnetyczne uważano za odrębne i dopiero dzięki badaniom prowadzonym przede wszystkim przez Faradaya i Maxwella okazało się, że w istocie istnieje tylko jedno oddziaływanie elektromagnetyczne. Obecnie prace nad podobną unifikacją oddziaływań słabych i elektromagnetycznych zostały zakończone. Na razie te dwa typy oddziaływań omawiamy oddzielnie.

Zasięg oddziaływania

Z wymienionych powyżej cech oddziaływań najłatwiejszy do zdefiniowania jest ich zasięg. Rozróżniamy oddziaływania długozasięgowe (których potencjał maleje dla dużych odległości jak $1/r$) oraz oddziaływania krótkozasięgowe (których potencjał maleje szybciej niż $1/r$). Znane z fizyki makroskopowej oddziaływania elementarne są długozasięgowe. Są to oddziaływania grawitacyjne (opisane w fizyce klasycznej prawem powszechnego ciążenia Newtona) oraz elektromagnetyczne, opisane w wypadku oddziaływania dwu spoczywających względem siebie ładunków elektrycznych prawem Coulomba. Natomiast oddziaływania krótkozasięgowe (silne i słabe) są znane wyłącznie w fizyce mikroświata, gdyż ujawniają się one dopiero na bardzo małych odległościach.

Prawo zaniku potencjału krótkozasięgowego na dużych odległościach ma postać $[\exp(-\mu r)]/r$, przy czym μ jest pewną stałą o wymiarze m^{-1}. Jako zasięg oddziaływania przyjmuje się odległość równą μ^{-1}. Oddziaływania długozasięgowe mają zatem zasięg nieskończony. Natomiast jako zasięg oddziaływań krótkozasięgowych (wartość stałej μ^{-1}) przyjmuje się długość fali comptonowskiej odpowiadającej najlżejszej cząstce zdolnej do przenoszenia danego rodzaju oddziaływania. W oddziaływaniach silnych taką najlżejszą cząstką jest mezon π. Wobec tego dla oddziaływań silnych $\mu = 0{,}7\ \mathrm{fm}^{-1}$. Oddziaływania słabe zaś mają zasięg znacznie mniejszy niż oddziaływania silne. Przez wiele lat uważano, że zasięg ten jest równy zeru. Odpowiadałoby to sytuacji, w której cząstka przenosząca oddziaływanie słabe miałaby masę nieskończenie dużą. Obecnie przypuszcza się, że zasięg oddziaływań słabych jest skończony, a bozonami przenoszącymi te oddziaływania (bozony Z i W) są cząstki o masie ok. 80 GeV/c^2, czyli cięższe ok. 600 razy od mezonu π. Tylokrotnie też krótszy byłby zasięg oddziaływań słabych od zasięgu oddziaływań silnych. Należy podkreślić, że zarówno hipoteza istnienia bozonów W i Z jak i oszacowanie wartości ich mas nie mają pełnego potwierdzenia eksperymentalnego.

Intensywność oddziaływania i stała sprzężenia

Trudniej jest wyjaśnić, co to znaczy, iż poszczególne typy oddziaływań różnią się intensywnością. Intuicyjnie jest zrozumiałe, że niektóre oddziaływania są silniejsze niż inne. Ilościowo intensywność oddziaływania charakteryzujemy wartością stałej sprzężenia.

stała sprzężenia

Fizyczny sens stałej sprzężenia jest łatwiej zrozumieć, rozważając znane z fizyki klasycznej prawa Newtona i Coulomba. We wzorach wyrażających te prawa oprócz odległości r między oddziałującymi obiektami fizycznymi występuje jeszcze — jako miara liczbowa zdolności tych obiektów do oddziaływania — wielkość taka jak masa (we wzorze Newtona) czy też ładunek (we wzorze Coulomba). Posługując się językiem fizyki cząstek elementarnych można powiedzieć, że wielkość ładunku elektrycznego lub grawitacyjnego (za który można uważać masę ciężką) charakteryzuje ilościowo zdolność danej cząstki elementarnej do emisji (lub absorpcji) cząstki przenoszącej oddziaływanie. W wypadku oddziaływania elektromagnetycznego cząstką tą jest dobrze znany foton (o masie spoczynkowej równej zeru, stąd stała μ dla tego oddziaływania ma także wartość równą zeru!), natomiast dla oddziaływań grawitacyjnych analogiczną funkcję pełniłby hipotetyczny na razie grawiton, o masie spoczynkowej równej także zeru.

Ponieważ oddziaływanie wzajemne polega na emisji oraz absorpcji cząstki przenoszącej oddziaływanie, więc liczbowo będzie ono charakteryzowane przez wielkość zawierającą kwadrat ładunku. Wielkość tę dobiera się tak, aby była bezwymiarowa. W rezultacie, jako stałą sprzężenia oddziaływania elektromagnetycznego dostajemy wielkość

$$\alpha_{el} = (4\pi\varepsilon_0)^{-1}e^2/(\hbar c), \qquad (1)$$

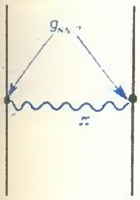
a oddziaływania grawitacyjnego wielkość

$$\alpha_{gr} = Gm_1m_2/(\hbar c), \qquad (2)$$

gdzie: e — ładunek elementarny, G — stała grawita-

cji, m_1 i m_2 — masy oddziałujących cząstek, ε_0 — przenikalność elektryczna próżni.

Jak widać ze wzorów (1) i (2) stałą sprzężenia oddziaływania elektromagnetycznego udało się zdefiniować w sposób uniwersalny, natomiast stała sprzężenia oddziaływania grawitacyjnego zależy od rodzaju oddziałujących cząstek. Dla cząstek elementarnych α_{gr} jest zawsze znacznie mniejsza od α_{el}. I tak np. dla dwu protonów α_{gr} wynosi ok. $5{,}9 \cdot 10^{-39}$, natomiast (uniwersalna) wartość α_{el} jest z dobrym przybliżeniem równa 1/137. Toteż oddziaływania grawitacyjne cząstek elementarnych są zawsze znacznie słabsze od elektromagnetycznych; tym się też tłumaczy fakt, dlaczego wolno pomijać oddziaływania grawitacyjne w opisie zachowania się cząstek elementarnych (przy obecnie osiąganych energiach).



Rys. 3. Oddziaływanie nukleon-nukleon przez wymianę wirtualnego mezonu (chmury zostały pominięte)

Przy oddziaływaniach krótkozasięgowych sytuacja jest trudniejsza, ponieważ nie znamy wzoru opisującego te oddziaływania, który by stanowił analogię do wzoru Newtona czy też Coulomba. (Wzór $e^{-\mu/r}$, którym się posługiwaliśmy przy omawianiu zasięgu oddziaływania, ma znaczenie jedynie poglądowe i półjakościowe, nie nadaje się jednak do rozważań ilościowych). Aby więc zdefiniować stałą sprzężenia, skorzystamy z analogii fizycznej do poprzednich przykładów oddziaływań.

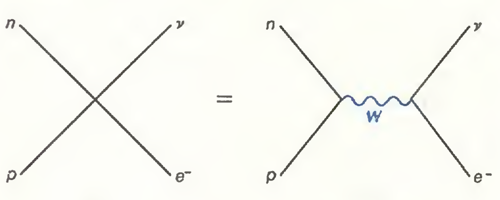
Rozpatrzmy dwa protony oddziałujące wzajemnie przez wymianę jednego (oczywiście wirtualnego) mezonu π. Aktowi emisji mezonu wirtualnego przez proton przyporządkujemy silny ładunek $g_{NN\pi}$ (rys. 3), podobnie jak aktowi emisji fotonu wirtualnego przez np. elektron został przyporządkowany ładunek elektryczny. Wielkość

$$\alpha_{sil} = (4\pi\hbar c)^{-1} g^2_{NN\pi} \qquad (3)$$

jest zwana stałą sprzężenia oddziaływań silnych nukleon-mezon π. Ta bezwymiarowa wielkość jest mierzalna (→ Oddziaływania silne) i wartość jej wynosi w przybliżeniu 15. Wobec tego, oddziaływania silne są rzeczywiście znacznie silniejsze od oddziaływań elektromagnetycznych, co uzasadnia ich nazwę.

Stosując tę terminologię należy jednak pamiętać, że oddziaływania silne są krótkozasięgowe i wobec tego dla dużych odległości są efektywnie znacznie słabsze od oddziaływań elektromagnetycznych. Dlatego też porównując intensywności różnych typów oddziaływań należy wyraźnie określić, w jakiej odległości od siebie znajdują się cząstki oddziałujące.

Sytuacja ta jest jeszcze wyraźniejsza w wypadku oddziaływań słabych. Początkowo sądzono, że oddziaływania te mają zasięg zerowy. Wtedy z fenomenologicznego punktu widzenia podstawowym aktem oddziaływania słabego nie byłaby emisja pojedynczej cząstki wirtualnej, lecz proces, w którym brałyby udział 4 cząstki fizyczne o spinach $^1/_2$ (np. w procesie rozpadu mionu — rys. 4, $\mu^- \to e^- + \bar{\nu}_e + \nu_\mu$, albo też pro-



Rys. 4. Oddziaływanie Fermiego między czterema fermionami odpowiadające wychwytowi elektronu przez proton z przejściem do stanu neutron + neutrino. Na rys. 1b ten sam proces przedstawiony jest przy założeniu istnienia hipotetycznego bozonu W przenoszącego oddziaływania słabe

ces zderzenia neutrina z elektronem, $\nu_e + e \to \nu_e + e$). Okazuje się (→ Oddziaływania słabe), że wielkość charakteryzująca taki proces nie jest bezwymiarowa i wynosi, jak wynika z danych doświadczalnych, $G = 1{,}02 \cdot 10^{-5} \cdot m_p^{-2}$ (gdzie m_p — masa protonu). Stałą G jest dość trudno porównać ze stałą sprzężenia np. oddziaływania elektromagnetycznego ze względu na zupełnie inny charakter zależności tych dwu rodzajów oddziaływania od odległości.

Jeśli, zgodnie z nowymi poglądami na oddziaływania słabe, są one przenoszone przez bozony Z i W, to można, znając stałą G, obliczyć stałą sprzężenia

$$\alpha_{st} = (4\pi\hbar c)^{-1} g^2_{pp} W. \qquad (4)$$

Wartość tej stałej jest rzędu stałej sprzężenia oddziaływań elektromagnetycznych.

Rozważania dotyczące uszeregowania oddziaływań cząstek elementarnych na podstawie wartości stałych sprzężenia można poprzeć innymi jeszcze argumentami ilościowymi.

masy cząstek elementarnych

Argumenty te czerpie się z danych podających różnice mas cząstek elementarnych oraz różnice ich czasów życia. Należy zwrócić uwagę, że energia samooddziaływania, jak każda w ogóle energia (zgodnie z zasadami szczególnej teorii względności), wiąże się z pewną masą. Ponieważ zaś cząstka jest zawsze w stanie samooddziaływania, przeto jej zmierzona masa zawiera już w sobie ów dodatek (masę polową) pochodzący z samooddziaływania. Można więc oczekiwać, że każde dwie cząstki, które się różnią samooddziaływaniem będą mieć różne wartości masy. Porównując zatem masy tych cząstek, co do których można żywić nadzieję, że ich masy niepolowe są dokładnie jednakowe, możemy uzyskać dane o sile oddziaływania, powodującego przyrost masy. Intensywność oddziaływania zależy zaś od wartości stałej sprzężenia. Tym samym różnice mas cząstek elementarnych są jakąś miarą wartości stałych sprzężenia poszczególnych rodzajów oddziaływań.

Mimo, że problem ten przedstawia się dość jasno od strony jakościowej, to jednak dotychczas nie opracowano konsekwentnej metody obliczania poprawek do masy pochodzących od samooddziaływania. Jest to więc zagadnienie, którego nie można jeszcze przedstawić od strony ilościowej.

różnice mas cząstek pokrewnych

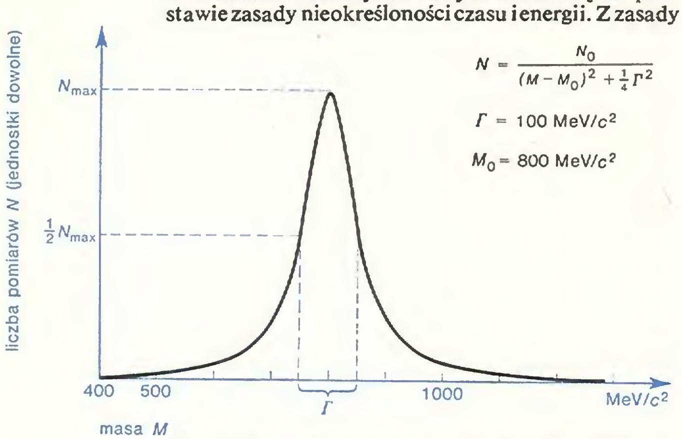
Dotychczas jest znany jeden przykład dwu cząstek elementarnych, różniących się jedynie swymi oddziaływaniami słabymi — są to K^0_L i K^0_S (patrz tabela na str. 84). Różnica ich mas jest bardzo mała, wynosi zaledwie $3{,}5 \cdot 10^{-12}$ MeV/c^2. Znacznie większe wartości mają różnice mas między pokrewnymi cząstkami wynikające z oddziaływań elektromagnetycznych. I tak np. neutron jest cięższy od protonu o 1,29 MeV/c^2, mezon π^+ (i π^-) od mezonu π^0 o 4,6 MeV/c^2, hiperon Σ^- od hiperonu Σ^0 o 4,8 MeV/c^2, ten ostatni zaś od hiperonu Σ^+ o 3,1 MeV/c^2 itd. Patrząc na tabelę, można się przekonać, że różnice te są z reguły mniejsze niż 10 MeV/c^2.

Jeszcze większe różnice mas występują między cząstkami pokrewnymi, różniącymi się jednak oddziaływaniem silnym. I tak np. różnica masy między hiperonem Λ a nukleonem wynosi ok. 175 MeV/c^2, między hiperonami Σ i Λ — ok. 75 MeV/c^2, między hiperonami Ξ i Σ — ok. 130 MeV/c^2, między mezonem K a mezonem π ok. 350 MeV/c^2 (wartość wyjątkowo duża!), między mezonem K^* a mezonem ϱ ok. 130 MeV/c^2 itd. Jak widać z tego zestawienia, różnice mas spowodowane oddziaływaniami silnymi zawierają się orientacyjnie między 100 a 400 MeV/c^2. Są to więc wartości mniej więcej 100 razy większe od różnic mas elektromagnetycznych, co się z grubsza zgadza ze stosunkiem odpowiednich ładunków: silnego i elektrycznego. Wszelkie dokładniejsze porównania ilościowe mają jednak niewielkie znaczenie, ponieważ nie ma konsekwentnej teorii mas polowych.

średnie czasy życia

Następnym kryterium służącym do rozróżniania rozmaitych rodzajów oddziaływań mogą być średnie czasy życia cząstek nietrwałych. Cząstki nietrwałe — z punktu widzenia struktury pola własnego — tym się różnią od cząstek trwałych, że mogą emitować także cząstki rzeczywiste (fizyczne), a nie tylko wirtualne; taki akt emisji nie zachodzi bowiem z pogwałceniem zasady zachowania czteropędu. Jednakże emisja cząstki czy to wirtualnej czy też rzeczywistej jest

uzależniona od zdolności cząstki emitującej do oddziaływania i wobec tego zależy od wartości stałej sprzężenia. Należy więc oczekiwać, że cząstki rozpadające się w wyniku oddziaływania silnego mają średnie czasy życia krótsze niż cząstki rozpadające się elektromagnetycznie, a te z kolei żyją średnio krócej niż cząstki rozpadające się w wyniku oddziaływań słabych. Jak wynika z tabeli na str. 84 dane doświadczalne potwierdzają to przypuszczenie. Istotnie, wszystkie średnie czasy życia cząstek nietrwałych można podzielić na trzy grupy. Warto tu dodać, że średnie czasy życia cząstek rozpadających się w rezultacie oddziaływań silnych ocenia się na podstawie zasady nieokreśloności czasu i energii. Z zasady



Rys. 5. Charakterystyczny przebieg krzywej rozkładu masy cząstki nietrwałej, z którego można odczytać wartość szerokości połówkowej. Na osi poziomej odłożona jest masa, a na pionowej liczba pomiarów, w których została zmierzona dana wartość masy

tej wynika m.in., że średni czas życia cząstek nietrwałych τ, pomnożony przez szerokość połówkową Γ, charakteryzującą niepewność w określeniu masy cząstki nietrwałej, wynosi $\hbar$. Na tej podstawie można zapisać zależność $\tau(s)\Gamma(\text{MeV}) \approx 6{,}58 \cdot 10^{-22}$ (rys. 5).

zależność czasu życia od typu oddziaływania

Jedną grupę stanowią średnie czasy życia rzędu 10^{-21}–10^{-24} s. Zaliczyć tu można czasy życia takich cząstek jak Δ ($\tau = 4{,}3 \cdot 10^{-24}$ s), ω ($\tau = 6{,}6 \cdot 10^{-23}$ s), ϱ ($\tau = 5{,}7 \cdot 10^{-24}$ s) itd. Najdłuższe czasy życia w tej grupie mają mezony φ ($1{,}6 \cdot 10^{-22}$ s) oraz niedawno odkryte mezony J/ψ ($9{,}8 \cdot 10^{-21}$ s) i ψ' ($2{,}9 \cdot 10^{-21}$ s). Wszystkie te rozpady przypisuje się oddziaływaniom silnym.

Drugą (nieliczną) grupę średnich czasów życia tworzą τ rzędu 10^{-16}–10^{-19} s. Zaliczyć tu można średnie czasy życia mezonu π^0 ($0{,}83 \cdot 10^{-16}$), mezonu η ($0{,}77 \cdot 10^{-18}$) czy wreszcie hiperonu Σ^0 ($0{,}58 \cdot 10^{-19}$ s). Wszystkie te rozpady przypisuje się oddziaływaniom elektromagnetycznym.

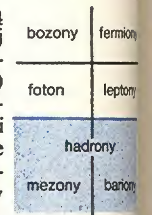
Trzecią natomiast grupę tworzą średnie czasy życia rzędu 10^{-6}–10^{-13} s. Zaliczyć tu można czasy życia mezonów $\pi^{\pm}$, $K^{\pm}$, K^0_L, K^0_S, hiperonów Λ, Σ^+, Σ^-, Ξ^0, Ξ^-, Ω. Zestaw danych liczbowych znajduje się w tabeli na str. 84. Można tu też zaliczyć średni czas życia mionu. Pozornym wyjątkiem jest rozpad neutronu, lecz wyjątkową powolność tego rozpadu wyjaśnia się działaniem czynników ubocznych. Wszystkie te rozpady powstają wskutek oddziaływań słabych.

Tak więc między czasami życia charakterystycznymi dla oddziaływań silnych i elektromagnetycznych istnieje różnica co najmniej dwu rzędów wielkości, szczególnie, jeśli pominąć wyjątkowo powolne — jak na oddziaływania silne — rozpady cząstek J/ψ i ψ'. Podobna, a nawet większa różnica (co najmniej pięć rzędów wielkości) dzieli czasy życia wywołane rozpadami elektromagnetycznymi i słabymi.

Podsumowując można powiedzieć, że dane doświadczalne przemawiają za istnieniem trzech dobrze wyodrębnionych fenomenologicznie rodzajów oddziaływań (nie biorąc pod uwagę oddziaływań grawitacyjnych).

Najtrudniejsza do wyjaśnienia jest różnica w symetrii poszczególnych rodzajów oddziaływań. Do tego zagadnienia przejdziemy dopiero po omówieniu głównych rodzajów cząstek i charakteryzujących je liczb kwantowych.

Klasyfikacja cząstek elementarnych



Podstawą klasyfikacji cząstek elementarnych są dwa niezależne kryteria — wartość liczby spinowej i zdolność do oddziaływań silnych. Wszystkie cząstki dzielimy na fermiony (ich spin wynosi $^1/_2$, $^3/_2$, ...) oraz bozony (mające spin całkowity — 0, 1, 2, ...). Jednocześnie wszystkie cząstki dzieli się na hadrony, czyli cząstki zdolne do oddziaływań silnych, oraz pozostałe cząstki, które nie mają odrębnej nazwy. Hadrony będące fermionami są nazywane barionami, a hadrony będące bozonami — mezonami. Natomiast fermiony, które nie są hadronami tworzą grupę leptonów, a odpowiednie bozony — grupę zawierającą jedną tylko znaną cząstkę elementarną, a mianowicie foton. Do tej samej grupy można by zaliczyć także bozony oddziaływań słabych, a wśród nich bozon W, gdyby ich istnienie zostało potwierdzone doświadczalnie. Liczba hadronów znacznie przewyższa liczbę cząstek pozostałych i główny problem klasyfikacyjny wiąże się właśnie z uporządkowaniem hadronów.

powszechne liczby kwantowe

Istnieje kilka wielkości fizycznych, które charakteryzują wszystkie bez wyjątku cząstki elementarne. Są to wielkości znane już z fizyki klasycznej, jak pęd, energia, masa, czas życia (niekiedy wyrażany przez swoją odwrotność, a mianowicie przez szerokość połówkową Γ), własny moment pędu (spin) czy ładunek elektryczny. Jednakże wartość pędu i energii nie odgrywa żadnej roli w klasyfikacji cząstek, gdyż może ona ulec zmianie w wyniku przyspieszenia lub spowolnienia ruchu tej samej cząstki.

Oprócz ładunku elektrycznego są jeszcze znane z fizyki klasycznej inne wielkości fizyczne charakteryzujące oddziaływanie obiektów fizycznych z polem elektromagnetycznym, jak np. momenty dipolowe (elektryczne i magnetyczne), kwadrupolowe, oktupolowe itd. W odniesieniu do cząstek elementarnych teoria pozwala na sformułowanie kilku twierdzeń ogólnych. Weźmy pod uwagę cząstkę elementarną o spinowej liczbie kwantowej J. Zgodnie z ogólnymi twierdzeniami teorii cząstek elementarnych oddziaływanie takiej cząstki z polem elektromagnetycznym scharakteryzowane jest $2J+1$ wielkościami. Twierdzenie to jest słuszne z taką dokładnością, z jaką można pominąć oddziaływania słabe, a w pewnych wypadkach nawet z lepszą (z dokładnością, z jaką zachowana jest parzystość kombinowana CP). Tak więc oddziaływania elektromagnetyczne cząstki o spinie $J = 0$ są opisywane tylko przez jedną wielkość, a mianowicie ładunek elektryczny; dla cząstek o $J = {}^1/_2$ może się pojawić druga wielkość — dipolowy moment magnetyczny; dla cząstek o $J = 1$ oprócz dwu wymienionych może wystąpić jeszcze niezerowa wartość kwadrupolowego momentu elektrycznego itd. Oczywiście nie wszystkie te wielkości, poczynając od ładunku, muszą być różne od zera. Ponieważ trudno jest zmierzyć te wielkości dla cząstek bardzo nietrwałych (wyjątkiem jest ładunek), przeto o elektromagnetycznych liczbach kwantowych cząstek elementarnych wiadomo na razie jeszcze niewiele.

Liczby kwantowe: barionowa B, leptonowa L, taonowa L_τ, mionowa L_μ i elektronowa L_e są także powszechnymi liczbami kwantowymi, które można przyporządkować wszystkim cząstkom. Przyporządkowania tego dokonuje się w ten sposób, że wszystkim barionom przypisuje się $B = 1$, wszystkim antybarionom $B = -1$ (szerzej o pojęciu antycząstki będzie mowa dalej), natomiast wszystkim nie-barionom $B = 0$. Podobnie są definiowane pozostałe liczby, np.

mionową: przypisujemy $L_\mu = 1$ mionowi μ^- oraz neutrinie mionowemu ν_μ, a ich antycząstkom — $L_\mu = -1$. Natomiast wszystkie inne cząstki mają z definicji $L_\mu = 0$.

addytywne i multiplikatywne liczby kwantowe

Wymienione dotychczas liczby kwantowe, takie jak np. pęd, moment pędu, ładunek, czy wreszcie liczby barionowa i leptonowa mają tę własność, że aby obliczyć wartość danej liczby kwantowej dla układu dwu lub więcej cząstek elementarnych należy po prostu dodać liczby kwantowe charakteryzujące poszczególne cząstki (przy momencie pędu należy oczywiście uwzględnić oprócz spinowego także orbitalny moment pędu i zastosować kwantowe prawo dodawania tej wielkości fizycznej). Takie liczby kwantowe są nazywane addytywnymi. Oprócz nich mamy też do czynienia z multiplikatywnymi liczbami kwantowymi, które należy mnożyć przez siebie, aby uzyskać wartość danej liczby kwantowej dla układu cząstek.

parzystość P

Przykładem multiplikatywnej liczby kwantowej jest parzystość P. Określa ona zachowanie się funkcji falowych opisujących układy kwantowe przy dokonaniu inwersji, tzn. zmiany znaku wszystkich trzech przestrzennych osi układu odniesienia. Jest oczywiste, że dwukrotne wykonanie inwersji przywraca stan pierwotny, można więc oczekiwać, iż parzystość P wszystkich cząstek może przybierać tylko dwie wartości, a mianowicie $+1$ i -1. Okazuje się jednak, że funkcja falowa fermionów nie ma określonego zachowania przy inwersji, co uniemożliwia bezwzględne przyporządkowanie parzystości cząstkom o spinie połówkowym. Można jednak określić parzystość względną dwu fermionów, gdyż para fermionów ma spin całkowity i jest bozonem. Bardziej szczegółowa analiza prowadzi do wniosku, że kilku wybranym fermionom należy umownie przyporządkować parzystość, a wtedy parzystości pozostałych cząstek będą jednoznacznie określone. Wybranymi fermionami są: proton, neutron, hiperon Λ, elektron, mion i taon. Cząstkom tym przypisuje się $P = +1$. Dodajmy, że określenie parzystości dla neutrin nie ma sensu, gdyż cząstki te biorą udział tylko w oddziaływaniach słabych, które nie zachowują parzystości (zostanie to omówione nieco dalej).

Należy jeszcze pamiętać o tym, że parzystość układu cząstek elementarnych nie jest po prostu równa iloczynowi parzystości tych cząstek, lecz temuż iloczynowi pomnożonemu dodatkowo przez parzystość związaną z ruchem orbitalnym składników układu. Dla układu dwucząstkowego mamy np.

$$P_{\text{układu}} = P_1 P_2 (-1)^l, \tag{5}$$

przy czym P_1, P_2 — parzystości składników układu, l — liczba kwantowa orbitalnego momentu pędu tych składników.

transformacja odwrócenia czasu

Powstaje pytanie, czy oprócz dotychczas omawianej parzystości, którą można by nazwać przestrzenną, istnieje także parzystość czasowa, związana z zachowaniem się funkcji falowej przy odwróceniu biegu czasu. Pozornie nie ma ona sensu, gdyż przecież rzeczywistego czasu odwrócić się nie da. Jednakże dla każdego ruchu można znaleźć ruch odwrócony w czasie, to znaczy taki, który zachodzi w ten sam sposób, w jaki zachodziłby ruch wyjściowy oglądany przez obserwatora poruszającego się wstecz w czasie.

Aby to lepiej zrozumieć, rozważmy ciało poruszające się w danym polu sił, zależnym tylko od odległości, z pewną — w ustalonej chwili i w ustalonym punkcie — prędkością. Nie zmieniając położenia ciała oraz sił działających na to ciało zmieńmy po prostu zwrot jego prędkości. Nastąpi wówczas ruch tego ciała w danym polu sił, lecz odwrócony w czasie. Gdyby zaś siły działające na to ciało były zależne od czasu, należałoby rozpatrywać ruch ciała w polu sił $\vec{F}(-t)$, przy jednoczesnym odwróceniu zwrotu jego prędkości. Widać więc, że transformacja odwrócenia czasu ma sens fizyczny.

W mechanice klasycznej z transformacją odwrócenia czasu, zresztą podobnie jak z operacją inwersji, nie wiąże się żadna stała ruchu. Jednakże w mechanice kwantowej sytuacja ta — teoretycznie rzecz biorąc — mogłaby się zmienić, podobnie jak dla inwersji, której wartość własna, a mianowicie parzystość przestrzenna, jest pewną liczbą kwantową charakteryzującą zarówno pojedyncze cząstki elementarne jak i ich układy. Wykażemy teraz, że sytuacja taka nie zachodzi dla odwrócenia czasu, a więc że nie istnieje żadna parzystość czasowa

Podstawowe w mechanice kwantowej równanie ruchu czyli równanie Schrödingera ma postać

$$\left(-\frac{\hbar^2}{2m}\Delta + V\right)\psi = -\frac{\hbar}{i}\frac{\partial\psi}{\partial t}, \tag{6}$$

przy czym: Δ — operator Laplace'a równy $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$,

$\hbar$ — stała Plancka, m — masa rozważanego obiektu kwantowego, a ψ — funkcja falowa opisująca ten obiekt. Z równania Schrödingera wynika, że zamiana t na $-t$ prowadzi do zmiany znaku prawej strony równania, natomiast nie powoduje żadnej zmiany po stronie lewej. Dąży się jednak do tego, aby ruch kwantowy odwrócony w czasie przebiegał wedle tych samych praw co i ruch nieodwrócony. Trzeba wobec tego przyjąć, że operacja odwrócenia czasu zawiera w sobie także operator sprzężenia zespolonego, przy którym liczby rzeczywiste nie zmieniają się, zmienia się natomiast znak jednostki urojonej i. Ta drobna na pozór sprawa ma poważne konsekwencje, uniemożliwia bowiem sformułowanie w ramach mechaniki kwantowej zagadnienia własnego dla operatora odwrócenia czasu. Zgodnie bowiem z prawami mechaniki kwantowej dwie funkcje ψ i $c\psi$ (c — dowolna stała zespolona) opisują dokładnie ten sam stan fizyczny. Jeśli ψ jest funkcją własną operatora odwrócenia czasu, to, oznaczając ten operator symbolem T, mamy $T\psi = T'\psi$, gdzie T' jest odpowiednią wartością własną charakteryzującą stan kwantowy ψ. Jednakże $c\psi$ musi być wtedy także funkcją własną operatora T przynależną do tej samej wartości własnej, a więc powinna zachodzić $Tc\psi = T'c\psi$; wszelako mamy też związek $Tc\psi = c^* T\psi = c^* T'\psi$. Dwie te równości nie mogą być ze sobą zgodne dla dowolnej zespolonej wartości c. Tak więc nie ma parzystości czasowej.

dziwność S

Bardzo wiele liczb kwantowych charakteryzuje wyłącznie (lub prawie wyłącznie) hadrony. Na pierwszym miejscu zostanie tu wymieniona dziwność S. Jest to addytywna liczba kwantowa przybierająca wartość 0 dla fotonu oraz dla hadronów niedziwnych. Do hadronów niedziwnych zalicza się proton i neutron (wraz z ich antycząstkami) oraz wszystkie cząstki, które mogą powstawać pojedynczo w reakcjach zachodzących między wymienionymi cząstkami wywołanych oddziaływaniami silnymi. Hiperonowi Λ jest przypisywana dziwność -1. Łatwo wobec tego określić dziwność wszystkich innych hadronów, przestrzegając reguły addytywności oraz warunku, aby cząstka, której dziwność chcemy określić, powstawała pojedynczo w reakcji wywołanej oddziaływaniami silnymi. Obecnie są znane bariony o dziwności $S = 0, -1, -2, -3$ oraz mezony o dziwności $S = -1, 0, +1$.

izospin I

Z dziwnością wiąże się ściśle hiperładunek $Y = B + S$. Inną liczbą kwantową charakteryzującą wyłącznie bariony jest izospin I. Podstawą fenomenologiczną do wprowadzenia tej liczby kwantowej była obserwacja grupowania się hadronów w pewne rodziny o następujących własnościach: a) wszystkie cząstki należące do tej samej rodziny mają w przybliżeniu takie same masy (ich różnice są rzędu kilku MeV/c^2, b) wszystkie cząstki należące do tej samej rodziny mają te same wartości spinu, parzystości, liczby barionowej i dziwności, c) ładunki elektryczne tych cząstek przybierają (w jednostkach e) wszystkie wartości z przedziału $-I + Y/2$ aż do $+I + Y/2$, gdzie I — pewna liczba całkowita lub połówkowa, nieujemna. Nazywa się ją izospinem całej rodziny cząstek, a tę rodzinę z kolei — multipletem izospinowym. Liczebność takiej rodziny wynosi oczywiście $N_I = 2I+1$.

Analogia formalna izospinu do zwykłego spinu, częściowo tłumaczy wybór nazwy dla tej wielkości fizycznej. Mianowicie (dla zwykłego spinu) danej wartości J odpowiada $2J+1$ stanów, różniących się wartością jednej ze składowych spinu, którą zwykle wybieramy jako z-ową (albo trzecią) składową J, tzn. J_3. Liczba kwantowa J_3 przybiera wszystkie bez wyjątku wartości ze zbioru $-J, -J+1, \ldots, J-1, J$. Na tej zasadzie możemy też wprowadzić trzecią składową izospinu, I_3, która także przybiera wszystkie wartości $-I, -I+1, \ldots, I-1, I$, po czym można zapisać wzór na ładunek cząstek należących do tego samego multipletu izospinowego w postaci

wzór Gell-Manna–Nishijimy

$$Q = I_3 + Y/2 \tag{7}$$

(wzór Gell-Manna–Nishijimy).

Oczywiście, rozpatrując składową izospinu mamy na myśli jakąś interpretację geometryczną izospinu jako wektora. Rozumie się jednak, że nie jest to wektor w zwykłej przestrzeni, lecz jakiejś dodatko-

wej przestrzeni ładunkowej, a lepiej — izosprzestrzeni. Tę izoprzestrzeń należy także uważać za trójwymiarową. W zwykłej przestrzeni oprócz spinowego istnieje też orbitalny moment pędu, związany z ruchem postępowym dwu obiektów względem siebie. Jednakże w przestrzeni izospinowej nie występuje ruch postępowy, nie ma więc także np. „izopędu" lub „izoenergii", a w konsekwencji nie ma też „izomomentu orbitalnego". Tak więc, chcąc znaleźć wartość izospinu układu dwu cząstek należy dodać izospiny składników układu (nie dbając o nie istniejący w tym wypadku moment orbitalny), pamiętając jedynie, że izospin należy składać zgodnie z kwantowymi regułami dodawania momentu pędu.

powab *C*

Niedawno pojawiła się w fizyce nowa liczba kwantowa, charakteryzująca wyłącznie hadrony, a mianowicie powab *C*. Wszystkie dotychczas wymienione cząstki, umieszczone w tabeli na str. 84, miałyby $C = 0$, byłyby więc cząstkami niepowabnymi. Powab jest liczbą kwantową addytywną, pod wieloma względami podobną do dziwności. Dla cząstek powabnych trzeba nieco zmodyfikować wzór Gell-Manna–Nishijimy, który przybierze wówczas postać

$$Q = I_3+(Y+C)/2. \tag{8}$$

Znamy też od niedawna, choć bardzo niedokładnie, jeszcze jedną liczbę kwantową charakteryzującą hadrony, a mianowicie B' (ang. *beauty*), piękno. Liczba ta miałaby podobny status jak powab.

antycząstki

Należy teraz dokładnie sprecyzować pojęcie antycząstki. Antycząstką danej cząstki *a* nazywamy taką cząstkę $\bar{a}$, która ma te same co *a* wartości masy, izospinu, spinu, czasu życia, przeciwne zaś B, L_τ, L_μ, L_e (a więc i L), Q, S i C. Moment magnetyczny antycząstki jest ustawiony w stosunku do jej spinu przeciwnie niż moment magnetyczny cząstki. Parzystość antyfermionu jest przeciwna parzystości fermionu, a parzystość antybozonu jest taka sama jak parzystość bozonu. To, co uznajemy za cząstkę, a co za antycząstkę, jest oczywiście umowne, gdyż $\bar{\bar{a}} = a$.

Z powyższego zestawienia liczb kwantowych antycząstek wynika, że jest możliwe istnienie takich cząstek, które są identyczne ze swymi antycząstkami, $a = \bar{a}$ gdyż mają zerowe wartości tych wszystkich liczb kwantowych, które zmieniają znak przy przejściu od cząstki do antycząstki. Cząstki te nazywać będziemy istotnie obojętnymi. Tak więc, dla cząstek istotnie obojętnych musi być $B = S = Q = C = 0$, $L_\tau = L_\mu = L_e = L = 0$. Cząstką istotnie obojętną jest np. foton oraz każdy nienaładowany, niedziwny, niepowabny mezon (np. mezon π^0). Operację polegającą na zmianie cząstki w antycząstkę nazywa się sprzężeniem cząstka-antycząstka. Z powyższych definicji wynika, że funkcja falowa opisująca stan cząstki istotnie obojętnej jest funkcją własną operatora tego sprzężenia. Ze względu na relację $\bar{\bar{a}} = a$ kwadrat operatora sprzężenia cząstka-antycząstka jest operatorem jednostkowym. Tym samym jego wartości własne muszą, podobnie jak dla parzystości przestrzennej, wynosić +1 lub −1. Wartość własną operatora sprzężenia cząstka-antycząstka P_C nazywa się parzystością ładunkową. Z powyższego wynika, że tylko cząstki istotnie obojętne mają określoną parzystość ładunkową. Parzystość ładunkowa jest multiplikatywną liczbą kwantową.

cząstki istotnie obojętne

sprzężenie cząstka-antycząstka

parzystość ładunkowa P_C

Warto zauważyć, że układ dwucząstkowy składający się z cząstki i jej antycząstki tworzy także układ istotnie obojętny, ma więc określoną parzystość ładunkową, nawet wtedy, gdy jego składniki nie są cząstkami istotnie obojętnymi. Wartość tej parzystości jest dana wyrażeniem

$$P_C = (-1)^{l+j}, \tag{9}$$

w którym l — liczba kwantowa orbitalnego momentu pędu układu, j — wartość spinu tego układu.

izoparzystość *G*

Inną multiplikatywną liczbą kwantową jest izoparzystość *G*. Jest ona wartością własną operatora izoparzystości, który jest zdefiniowany jako iloczyn operatora sprzężenia cząstka-antycząstka i operatora obrotu w izosprzestrzeni o kąt π wokół drugiej osi w tej przestrzeni. Izoparzystość ma określoną wartość dla każdej cząstk i należącej do takiego multipletu izospinu, do którego należy także cząstka istotnie obojętna. W multiplecie takim bowiem wszystkie cząstki mają $B = S = C = 0$, a zatem wzór Gell-Manna–Nishijimy przybiera szczególnie prostą postać $Q = I_3$. Sprzężenie cząstka-antycząstka zmienia Q na $-Q$, co zgodnie z powyższą relacją jest równoważne ze zmianą I_3 na $-I_3$. Jednakże obrót o kąt π wokół osi 2 w izoprzestrzeni powoduje odwrócenie znaku osi trzeciej w tejże przestrzeni, a więc automatycznie zmianę znaku I_3. Te dwie zmiany znaku kompensują się więc i łącznie sprzężenie *G* zamienia daną cząstkę w samą siebie. I tu znów $G^2 = 1$, więc izoparzystość może mieć wartości jedynie +1 i −1. Dla cząstki istotnie obojętnej można ją obliczyć ze wzoru

$$G = C(-1)^I, \tag{10}$$

zaś inne cząstki wchodzące w skład tego samego izomultipletu mają tę samą izoparzystość.

Łatwo dostrzec, że inne cząstki nie mają określonej wartości *G*. Na przykład proton pod działaniem sprzężenia cząstka-antycząstka przechodzi w antyproton, a ten pod wpływem obrotu w przestrzeni izospinu zmienia znak I_3, stając się antyneutronem. Operator *G* przeprowadza więc proton nie w proton, lecz w antyneutron. Proton nie jest więc stanem własnym *G* i nie ma określonej izoparzystości.

parzystość kombinowana

Należy przy tym dodać, że w teorii oddziaływań słabych używa się jeszcze jednej multiplikatywnej liczby kwantowej, zwanej parzystością kombinowaną, będącej wartością własną operatora zdefiniowanego jako iloczyn operatora inwersji i sprzężenia cząstka-antycząstka.

Rodzaje oddziaływań a prawa zachowania

Rozważmy dowolny proces z udziałem cząstek elementarnych. W takim procesie na ogół ulega zmianie natura a nawet liczba cząstek. Doświadczenie wykazuje, że nie wszystkie możliwe a priori procesy zachodzą w przyrodzie, a także, że niektóre z nich zachodzą powoli, a inne — szybko. Aby zdać sobie sprawę z tych regularności wprowadzamy pojęcie zachowania wielkości fizycznych charakteryzujących cząstki elementarne i ich układy.

Wykorzystując podane w poprzednim punkcie reguły, możemy obliczyć wartość określonej wielkości fizycznej dla układu cząstek wyjściowych oraz oddzielnie dla układu cząstek końcowych. Mówimy, że pewna wielkość fizyczna jest zachowana w danym procesie, jeżeli dwie tak obliczone wartości są jednakowe oraz, że wielkość ta jest zachowana przez oddziaływania silne (elektromagnetyczne, słabe), jeżeli jest ona zachowana we wszystkich procesach przebiegających w czasie charakterystycznym dla oddziaływań silnych (elektromagnetycznych, słabych), tj. w czasie 10^{-21}–10^{-24} s (lub też 10^{-16}–10^{-19} s, lub wreszcie 10^{-6}–10^{-10} s). Wielkości zachowywane we wszystkich w ogóle oddziaływaniach nazywa się wielkościami bezwzględnie zachowanymi.

wielkości zachowywane bezwzględnie

Z doświadczenia wynika niesłychanie interesująca zależność. Okazuje się, że jeżeli jakaś wielkość fizyczna jest zachowana przez oddziaływania słabe, to na pewno jest też zachowana przez oddziaływania elektromagnetyczne, oraz jeśli jest zachowana przez oddziaływania elektromagnetyczne, to jest też zachowana przez oddziaływania silne.

Wielkościami zachowanymi bezwzględnie są: pęd, energia i moment pędu, oraz ładunek elektryczny, liczba barionowa i liczby leptonowe. Oddziaływania słabe zachowują jeszcze dodatkowo parzystość kombinowaną. Występują jednak procesy, które zachodzą jeszcze wolniej niż procesy uwarunkowane oddziały-

waniami słabymi, w których parzystość kombinowana nie jest zachowana. Są to pewne rozpady mezonów K_L^0. Jak wynika z tabeli na str. 84 jednym z głównych kanałów rozpadu tych mezonów jest rozpad na trzy mezony π, zgodny z prawem zachowania parzystości kombinowanej. Jednakże przekonano się, że w pewnym bardzo niewielkim procencie wypadków, następuje też rozpad K_L^0 na dwa mezony π z pogwałceniem tego prawa. Fenomenologicznie odpowiada to istnieniu oddziaływań podsłabych, słabszych od słabych, nie zachowujących nawet parzystości kombinowanej. Rzeczywiste jednak wyjaśnienie tego zjawiska jest inne.

oddziaływania podsłabe

Poza wielkościami zachowywanymi przez oddziaływania słabe, oddziaływania elektromagnetyczne zachowują jeszcze parzystość i parzystość ładunkową (z osobna) oraz S i I_3 (te dwie ostatnie wielkości dla procesów, w których uczestniczą wyłącznie hadrony i foton, przy założeniu, że dla fotonu $I_3 = 0$, aby utrzymać w mocy prawo Gell-Manna–Nishijimy). Oddziaływania silne zachowują ponadto izospin i izoparzystość. Powab byłby zachowany już przez oddziaływania elektromagnetyczne.

Ta degradacja oddziaływań ze względu na liczbę zachowywanych przez nie wielkości fizycznych, pokrywająca się przy tym ze spadkiem intensywności oddziaływania, nie jest w pełni jasna. Lepsze zrozumienie znaczenia tego faktu można uzyskać po zbadaniu związku między prawami zachowania a symetrią teorii fizycznej.

Symetrie a prawa zachowania

Jedną z najgłębszych i najciekawszych idei już nie tylko fizyki cząstek elementarnych, ale fizyki w ogóle jest związek między symetrią a prawami zachowania (→ Zasady zachowania). Układy fizyczne wykazują na ogół pewne własności symetrii. Na przykład, pojedyncza cząstka elementarna swoimi własnościami symetrii nie wyróżnia w przestrzeni żadnego kierunku dzięki czemu można, nie zmieniając jej opisu, w dowolny sposób skierować osie układu odniesienia. Jest też możliwe dowolne wybranie początku układu odniesienia, gdyż każdy punkt w przestrzeni jest równoważny innemu, skoro w przestrzeni tej nie ma innych cząstek, jak również dowolne wybranie chwili początkowej jako punktu zerowego na osi czasu. Co więcej, taka pojedyncza, swobodna cząstka porusza się, zgodnie z I zasadą dynamiki, ruchem jednostajnym prostoliniowym; prędkość tego ruchu jest jednak dowolna. Ta dowolność prędkości odpowiada swobodzie przy wyborze do opisu cząstki jednego z nieskończenie wielu układów inercjalnych, poruszających się względem siebie ze stałą prędkością.

własności symetrii cząstki swobodnej

Intuicyjnie jest dość zrozumiałe, że ze swobodą przy wyborze tego czy innego układu odniesienia powinna się wiązać jakaś niezmienna własność rozważanego ciała lub układu ciał. Ścisłe uzasadnienie tego przekonania jest jednak dość trudne. Dokładne rozważania prowadzą do twierdzenia Noether, które orzeka, że każda wielkość, będąca stałą ruchu (a więc spełniająca prawo zachowania), wiąże się z jakimś stopniem swobody, który mamy do dyspozycji przy wyborze sposobu opisu zachowania się układu fizycznego. Twierdzenie odwrotne nie byłoby jednak zawsze prawdziwe.

twierdzenie Noether

W fizyce klasycznej występują prawa zachowania pędu energii, momentu pędu i ładunku elektrycznego. Twierdzenie Noether pozwala na ustalenie, że prawo zachowania pędu wynika ze swobody przy wyborze początku układu odniesienia (w przestrzeni trójwymiarowej są to trzy stopnie swobody, odpowiada więc im zachowana wielkość trójskładnikowa — wektor). Prawo zachowania energii wynika ze swobody przy wyborze chwili początkowej na osi czasu, a prawo zachowania momentu pędu — ze swobody przy wyborze kierunków osi układu odniesienia.

Nieco trudniej opisać symetrię układu fizycznego, z której wynika prawo zachowania ładunku. Przypomnijmy, że podstawowe równania teorii elektromagnetyzmu, a mianowicie równania Maxwella, dotyczą dwu wektorów — natężenia pola elektrycznego $\vec{E}$ i indukcji magnetycznej $\vec{B}$. Okazuje się, że wektory te przez pewne operacje różniczkowe można otrzymać z dwu potencjałów — skalarnego φ i wektorowego $\vec{A}$. Potencjały te jednak nie są wyznaczone jednoznacznie przez wartości pól $\vec{E}$ i $\vec{B}$; można je w pewien sposób przekształcić nie zmieniając samych pól. Każde takie przekształcenie nazywano przekształceniem cechowania. Zakładając, że potencjał skalarny można traktować jako zerową (czasową) składową czterowektora potencjału, którego składowymi przestrzennymi są składowe potencjału wektorowego, można przyjąć, że przekształcenie cechowania sprowadza się w przybliżeniu do dodania do każdej składowej czterowektora potencjału pochodnej cząstkowej pewnej (jednakowej dla wszystkich składowych) funkcji skalarnej β, wziętej względem tej zmiennej, która odpowiada danej składowej. Na przykład do A_x można dodać $\partial\beta/\partial x$, a do A_0 (czyli φ/c) — $(\partial\beta/\partial t)/c$.

przekształcenia cechowania

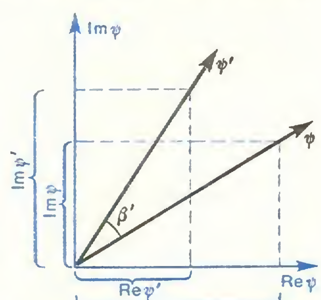
Wyrażenie na gęstość energii oddziaływania cząstki naładowanej z polem elektromagnetycznym ma postać

$$ej_\mu A^\mu = e(j_0 A_0 - \vec{j}\vec{A}), \tag{11}$$

gdzie j_μ — czterowektor gęstości prądu (jego składową zerową jest gęstość ładunku). W mechanice kwantowej analogiczna wielkość jest zbudowana z funkcji falowych cząstek naładowanych w sposób zależny od natury tych cząstek, jednakże zawsze w taki sposób, że każda składowa prądu zawiera funkcję falową cząstki i funkcję z nią sprzężoną (zespoloną). Jeśli chcemy, aby przekształcenie cechowania nie zmieniało postaci równań opisujących ruch cząstki naładowanej w polu elektromagnetycznym, to nie może się ono ograniczyć do pola elektromagnetycznego, lecz musi też objąć prąd, a więc i funkcje falowe cząstek naładowanych. Okazuje się, się przekształcenie to, wykonane na tych funkcjach sprowadza się do ich obrotu w płaszczyźnie zespolonej (Re ψ, Im ψ) o pewien kąt $\beta' = (e/c)\,\beta$:

$$(\mathrm{Re}\,\psi)' = \mathrm{Re}\,\psi\,\cos\beta' + \mathrm{Im}\,\psi\,\sin\beta',$$
$$(\mathrm{Im}\,\psi)' = -\mathrm{Re}\,\psi\sin\beta' + \mathrm{Im}\,\psi\,\cos\beta'.$$

Kąt ten (rys. 6), co widać już z powyższego wzoru, jest proporcjonalny do funkcji β, której pochodne dodaje się do składowych wektora czteropotencjału.



Rys. 6

W powyższych wzorach założono, że β jest funkcją położenia i czasu. Takie przekształcenie cechowania jest nazywane przekształceniem drugiego rodzaju. Zakładając, że β nie zależy od r i t, a więc jest pewną stałą, otrzymujemy cechowanie pierwszego rodzaju. Okazuje się, że zachowanie ładunku (jedna wielkość!) wynika ze swobody wyboru kierunków osi w płaszczyźnie (Re ψ, Im ψ), co się wyraża w dowolności kąta obrotu β'. Ładunek elektryczny pełni w fizyce

cechowanie pierwszego i drugiego rodzaju

podwójną rolę: jest stałą ruchu oraz źródłem pola elektromagnetycznego. Jak się okazuje należy używać w tym wypadku przekształcenia cechowania drugiego rodzaju.

W przyrodzie istnieją inne stałe ruchu, których formalne podobieństwo do ładunku elektrycznego wyraża się tym, że ich zachowanie jest także konsekwencją pewnej transformacji cechowania, ale które nie są źródłami żadnego pola fizycznego. Wielkościami takimi są liczba barionowa i liczba leptonowa. Tak też, w pewnym ujęciu, można traktować dziwność i powab. W tych wszystkich wypadkach mamy do czynienia z przekształceniem cechowania pierwszego rodzaju.

Prawo zachowania parzystości wiąże się ze swobodą przy wyborze lewo- i prawoskrętnego układu odniesienia, a więc z symetrią względem inwersji. Podobnie prawo zachowania parzystości ładunkowej wynika ze swobody przy wyborze tego, co uznamy za cząstkę, a co za antycząstkę, a więc z symetrii względem sprzężenia cząstka-antycząstka.

Przekształcenie odwrócenia czasu jest przykładem sytuacji, w której symetria teorii względem pewnego przekształcenia nie prowadzi do żadnego prawa zachowania. Istotnie, prawa fizyki klasycznej, czy też mechaniki kwantowej są (z pewnym wyjątkiem, o czym dalej) symetryczne ze względu na odwrócenie czasu. Nie wiąże się z tym jednak zachowanie parzystości czasowej, gdyż po prostu wielkości takiej nie można zdefiniować, jak to zostało zaznaczone powyżej.

sprzężenie *PCT*

Do podstawowych twierdzeń teorii cząstek elementarnych należy hipoteza (w pewnych konkretnych wypadkach udało się ją udowodnić), że wszystkie rodzaje oddziaływań wykazują symetrię względem przekształcenia będącego iloczynem inwersji, sprzężenia cząstka-antycząstka i odwrócenia czasu (zwanego sprzężeniem *PCT*). Mimo, że z symetrią tą, skoro zawiera ona odwrócenie czasu, nie wiąże się żadna stała ruchu, to jest ona bardzo ważnym składnikiem teorii cząstek. Wymienione wyżej twierdzenie zapewnia bowiem równość mas cząstki i antycząstki oraz równość ich czasów życia.

Warto zauważyć, że wykonanie po sobie dwu przekształceń symetrii jest znów pewnym przekształceniem symetrii. Przez przekształcenie symetrii rozumie się takie przekształcenie, które nie zmienia opisu badanego obiektu fizycznego, a więc jest zgodne z symetrią tego obiektu. Za przekształcenie symetrii należy też uważać przekształcenie tożsamościowe. Nasuwa się więc wniosek, że wszystkie przekształcenia symetrii danego układu tworzą grupę, zwaną jego grupą symetrii.

grupa symetrii

Wspomnieliśmy poprzednio, iż najwięcej praw zachowania wiąże się z oddziaływaniami silnymi, a na podstawie obecnych rozważań wnioskujemy, że układom cząstek oddziałujących silnie odpowiada najwięcej przekształceń symetrii. Można zatem powiedzieć, że oddziaływania silne są najbardziej symetryczne, mniej symetryczne są oddziaływania elektromagnetyczne, a jeszcze mniej — oddziaływania słabe. Najobszerniejsza grupa symetrii charakteryzuje oddziaływania silne, a grupa symetrii oddziaływań elektromagnetycznych jest jej podgrupą; podgrupą zaś jej z kolei jest grupa symetrii oddziaływań słabych. Wniosek ten wyciągnęliśmy już teraz, mimo, że nie jest jeszcze znana pełna grupa symetrii oddziaływań silnych, do czego przechodzimy w następnym punkcie.

Izospin i wyższe grupy symetrii

Przy wprowadzaniu pojęcia izospinu stwierdzono, że ma on własności podobne do zwykłego spinu. Spin jest momentem pędu, a zachowanie momentu pędu wiąże się z symetrią względem obrotów w trójwymiarowej przestrzeni zmiennych x, y, z. Analogicznie zachowanie izospinu wiąże się z symetrią względem obrotów w trójwymiarowej izoprzestrzeni.

Okazuje się jednak, że takie ujęcie teorii zarówno momentu pędu jak i izospinu nie jest dogodne. Niedogodność ta pojawia się z chwilą wprowadzenia cząstek o spinie połówkowym. Wydaje się oczywiste, że obroty o kąt 0 i 2π są geometrycznie równoważne. Dla cząstek o spinie całkowitym obroty te są sobie równoważne także fizycznie, gdyż funkcja falowa takich cząstek po obrocie o kąt 2π nie zmienia się. Inaczej jest dla cząstek o spinie połówkowym, gdyż obrót o kąt 2π zmienia znak funkcji falowej tych cząstek i dopiero obrót o kąt 4π jest równoważny nie tylko geometrycznie ale i fizycznie obrotowi o kąt 0. W celu uniknięcia tej dwuznaczności wprowadzono zamiast grupy obrotów w trójwymiarowej przestrzeni zmiennych rzeczywistych x, y, z, grupę przekształceń unitarnych unimodularnych w przestrzeni dwuwymiarowej zmiennych zespolonych, które można oznaczyć ξ, η. Każde przekształcenie tego typu zapisuje się w postaci

$$\xi' = A_{\xi\xi}\xi + A_{\xi\eta}\eta, \quad \eta' = A_{\eta\xi}\xi + A_{\eta\eta}\eta,$$

przy czym macierz współczynników w tym przekształceniu A,

$$A = \begin{pmatrix} A_{\xi\xi}, & A_{\xi\eta} \\ A_{\eta\xi}, & A_{\eta\eta} \end{pmatrix}$$

spełnia warunki $A^{\dagger}A = 1$ (unitarność),

gdzie $A^{\dagger} = \begin{pmatrix} A^*_{\xi\xi}, & A^*_{\eta\xi} \\ A^*_{\xi\eta}, & A^*_{\eta\eta} \end{pmatrix}$

symetria *SU*(2)

i $\mathrm{Det}(A) = 1$ (unimodularność), $\mathrm{Det}(A)$ oznacza wyznacznik macierzy A. Grupę takich przekształceń nazywamy grupą $SU(2)$.

Grupa $SU(2)$ umożliwia matematycznie poprawne sformułowanie teorii momentu pędu i izospinu. Zachowanie izospinu jest konsekwencją symetrii teorii względem przekształceń należących do izospinowej grupy $SU(2)$.

W kwantowej teorii momentu pędu wykazuje się, że przy dodawaniu dwu momentów pędu, $\vec{J_1}$ i $\vec{J_2}$, $\vec{J_1}+\vec{J_2} = \vec{J}$, liczba kwantowa J kwadratu momentu pędu J musi spełniać warunek

$$|J_1 - J_2| \leqslant J \leqslant J_1 + J_2. \tag{12}$$

Dodając zatem dwa momenty pędu $^1/_2$ dostajemy J wypadkowe równe 0 lub 1. Podobnie jest oczywiście z izospinem. Łatwo można się przekonać, że dodając do siebie co najmniej $2I$-krotnie izospin $^1/_2$ dostajemy, jako największą wartość wypadkowego izospinu, izospin I. Izospin $^1/_2$ odgrywa podstawową rolę w teorii izospinu jako elementarna cegiełka, z której można uzyskać wszystkie inne wartości I. Innymi słowy, każdy multiplet izospinu może być zbudowany z dubletów.

Omówioną przed chwilą operację składania dwu izospinów $^1/_2$ można wyrazić w następującej formie

$$2 \otimes 2 = 1 \oplus 3, \tag{13}$$

przy czym znak mnożenia i dodawania zostały wzięte w kółka dla zaznaczenia, że nie chodzi tu o mnożenie i dodawanie liczb naturalnych lecz o składanie izospinów; liczby 2, 1 i 3 odpowiadają liczebnościom izomultipletów o $I = {}^1/_2$, 0 i 1. Podobnie można napisać

$$2 \otimes 2 \otimes 2 = 2 \oplus 2 \oplus 4. \tag{14}$$

symetria *SU*(3)

W miarę wykrywania coraz to nowych cząstek elementarnych, głównie hadronów, zaczęły się ujawniać pewne dalsze podobieństwa między nimi, które doprowadziły, po wielu wcześniejszych nieudanych próbach, do sformułowania teorii symetrii $SU(3)$ oddziaływań silnych.

Podstawę stanowiło tu, podobnie jak w wypadku izospinu, grupowanie się hadronów w pewne rodziny, charakteryzujące się następującymi własnościami:

rodziny hadronów

a) do tej samej rodziny należą hadrony o tych samych wartościach J, P, B; b) w skład takiej rodziny wchodzi zawsze cały multiplet izospinowy; c) wchodzące w skład takiej rodziny izomultiplety mają na ogół różne wartości I oraz Y, jednakże zbiór tych wartości nie wykazuje żadnych luk, tj. I zmienia się co $^1/_2$, a Y co 1; d) charakterystyczne różnice mas między poszczególnymi izomultipletami są w danej rodzinie rzędu ok. 100–300 MeV/c^2; e) hadrony należące do tej samej rodziny wykazują analogiczne własności z punktu widzenia oddziaływań silnych (np. nukleon wiąże się z mezonem π w nietrwały stan barionu Δ o spinie $^3/_2$ i parzystości dodatniej, a podobne stany nietrwałe powstają też z oddziaływania z mezonem π hiperonów Λ i Ξ). Liczebność zaobserwowanych rodzin wynosi dla wszystkich hadronów 1 i 8, a dla barionów ponadto 10.

W 1961 r. Gell-Mann odkrył, że multiplety o tej liczebności pojawiają się właśnie w symetrii $SU(3)$. Jest to, jak łatwo odgadnąć przez analogię z $SU(2)$, grupa przekształceń unitarnych i unimodularnych, lecz w przestrzeni zespolonej trójwymiarowej. Ten dodatkowy wymiar przestrzeni jest potrzebny, aby włączyć hiperładunek do schematu klasyfikacyjnego.

multiplety podstawowe

W grupie symetrii $SU(3)$ multipletem podstawowym jest tryplet podobnie jak w grupie $SU(2)$ — dublet. Okazuje się jednak, że w odróżnieniu od grupy $SU(2)$ multiplet podstawowy istnieje w dwu odmianach — zwykłej, oznaczonej symbolem 3 oraz sprzężonej, oznaczonej przez $\bar{3}$. Konieczność wprowadzenia takich dwu multipletów można uzasadnić następująco. W multiplecie podstawowym symetrii $SU(3)$ powinien tkwić, jako jego część, multiplet podstawowy izospinu, a więc dublet. Zatem podstawowy tryplet $SU(3)$ musi się składać z dubletu i singletu $SU(2)$. Jednakże, należy tu jeszcze dodać hiperładunek lub dziwność, jako wielkość dodatkowo definiującą grupę $SU(3)$. Chcąc budować z multipletu podstawowego $SU(3)$ wyższe multiplety o różnych wartościach nie tylko I ale i S, trzeba składowym multipletu 3 przyporządkować jakąś niezerową wartość dziwności. Gdyby wszystkie trzy składowe 3 miały tę samą dziwność, równą np. S_1, to izospinowi $I = 0$ lub też $I = 1$ towarzyszyłaby zawsze dziwność, co najmniej równa $2S_1$, izospinowi 3/2 dziwność co najmniej $3S_1$ itd. Tak jednak nie jest. Poza tym powyższe przyporządkowanie spowodowałoby, że pojawiłyby się tylko multiplety o dodatnim S (jeśli $S_1 > 0$) lub ujemnym S (w przeciwnym razie). Tymczasem dziwności ujemne i dodatnie występują w przyrodzie symetrycznie, gdyż jeśli cząstka ma dziwność S, to jej antycząstka ma dziwność $-S$. Tak więc powinno się skonstruować multiplet 3 w taki sposób, aby wartość S dla izodubletu była inna niż dla izosingletu, oraz oprócz 3 należy wprowadzić tryplet $\bar{3}$, o przeciwnych wartościach dziwności. Łatwo też zauważyć, że wartości S dla izodubletu i izosingletu należącego do 3 powinny się różnić o 1. Ostatecznie dokonujemy przyporządkowania tak, iż izodublet w 3 ma $S = 0$, a izosinglet w 3 ma $S = -1$, natomiast w $\bar{3}$ mamy izodublet o $S = 0$ i izosinglet o $S = +1$.

W ramach symetrii $SU(3)$ można podać wzory analogiczne do (13) i (14). Mamy mianowicie:

$$3 \otimes 3 = 1 \oplus 8 \tag{15}$$

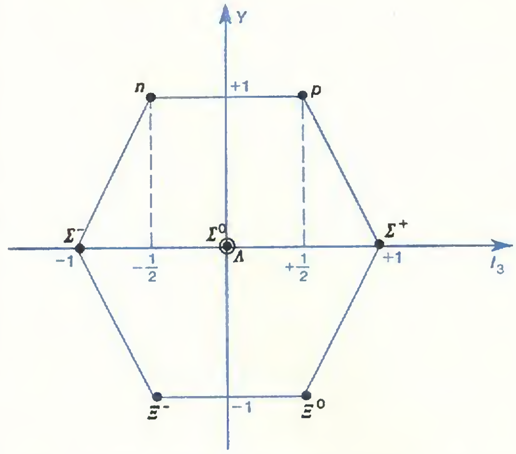
oraz

$$3 \otimes 3 \otimes 3 = 1 \oplus 8 \oplus 8 \oplus 10. \tag{16}$$

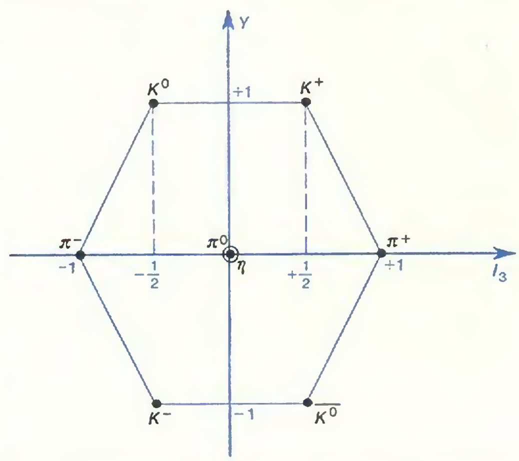
Z powyższych wzorów wynika, że istotnie, w ramach symetrii $SU(3)$ pojawiają się — obok wielu innych — także multiplety o liczebności 1, 8 i 10.

Przyporządkowanie znanych hadronów poszczególnym multipletom $SU(3)$ na ogół nie nastręcza większych trudności. I tak np. znamy 8 barionów o $J^P = \frac{1}{2}^+$, a mianowicie neutron i proton (o $S = 0$ czyli $Y = +1$), tworzące dublet izospinu, hiperon Λ i trzy hiperony Σ o $S = -1$ (więc $Y = 0$), oraz dwa hiperony Ξ o $S = -2$ (więc $Y = -1$, rys. 7). Podobnie można sklasyfikować razem — w ramach tego samego oktetu — osiem mezonów o $J^P = 0^-$ (rys. 8). Należą do niego dwa mezony K o $S = Y = +1$, trzy mezony π i jeden mezon η o $S = Y = 0$

oktet barionów i mezonów



Rys. 7. Oktet barionów $J^P = \frac{1}{2}^+$ przedstawiony na płaszczyźnie (I_3, Y)



Rys. 8. Oktet mezonów $J^P = 0^-$ przedstawiony w płaszczyźnie (I_3, Y)

oraz dwa mezony $\bar{K}$ o $S = Y = -1$. Okazuje się, że struktura wewnętrzna oktetu barionowego i mezonowego jest identyczna, jeśli do klasyfikacji użyje się nie dziwności lecz hiperładunku. Ta sama struktura cechuje też wszystkie inne oktety symetrii $SU(3)$.

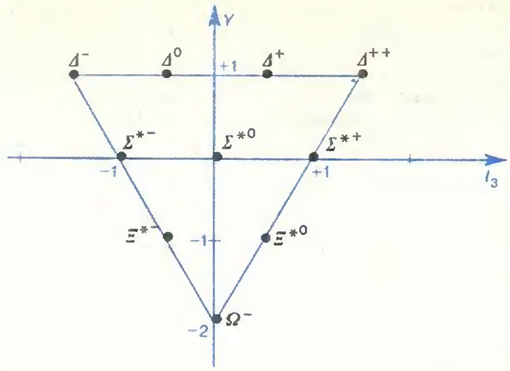
W pewnych wypadkach obserwujemy więcej cząstek niż np. osiem, o tych samych wartościach J, P, B. Powstaje wtedy pytanie, jakimi kryteriami należy się kierować, aby we wspólnym multiplecie umieścić te a nie inne cząstki. Wówczas należy uciec się do rozważań ilościowych opartych na pewnych założeniach dynamicznych.

dekuplet barionów

Jeśli chodzi o dekuplet 10, to zaliczamy tu cztery bariony $\Delta (Y = +1)$, trzy bariony Y_1^* (1385) o $Y = 0$, dwa bariony Ξ^* (1520) o $Y = -1$ i wreszcie jeden barion Ω^- o $Y = -2$ (rys. 9). Wtedy gdy sformułowano hipotezę że symetria $SU(3)$ jest symetrią świata hadronów, hiperon Ω nie był jeszcze znany z doświadczenia. Podobnie jednak, jak to było z tablicą Mendelejewa, symetria $SU(3)$ przewidywała dla tego hiperonu wolne miejsce właśnie w ramach dekupletu. Późniejsze odkrycie tej cząstki było więc poważnym sukcesem teorii, tym bardziej, że była ona w stanie przewidzieć także, i to z bardzo dobrą

przewidzenie hiperonu Ω^-

dekuplet
10



Rys. 9. Dekuplet barionów $J^P = \frac{3}{2}^+$ przedstawiony na płaszczyźnie (I_3, Y)

dokładnością masę Ω^- na podstawie wzoru masowego Gell-Manna–Okubo.

symetria
$SU(4)$

Wiele przesłanek przemawia za tym, że w rzeczywistości grupa symetrii $SU(3)$ nie jest najszerszą grupą symetrii hadronów. Ogólniejszą mogłaby być grupa $SU(4)$ lub może nawet jeszcze wyższa ($SU(6)$). Rozważając grupę $SU(4)$, musimy znów wprowadzić pewien wymiar cząstek elementarnych, a więc nową, analogiczną do hiperładunku liczbę kwantową; taką liczbą może być powab C. Zatem podstawowy multiplet symetrii $SU(4)$ byłby czeroskładnikowy, i zawierałby — oprócz podstawowego trypletu symetrii $SU(3)$ — jeszcze singlet o $C = 1$. Prawa analogiczne do (15) i (16) miałyby wtedy postać:

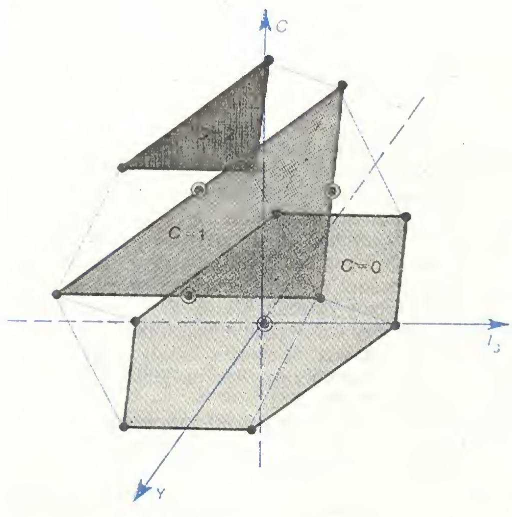
$$4 \otimes 4 = 1 \oplus 15 \qquad (17)$$

oraz

$$4 \otimes 4 \otimes 4 = 4 \oplus 20_1 \oplus 20_1 \oplus 20_2. \qquad (18)$$

I w tym wypadku trzeba wprowadzić oprócz 4 multiplet sprzężony $\bar{4}$, w którym dodatkowy izosinglet ma $C = -1$. Zauważmy przy tym, że w prawie wyrażonym wzorem (18) występują dwa różne multiplety dwudziestowymiarowe, odróżnione od siebie dolnym wskaźnikiem 1 i 2. Multiplety te mają całkowicie odmienną strukturę. W 20_1 jest zawarty oktet symetrii $SU(3)$ o $C = 0$, a poza tym sekstet $SU(3)$

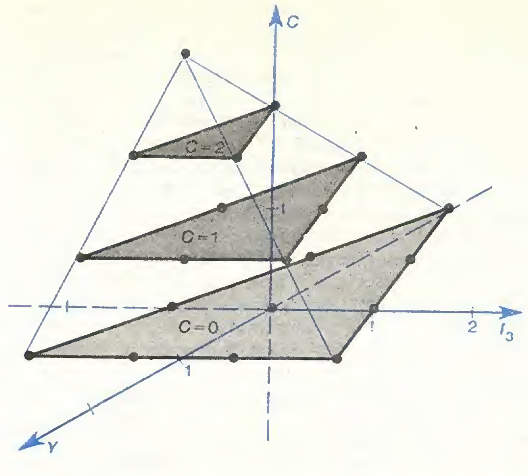
multiplet
20_1



stany obsadzone przez 2 cząstki

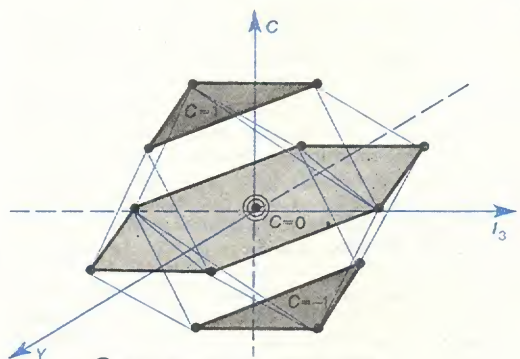
Rys. 10. Multiplet 20_1 symetrii $SU(4)$ przedstawiony w przestrzeni zmiennych (I_3, Y, C). Multiplet ten obejmuje oktet barionów, położony w płaszczyźnie $C = 0$, oraz jeden sekstet o $C = +1$, jeden tryplet o $C = 1$ i wreszcie jeden tryplet o $C = 2$. Niektóre cząstki należące do multipletów $C \neq 0$ zostały już z zupełną pewnością zidentyfikowane w doświadczeniu. Wszystkie bariony należące do tego multipletu miałyby $J^P = \frac{1}{2}^+$

multiplet
20_2



Rys. 11. Multiplet 20_2 symetrii $SU(4)$ obejmujący bariony o $J^P = \frac{3}{2}^+$. W multiplecie tym w płaszczyźnie $C = 0$ znajduje się dekuplet 10 symetrii $SU(3)$. Inne cząstki należące do 20_2 nie zostały jeszcze wykryte

multiplet
15



stan obsadzony przez 3 cząstki

Rys. 12. Multiplet 15 symetrii $SU(4)$. Ma on tę samą budowę niezależnie od wartości J^P. Dla mezonów o $J^P = 1^-$ w płaszczyźnie $C = 0$ znajdowałyby się następujące cząstki: ρ^-, ρ^0, ρ^+, φ, ω, J/ψ K^{*0}, K^{*+}, K^{*0}, $\bar{K}^{*-}$

o $C = +1$, i dwa tryplety odpowiednio o $C = +1$ i $C = +2$ (rys. 10). Natomiast w 20_2 mieści się dekuplet $SU(3)$ o $C = 0$, a poza tym sekstet $SU(3)$ o $C = +1$, tryplet $SU(3)$ o $C = +2$ i wreszcie singlet $SU(3)$ o $C = +3$ (rys. 11). Struktura multipletu 15 obejmującego mezony (podczas gdy 20_1 i 20_2 obejmowałyby bariony) jest przedstawiona na rys. 12. Multiplet ten zawierałby oprócz oktetu symetrii $SU(3)$ o $C = 0$ także dwa tryplety odpowiednio o $C = +1$ i $C = -1$ oraz dodatkową cząstkę niepowabną, singlet o $C = 0$. Gromadzenie bezpośrednich dowodów doświadczalnych występowania symetrii $SU(4)$ rozpoczęło się w gruncie rzeczy od wykrycia mezonu J/ψ, który interpretuje się jako cząstkę związaną z tym właśnie dodatkowym singletem niepowabnym symetrii $SU(4)$.

W poprzednich punktach wspominaliśmy o tym, że różnice mas wiążące się z poszczególnymi oddziaływaniami powiększają się wraz ze wzrostem siły oddziaływania. Schemat ten można by rozbudowywać po przejściu do symetrii $SU(4)$. Istotnie, różnice mas mogą być jeszcze większe. Na przykład cząstka J/ψ, która byłaby analogiem cząstki φ, lecz należałaby do innego multipletu symetrii $SU(3)$, pozostając w tym samym multiplecie symetrii $SU(4)$ różniłaby się od niej masą aż o ok. 2 GeV/c^2.

Podsumowując te rozważania, można obecnie lepiej uzasadnić pogląd, że im większe jest natężenie oddziaływania, tym większa też jest jego symetria. Mogliśmy tę zastanawiającą regularność śledzić

poczynając od oddziaływań podsłabych do oddziaływań bardzo silnych.

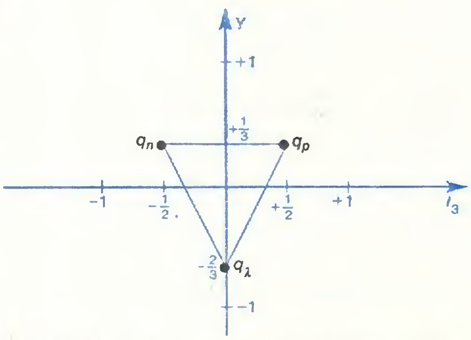
Kwarki

Jest rzeczą zastanawiającą, że znane z doświadczenia cząstki wypełniają tak niewiele różnych multipletów symetrii $SU(3)$: mezony tylko 1 i 8, a bariony tylko 1, 8 i 10. Nie są zatem wykorzystane przez przyrodę (przynajmniej w obrębie cząstek niepowabnych) inne liczne możliwości, włączając tu, co jest może szczególnie dziwne, także multiplet podstawowy $SU(3)$, a mianowicie triplety 3 i $\bar{3}$. Z wzorów (15) i (16) wynika, że te i tylko te multiplety, które są obsadzone w naturze można zbudować w wypadku mezonów z multipletów 3 i $\bar{3}$, a w wypadku barionów z trzech trypletów 3. Załóżmy więc, że multiplet 3 jest obsadzony przez hipotetyczne cząstki o spinie $^1/_2$ zwane kwarkami ($\bar{3}$ obsadzałyby więc antykwarki). B dla kwarków wynosiłoby $^1/_3$. Wówczas, trzy kwarki miałyby łącznie $B = 1$, a kwark i antykwark — $B = 0$, przy czym układy trzykwarkowe miałyby spin połówkowy, a dwukwarkowe — całkowity.

własności kwarków

Założenie to wywołuje dalsze nieoczekiwane konsekwencje. Po pierwsze, dostajemy wówczas także ułamkowe wartości dla hiperładunku i ładunku elektrycznego kwarków. I tak np. dla dubletu kwarków o $I = {}^1/_2$ (oraz $S = 0$) mamy $Y = {}^1/_3$, natomiast dla singletu $I = 0$ (i $S = -1$) — $Y = -{}^2/_3$. Po drugie, ze wzoru Gell-Manna–Nishijimy wynika zaś, że ładunki kwarków są wielokrotnościami nie ładunku elementarnego e, lecz $e/3$ (patrz tabela poniżej, rys. 13).

tryplet kwarków



Rys. 13. Multiplet podstawowy symetrii $SU(3)$ — trypletu kwarków, przedstawionego w płaszczyźnie (I_3, Y)

Liczby kwantowe kwarków

Symbol kwarku	B	S	Y	C	B'	I	I_3	Q
u	1/3	0	1/3	0	0	1/2	+1/2	2/3
d	1/3	0	1/3	0	0	1/2	−1/2	−1/3
s	1/3	−1	−2/3	0	0	0	0	−1/3
c	1/3	0	1/3	1	0	0	0	2/3
b	1/3	0	1/3	0	−1	0	0	−1/3

Fakt ten rokuje nadzieje na możliwość wykrycia kwarków swobodnych, gdyż, zgodnie z prawem zachowania ładunku, cząstki o ładunku ułamkowym nie mogą się rozpadać na cząstki o ładunkach wyłącznie całkowitych. Tak więc co najmniej jeden kwark powinien być cząstką trwałą, stosunkowo łatwą do wykrycia. Jednakże dotychczasowe poszukiwania nie dały żadnego wyniku. Brak danych potwierdzających istnienie kwarków swobodnych można wyjaśniać jedną z czterech możliwych hipotez.

Po pierwsze, kwarki mogą być cząstkami bardzo ciężkimi (obecnie dolna granica eksperymentalna byłaby dla cząstek silnie oddziałujących rzędu 10 GeV/c^2). Gdyby tak było używane obecnie akceleratory nie mogłyby ich wyprodukować. Po drugie, kwarki mogłyby być cząstkami stosunkowo lekkimi, ale nie oddziałującymi dostatecznie silnie (wówczas byłyby one wytwarzane rzadko i szansa ich zaobserwowania byłaby odpowiednio mniejsza). Po trzecie, być może kwarki w ogóle nie istnieją. Po czwarte wreszcie, być może kwarki mogą istnieć wyłącznie w stanie związanym, a więc wewnątrz hadronów.

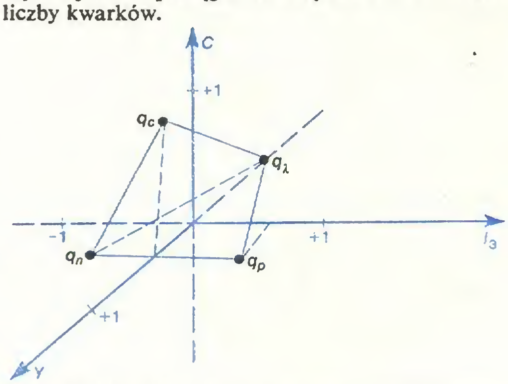
Żadna z wymienionych hipotez nie może być w tej chwili odrzucona. Jednakże druga z nich przedstawia się dziwacznie: oto cząstki zbudowane z kwarków mogłyby oddziaływać silniej niż ich składniki. Także i pierwsza hipoteza nie jest najlepsza. Bardzo ciężki kwark wiążąc się w stosunkowo lekki barion musiałby tracić ogromną część swej masy na wiązanie. Tak duża energia wiązania wskazywałaby na ogromną siłę oddziaływań wiążących kwarki. Kwarki takie musiałyby mieć same bardzo skomplikowaną strukturę chmury, co trochę nie jest zgodne z przedstawianiem ich jako obiektów elementarnych. W każdym razie wtedy teoria kwarka musiałaby być może nawet bardziej skomplikowana niż teoria hadronu. Poza tym taka chmura dawałaby zapewne bardzo duży wkład do momentu magnetycznego kwarka (→ Oddziaływania elektromagnetyczne), i to momentu anomalnego, polowego. Ten wniosek jest, jak się wydaje, niezgodny z prostym modelem, z którego można obliczyć momenty magnetyczne nukleonów, zakładając pewien efektywny moment magnetyczny kwarków. Trzecia koncepcja także nie znalazłaby obecnie zbyt wielu zwolenników, gdyż wiele prostych, nawet jakościowych przewidywań modelu kwarkowego jest zdumiewająco zgodnych z faktami doświadczalnymi.

hipoteza kwarków uwięzionych

Pozostaje więc hipoteza czwarta, kwarków uwięzionych wewnątrz hadronu. Jaka miałaby być jednak przyczyna tego uwięzienia, nie jest łatwo ustalić, toteż sformułowanie konsekwentnej teorii opartej na tej hipotezie jest dalekie od zakończenia.

Przejście do modelu opartego na symetrii $SU(4)$ wymaga powiększenia liczby kwarków do czterech. Czwarty kwark byłby kwarkiem powabnym, o $C = +1$ (tabela obok i rys. 14). Przejście do jeszcze wyższych symetrii pociąga za sobą dalsze rozszerzenie liczby kwarków.

kwartet kwarków



Rys. 14. Multiplet podstawowy symetrii $SU(4)$ — kwartetu kwarków przedstawionego w przestrzeni zmiennych (I_3, Y, C)

Patrząc na rozmaite możliwe podklasy oddziaływań silnych z punktu widzenia modelu kwarkowego, dostrzegamy inną możliwość wyjaśnienia łamania symetrii $SU(3)$ a także $SU(4)$. Można bowiem przyjąć, że oddziaływania kwarków są w pełni zgodne z zasadami symetrii $SU(3)$ czy też $SU(4)$, ale że kwarki te mają różne masy niepolowe. Kwark dziwny byłby wówczas cięższy od kwarków niedziwnych o ok. 200–300 MeV/c^2, kwark powabny od kwarków niepowabnych o ok. 1–2 GeV/c^2.

Gdyby hipoteza kwarków uwięzionych okazała się poprawna, pytanie, która z wyżej wymienionych koncepcji łamania wyższych symetrii jest słuszna, miałoby charakter w znacznym stopniu akademicki. Nie ma-

jąc do dyspozycji kwarków swobodnych, nie moglibyśmy z sensem mówić o ich masie, a tylko o ich masie efektywnej wewnątrz hadronu.

Jednym ze sprawdzianów ogólnej poprawności obrazu kwarkowego jest zagadnienie istnienia cząstek egzotycznych. Mianem tym określa się ogólnie takie cząstki, które by były składnikami innych multipletów symetrii $SU(3)$ poza singletem i oktetem dla mezonów oraz singletem, oktetem i dekupletem dla barionów. Przykładami cząstek egzotycznych byłyby: a) mezon niedziwny podwójnie naładowany o $I = 2$ lub więcej; b) mezon dziwny o $I = {}^3/_2$; c) barion o $S = +1$. Żadna jednak z tych cząstek (ani innych cząstek egzotycznych) nie została wykryta. Gdyby to w przyszłości nastąpiło, to nie musielibyśmy rezygnować z modelu kwarkowego, a tylko rozszerzyć schemat budowy hadronów z kwarków. Oprócz układów kwark-antykwark lub też trzy kwarki należałoby dopuścić np. układy dwa kwarki i dwa antykwarki czy też cztery kwarki i jeden antykwark. Wobec tego niewystępowanie w przyrodzie cząstek egzotycznych przemawia nie tylko za modelem kwarkowym, ale także dodatkowo za pewnym konkretnym schematem budowy hadronów z kwarków.

cząstki egzotyczne

Sytuacja jednak nie może być tak prosta, jakby to wynikało z poprzednich uwag. I rzeczywiście, między kwarkami muszą przecież działać jakieś siły, skoro łączą się one w układy o takim stopniu trwałości jak hadrony. Musi to więc być jakieś nowe pole odpowiadające tym siłom oraz cząstki, które by mogły być wymieniane między kwarkami. Takie hipotetyczne obiekty nazywa się gluonami. Byłyby to cząstki o $B = Q = C = S = I = L = 0$ oraz $J^P = 1^-$. Ta ostatnia własność gluonów wynikałaby z założenia, iż fundamentalną rolę w przenoszeniu oddziaływań odgrywają właśnie cząstki fotonopodobne. (Również bozon, czy też bozony przenoszące oddziaływania słabe miałyby $J^P = 1^-$). Posuwając się o krok dalej należałoby powiedzieć, iż skoro istnieją gluony, zdolne do oddziaływań z kwarkami, to niewątpliwie wewnątrz hadronu zachodzą też procesy wirtualnej kreacji par kwark-antykwark z gluonów. Hadron byłby więc bardzo skomplikowaną strukturą, w której występowałyby: a) zespół kwarków i antykwarków decydujący o własnościach hadronu z punktu widzenia symetrii $SU(3)$, a więc o jego zdolności do wchodzenia w oddziaływania silne (te kwarki nazywamy często walencyjnymi); b) gluony; c) wiele („morze”) par kwark-antykwark w takich samych stanach jak gluony, a więc o $C = S = Q = I = 0$. Oczywiście pary te, będąc singletami $SU(3)$, nie naruszałyby przyporządkowania danego hadronu do określonego multipletu tej symetrii.

gluony

Przedstawienie hadronów jako układów zbudowanych z kwarków walencyjnych, gluonów oraz par kwark-antykwark należących do „morza”, jest w ogólnych zarysach potwierdzone przez dane doświadczalne dotyczące zderzeń głęboko nieelastycznych leptonów z hadronami.

Kolor jako nowa liczba kwantowa

Do rozwiązania pozostał jeszcze niezmiernie skomplikowany problem dynamiczny: jak kwarki wiążą się w hadronie i czy rzeczywiście powinny one być w nim całkowicie uwięzione. Pochodzenie sił działających między kwarkami jest trudne do rozszyfrowania. Jedną z trudności stanowi konieczność zrozumienia, dlaczego układy dwukwarkowe są wiązane słabo (jeśli już w ogóle), natomiast układy trzykwarkowe — mocno, tak że tworzą bariony. Siły kwark-kwark nie są więc z jakichś powodów wystarczająco silne w układach dwukwarkowych. Czy są to wobec tego siły głównie trzyciałowe? Jak fakt ten uwzględnić w dynamicznej teorii kwarków?

Dodatkową komplikacją są niezwykłe własności symetrii funkcji falowej w związanych układach trzykwarkowych. Wszystkie cząstki o spinie połówkowym, a więc i kwarki powinny spełniać zasady statystyki Fermiego–Diraca. Funkcja falowa układu fermionów powinna być całkowicie antysymetryczna względem zamiany dwu fermionów należących do tego układu. Na te własności symetrii funkcji falowej składają się w wypadku kwarków związanych w barionie własności symetrii: — względem przestawienia liczb kwantowych odpowiadających symetrii $SU(3)$, — względem przestawienia zmiennych spinowych oraz — względem zamiany położeń. Z punktu widzenia dynamicznej struktury oddziaływań kwark-kwark interesujący jest przede wszystkim ten ostatni czynnik w funkcji falowej. Pewne informacje na ten temat można uzyskać w następujący sposób. Załóżmy, że wszystkie hadrony zbudowane z kwarków niepowabnych tworzą multiplety symetrii $SU(6)$, na którą składają się symetria $SU(3)$ oraz symetria $SU(2)$ zwykłego spinu. Podstawowy multiplet tej symetrii $SU(6)$ — zgodnie z ogólnymi regułami — będzie mieć 6 składowych, przy czym jako jedną składową traktuje się jeden kwark znajdujący się w określonym stanie spinowym. Ponieważ spin kwarku z założenia wynosi $^1/_2$, przeto każdy kwark może występować w dwu stanach spinowych $J_z = +{}^1/_2$ oraz $J_z = {}^1/_2$.

trudności dynamicznej teorii kwarków

Okazuje się, że:

$$6 \otimes \bar{6} = 1 \oplus 35, \tag{19}$$

$$6 \otimes 6 \otimes 6 = 20 \oplus 56 \oplus 70 \oplus 70. \tag{20}$$

Tak więc wszystkie stany mezonowe powinny należeć do multipletów 1 i 35 symetrii $SU(6)$, a wszystkie stany barionowe do jednej lub więcej rodziny spośród 20, 56 i 70. Na multiplet 35 składa się oktet $SU(3)$ o spinie 0 (8 stanów), plus singlet $SU(3)$ o spinie 1 (3 stany, bo spin 1 ma trzy możliwe ustawienia), i wreszcie oktet $SU(3)$ też o spinie 1, a więc łącznie 24 stany. Spośród multipletów barionowych na szczególną uwagę zasługuje multiplet 56, mieszczący w sobie oktet $SU(3)$ o spinie $^1/_2$ (16 stanów) i dekuplet $SU(3)$ o spinie $^3/_2$ (40 stanów). Jest rzeczą naturalną zaliczenie do tego multipletu symetrii $SU(6)$ znanego dekupletu barionowego zawierającego m.in. bariony Δ oraz oktetu barionowego zawierającego m.in. nukleony.

Można jednak sprawdzić, że multiplet symetrii $SU(6)$ o wymiarze 56 jest opisywany funkcją falową całkowicie symetryczną względem przestawień stopni swobody kwarków odpowiadających grupie $SU(3)$ i grupie $SU(2)$ zwykłego spinu. Tym samym funkcja falowa zarówno nukleonów jak i barionów z dekupletu musi mieć część przestrzenną antysymetryczną względem przestawień położeń kwarków. Z drugiej jednak strony jest to stan o najniższej energii układu trzech kwarków (nie ma lżejszych barionów), a więc stan zwany stanem podstawowym. Największa trudność polega na tym, że nikomu nie udało się skonstruować takiego potencjału oddziaływania, który by dawał antysymetryczną funkcję falową (przestrzenną) w stanie podstawowym.

Tę trudność dynamicznej teorii kwarków można ominąć, zakładając, że mają one jeszcze jeden stopień swobody, który należy włączyć do rozważań nad symetrią układów trójkwarkowych. Ten dodatkowy stopień swobody nazywa się kolorem. Zgodnie z tą hipotezą kwarki mogłyby istnieć w trzech kolorach, których zespół tworzyłby multiplet podstawowy nowej grupy $SU(3)$, zwanej grupą koloru, niezależnej od poprzednio wprowadzonej grupy $SU(3)$ izospinowo-hiperładunkowej.

grupa koloru

Najczęściej zakłada się, że wszystkie cząstki obserwowane w przyrodzie są singletami symetrii $SU(3)$ koloru, a więc, są cząstkami białymi. Zagadnienie istnienia hadronów kolorowych i możliwości wykrycia ich doświadczalnie w przyszłości jest otwarte, choć obecnie przeważa pogląd, że kolor jest liczbą kwantową wewnątrzhadronową, która na zewnątrz nie może się ujawnić. Natomiast koncepcja koloru jest właściwym rozwiązaniem zagadnienia budowy bario-

nów białych: singlet $SU(3)$ jest bowiem opisywany funkcją falową całkowicie antysymetryczną. Antysymetria funkcji falowej układu trzech kwarków byłaby więc przesunięta z części przestrzennej na część kolorową tej funkcji. Hipoteza koloru znajduje także pośrednie potwierdzenie w danych doświadczalnych dotyczących anihilacji par elektron-pozyton w obszarze bardzo wysokich energii. Wprowadzenie grupy $SU(3)$ koloru jako symetrii (ścisłej!) świata hadronów jest dodatkowym elementem świadczącym o wysokiej symetrii oddziaływań silnych.

kolor leptonowy

Często uogólnia się pojęcie koloru, tworząc kolor czwarty, leptonowy. W takiej koncepcji podstawowe dwie grupy symetrii cząstek elementarnych byłyby grupą $SU(4)$ (grupą koloru — trzy kolory kwarkowe i jeden leptonowy) oraz grupą izospinowo-hiper-ładunkowo-powabną $SU(n)$ ($n = 6$?). W tym jednolitym obrazie świat byłby zbudowany z 16 fermionów (4 leptonów oraz 4 kwarków w trzech odmianach koloru każdy) oraz z kilku cząstek o $J^P = 1^-$ (byłyby to fotony, gluony oraz bozony przenoszące oddziaływania słabe), składających się na dynamikę układów fermionów.

Wspominaliśmy już o próbach dodania do 4 kwarków jeszcze dwu dalszych kwarków, cięższych nawet od kwarka powabnego. Wówczas musielibyśmy oczekiwać wykrycia nowych cząstek o niezwykłych własnościach. Na ślad jednej z nich (odpowiednika cząstki J/ψ) udało się już natrafić. Jest to bardzo ciężki mezon o masie ponad 9,4G eV/c^2, niedawno zaobserwowany. Niezależnie od tych sugestii rozszerzyła się lista leptonów do 6 przez dodanie do czterech znanych leptonów jeszcze dwu leptonów, jednego o masie 1782 GeV/c^2 (lepton τ) i drugiego o masie bardzo małej (neutrino taonowe). Nie jest więc wykluczone, że należy konstruować świat z 24 rozmaitych fermionów, a mianowicie 6 leptonów i 6 kwarków, z których każdy występowałby w trzech stanach koloru.

hipoteza: 6 leptonów i 6×3 kwarków

Jak wynika z powyższych danych, fizyka cząstek elementarnych znajduje się wciąż na etapie wstępnym, mimo ogromnych osiągnięć. Podstawową jej wadą jest brak aparatu teoretycznego, który by umożliwiał wykonanie obliczeń, sprawdzających w jednoznaczny sposób konkretne hipotezy budowy cząstek i ich oddziaływań. Istnieje oczywiście wiele rozmaitych metod rachunkowych opierających się na mniej czy bardziej wiarygodnych założeniach. Często mają one zastosowanie do jednego tylko rodzaju oddziaływań, a nawet do jeszcze węższego kręgu zjawisk.

Mimo tego braku, z istniejących danych doświadczalnych oraz koncepcji teoretycznych wyłania się coraz lepiej uporządkowany obraz świata cząstek elementarnych, zawierających więcej czynników unifikacyjnych. Niewątpliwie największym osiągnięciem w tej dziedzinie jest sformułowanie teorii ujmującej wspólnie oddziaływania słabe i elektromagnetyczne. To postępujące porządkowanie koncepcji dotyczących cząstek elementarnych jest nierozerwalnie związane z hipotezą kwarków. Jest ona, niezależnie od wciąż jeszcze ogromnych trudności pojęciowych i matematycznych, niewątpliwie jednym z najważniejszych składników obecnego obrazu cząstek elementarnych i ich oddziaływań. Na razie trudno sobie wyobrazić, aby przyszła teoria cząstek elementarnych mogła się obejść bez pojęcia kwarku.

G. Białkowski, R. Sosnowski *Cząstki elementarne*, Warszawa 1971; B.H. Bransden, D. Evans, J.V. Major *Cząstki elementarne*, Warszawa 1981; J.J.J.Kokkedee *The Quark Model*, New York 1969 (ros. Moskwa 1971); *Mezony, grawitacja, antymateria*, Warszawa 1962; J. Nowożyłow *Cząstki elementarne*, Warszawa 1961.

Struktura cząstek elementarnych

Michał Święcki

Mówiąc o strukturze badanego ciała, mamy zwykle na myśli jego skład. Wiemy np., że kawałek żelaza składa się z atomów, które z kolei składają się z elektronów i jąder atomowych. Te ostatnie są również obdarzone strukturą — składają się z protonów i neutronów. Wydaje się, że moglibyśmy rozkładanie struktur na ich składniki kontynuować bez istotnych zmian natury metodologicznej. Jednak gdy przechodzimy do obiektów niesłychanie małych, jakimi są cząstki elementarne (→ Cząstki elementarne i ich oddziaływania), ta wydawałoby się naturalna metoda postępowania przestaje być prosta i łatwa. Jest tak dlatego, że dla obiektów o rozmiarach rzędu np. 10^{-15} m i charakterystycznych czasach trwania oddziaływań rzędu np. 10^{-24} s intuicyjny opis klasyczny zupełnie zawodzi. Obserwowane w doświadczeniach zachowanie się cząstek elementarnych ma, jak się okazuje, charakter czysto statystyczny. Statystyczna jest też teoria, za pomocą której z ogromnym powodzeniem opisujemy wyniki tych doświadczeń. Teorią, którą mamy na myśli jest oczywiście mechanika kwantowa a właściwie kwantowa teoria pola (→ Teoria pola, Oddziaływania elektromagnetyczne).

Statystyczny charakter struktury cząstek

Wróćmy do kawałka żelaza. Podgrzany do odpowiedniej temperatury zaczyna świecić, czyli promieniować świetlne fale elektromagnetyczne. Emisja, a także absorpcja światła odbywa się zawsze porcjami, które zwiemy fotonami. Tak więc rozżarzony (i nie tylko) kawałek żelaza promieniuje fotony. Można by się więc zapytać, z ilu fotonów składa się ten kawałek. Jest rzeczą oczywistą, że nie potrafimy podać żadnego rozsądnego sposobu pomiaru tej liczby. Możemy ją określić jedynie statystycznie, szacując np. średnią liczbę fotonów na jednostkę czasu. Zupełnie podobnie jest w wypadku struktury protonów, neutronów i innych cząstek. Skład ich możemy określić jedynie statystycznie, podając prawdopodobieństwo (na jednostkę czasu) wystąpienia pewnej struktury. Dlaczego więc mówiąc o jądrze atomowym mamy prawo twierdzić z całym przekonaniem, że składa się ono z określonej liczby protonów i neutronów, skoro struktury samych nukleonów nie możemy poznać jednoznacznie?

Popatrzmy na te zagadnienia nieco bardziej formalnie. Podstawowymi regułami obowiązującymi w fizyce są tzw. zasady nieokreśloności, które określają warunki, przy których zjawisko fizyczne ma przebieg klasyczny, a także warunki, w których istotne stają się cechy statystyczne, kwantowe. Jedną z owych zasad jest zasada nieokreśloności czasu i energii:

zasada nieokreśloności czasu i energii

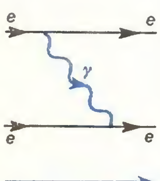
$$\Delta E \cdot \Delta t \geqslant {}^1\!/_2 \hbar,$$

która wiąże minimalny czas Δt, po jakim można przy użyciu dowolnych metod zaobserwować zmianę stanu badanego obiektu z nieokreślonością ΔE jego energii (a więc i masy, bo $E = mc^2$). Stała Plancka h jest bardzo mała ($\hbar = h/2\pi = 1{,}05 \cdot 10^{-34}$ J·s) i dla obiektów makroskopowych iloczyn $\Delta E \cdot \Delta t$ wielokrotnie ją przewyższa (np. 10^{30} razy), co prowadzi do klasycznego, jednoznacznego ich zachowania się. Dla cząstek elementarnych natomiast iloczyn ten niewiele różni się od $\hbar$ i powyższa zasada mocno ogranicza własności tych cząstek. Gdybyśmy np. obserwowali atomy żelaza w przedziałach czasu rzędu 10^{-10} s, to tak dokładnymi pomiarami wywołalibyśmy nieokreśloność ich energii rzędu 10 eV.

Doprowadziłoby to do niekontrolowanej emisji fotonów o takiej właśnie energii, a więc o długościach fal rzędu charakterystycznego dla fal świetlnych (gdyż $E = h\nu = hc/\lambda$). Pomiary przeprowadzane z mniejszą dokładnością nic by nie dały, gdyż już czas 10^{-10} s wielokrotnie przewyższa typowy czas emisji fotonu. Dlatego właśnie skład fotonowy kawałka żelaza możemy określić jedynie statystycznie. Przy opisie struktury kawałka żelaza nie jest to jednak ważne, gdyż wkład fotonów do energii (masy) żelaza jest zupełnie nieistotny. Przy opisie struktury nukleonów zagadnienia podobnego typu stają się znacznie bardziej istotne, gdyż masy składników nukleonów mogą być rzędu masy samych nukleonów. Nukleonowa struktura jąder atomowych jest z kolei równie prosta (?), jak struktura atomowa kawałka żelaza. Nukleony bowiem mają wewnątrz jąder masy rzędu 1 GeV (ubytek ich masy na koszt energii wiązania jest niewielki i wynosi ok. 8 MeV) i dopiero pomiary przeprowadzane z niewiarygodną dokładnością 10^{-24} s mogłyby doprowadzić do wykrycia w jądrze dodatkowych par nukleonów, czy też innych hadronów (masa najlżejszych z nich — mezonów π — wynosi $m_\pi \approx 140$ MeV). Wykonując pomiary z mniejszą dokładnością będziemy zawsze obserwować typowy nukleonowy skład jąder. Nie jest to przypadek odosobniony. Okazuje się, że dla każdej cząstki elementarnej można określić takie warunki doświadczalne, przy których skład cząstki staje się prawie jednoznaczny. Dzięki temu możemy w ogóle mówić o jakimkolwiek ich składzie.

Kwantowopolowy opis własności cząstek

Przejdźmy do zagadnienia struktury cząstek. Jak już mówiliśmy, zgodnie z zasadą nieokreśloności w dostatecznie krótkim przedziale czasu nieokreśloność energii (masy) każdego obiektu fizycznego, a więc i cząstki, jest tak duża, że może nastąpić niekontrolowana emisja innych cząstek (nie tylko fotonów). Taki krótkotrwały proces wysyłania cząstek przez inne cząstki jest podstawą występowania wszelkich rodzajów oddziaływań w przyrodzie. Na rys. 1 przedstawiliśmy proces, w którym elektron wysłał foton (tzw. foton wirtualny), pochłonięty następnie po krótkiej chwili przez inny elektron. W wyniku tej wymiany fotonu nastąpiło przekazanie energii oraz pędu od jednego elektronu do drugiego, czyli ich oddziaływanie (rozpraszanie). Szczegółowa teoria tego rodzaju oddziaływań została opisana w artykule „Oddziaływania elektromagnetyczne". Tam też znajduje się opis idei trudnych a zarazem pięknych doświadczeń, które umożliwiły poznanie struktury cząstek. Doświadczalne zbadanie owej struktury nie było bowiem wcale trywialne. Nie będziemy tu powtarzać wszystkich argumentów doświadczalnych dotyczących składu cząstek elementarnych. Przedstawimy jedynie współczesny stan wiedzy na ten temat. Niestety nie jest to jeszcze wiedza kompletna. Mimo to w ciągu ostatnich lat zrobiono ogromny krok w kierunku pełnego zrozumienia świata cząstek. Wystarczy powiedzieć, że przed 1969 rokiem artykuł ten nie mógłby zawierać niemal żadnych informacji na temat struktury hadronów.



Rys. 1. Rozpraszanie elektronu na elektronie zachodzi przez wymianę fotonu wirtualnego

Podstawą opisu oddziaływań cząstek elementarnych jest kwantowa teoria pola, której najlepiej zbadaną częścią jest elektrodynamika kwantowa (→ Teoria pola). Jest to teoria lokalna, co znaczy, że np. proces emisji (lub absorpcji) fotonu z elektronu zachodzi lokalnie, w jednym miejscu i jednej chwili, choć zarówno owo miejsce, jak i chwila są nieokreślone statystycznie. W związku z tym podstawą badań fizyki cząstek elementarnych są cząstki bez struktury, punktowe. Takie cząstki mogą bowiem punktowo wysyłać inne, również punktowe cząstki. Przymiotnik „punktowy" należy tu rozumieć umownie. Żadnej cząstki nie można dostatecznie dobrze zlokalizować doświadczalnie, gdyż zgodnie z zasadą nieokreśloności pędu i położenia,

cząstki punktowe

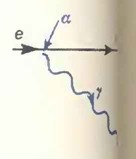
$$\Delta p \cdot \Delta x \geqslant {}^1/_2 \hbar,$$

zasada nieokreśloności pędu i położenia

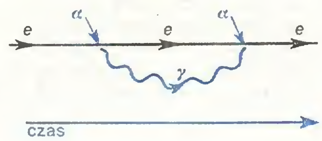
doprowadziłoby to do nadania jej dużego pędu, a więc i energii, co w konsekwencji wywołałoby emisję innych cząstek. Dokładnie zlokalizowana cząstka nie może być cząstką pojedynczą. Zwróćmy przy okazji uwagę na fakt, że daleko nam do osiągnięcia dokładności 10^{-15} m (rozmiary nukleonów). Wracając do struktury cząstek elementarnych — próbujemy opisywać ją poprzez składniki nie mające już żadnej struktury. Na szczęście okazuje się, że opis taki jest możliwy.

Struktura leptonów

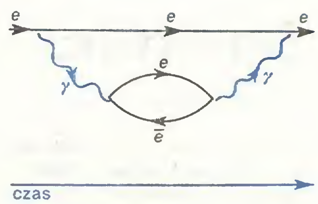
Najprostszym obiektem badań są leptony (e, μ, ν_e, ν_μ, τ) nie podlegające oddziaływaniom silnym. Najsilniejszymi są dla nich oddziaływania elektromagnetyczne, które charakteryzują się stałą sprzężenia $\alpha_{el} \approx {}^1/_{137}$. Znaczy to, że prawdopodobieństwo emisji fotonu z elektronu bądź mionu (neutrina są pozbawione ładunku elektrycznego) jest proporcjonalne właśnie do α_{el} (rys. 2). Wysłany foton musi po krótkim czasie być pochłonięty, np. przez ten sam elektron (rys. 3), a prawdopodobieństwo tego jest rzędu $\alpha_{el}^2 \approx ({}^1/_{137})^2$. Wysłany foton może też wytworzyć parę elektron-pozyton, która następnie, wciąż bardzo szybko, zamieni się z powrotem w foton. Prawdopodobieństwo każdego z tych procesów jest znów rzędu α_{el}. Prawdopodobieństwo, że elektron na krótką chwilę przejdzie w elektron oraz parę elektron-pozyton, jest więc rzędu $\alpha_{el}^4 \approx ({}^1/_{137})^4$ (rys. 4). To już bardzo mało. Możemy więc powiedzieć, że fizyczny elektron składa się z prawdopodobieństwem (na jednostkę czasu) bliskim jedności z jednego elektronu bez struktury, z prawdopodobieństwem rzędu $({}^1/_{137})^2$ z pozbawionych struktury elektronu i fotonu oraz z prawdopodobieństwem rzędu $({}^1/_{137})^4$ z elektronu i pary elektron-pozyton itd. (rys. 5). Elektron i wszystkie inne leptony są zatem niemal pozbawione struktury. Jest to wywołane małą



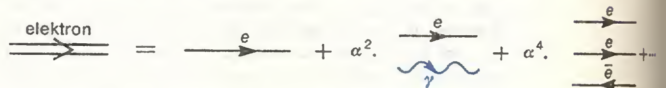
Rys. 2. Prawdopodobieństwo zajścia procesu emisji fotonu jest rzędu $\alpha_{el} \approx {}^1/_{137}$



Rys. 3. Prawdopodobieństwo pochłonięcia fotonu przez ten sam elektron jest rzędu α_{el}^2



Rys. 4. Prawdopodobieństwo zajścia takiego procesu jest rzędu α_{el}^4



Rys. 5. Różny skład elektronu fizycznego ma różne prawdopodobieństwo na jednostkę czasu

wartością elektromagnetycznej stałej sprzężenia, α_{el}. Oddziaływania słabe leptonów, jako jeszcze słabsze, nie zmieniają tego obrazu. Neutrina więc (jak i fotony) też są bez struktury.

Kwarkowa struktura hadronów

W przypadku leptonów sytuacja była prosta, komplikuje się jednak, gdy przechodzimy do opisu hadronów. Cząstki te podlegają oddziaływaniom silnym,

których stała sprzężenia nie jest mała. Dlatego poświęcono wiele lat na zbadanie struktury hadronów. W wyniku tych badań prawdziwy obraz rysuje się dosyć jasno.

Podstawą struktury wszystkich hadronów są pozbawione już struktury kwarki. Wprawdzie nikomu nie udało się wybić kwarku ze struktury hadronu, ale niewielu jest fizyków, którzy nie wierzyliby w kwarkową naturę oddziaływań silnych. Kwarki tkwią więc wewnątrz hadronów jak głaz w studni i trzeba ogromnej (prawdopodobnie nieskończonej) energii, żeby je stamtąd wydobyć. Podstawowym mankamentem teorii kwarków jest niekompletne poznanie natury owych ogromnych sił, które nie pozwalają kwarkom na wydostanie się z wnętrza hadronów. Dlatego należy być może traktować tę teorię z pewną rezerwą. Mimo to teoria kwarków wyjaśnia tyle faktów i jest potwierdzona przez tyle niezależnych od siebie doświadczeń, że rezerwa ta powinna być naprawdę niewielka.

Zobaczymy dalej, że hadrony składają się nie tylko z kwarków, ale i z pewnych innych cząstek — nośników oddziaływań międzykwarkowych. Jednak kwarki pełnią w strukturze hadronów rolę szczególną. Tylko bowiem kwarki biorą udział w oddziaływaniach elektromagnetycznych i słabych. Różne masy kwarków decydują też o różnicach mas samych hadronów, a obserwowane w doświadczeniach prawo zachowania liczby barionów ma swoje źródło w ścisłym prawie zachowania liczby kwarków (liczba kwarków minus liczba antykwarków nie zmienia się w żadnym procesie). Można powiedzieć, że własności kwarków (przede wszystkim ich masy) są odpowiedzialne za wszelkiego rodzaju symetrie obowiązujące w świecie hadronów.

hipoteza kwarkowa

Jakie więc są same kwarki? Najważniejsza, sprawdzona w wielu doświadczeniach hipoteza teorii kwarków mówi, że wszystkie bariony (hadrony o spinach połówkowych: $^1/_2$, $^3/_2$ itd.) składają się z trzech kwarków, zaś mezony (hadrony o spinach całkowitych: 0, 1 itd.) z pary kwark-antykwark. Antybariony składają się oczywiście z trzech antykwarków. Wynika z tego, że kwarki są fermionami o spinie $^1/_2$ (z trzech fermionów można zbudować tylko fermion o spinie połówkowym, a z dwóch jedynie bozon o spinie całkowitym) i podlegają zakazowi Pauliego, podobnie jak bariony i leptony. W przyrodzie obserwujemy jedynie układy związane trzech kwarków lub par kwark-antykwark. Inne układy, podobnie jak i same kwarki, nie zostały dotychczas znalezione. Liczba obserwowanych kwarkowych układów związanych, hadronów, zależy oczywiście od liczby samych kwarków, dla której nie znamy żadnego ograniczenia teoretycznego. Liczba kwarków może więc być dowolna, choć doświadczalnie dowiedzieliśmy się o istnieniu jedynie kilku pierwszych, najlżejszych. Znamy z pewnością cztery rodzaje kwarków, a ostatnio odkryto hadrony, w których składzie znajduje się prawdopodobnie kwark piąty. Zapomnijmy na chwilę o spinach kwarków. Wtedy z czterech kwarków możemy zbudować $4\times4\times4 = 64$ bariony oraz $4\times4 = 16$ mezonów. Po dodaniu spinów i uwzględnieniu istnienia stanów wzbudzonych liczba ta znacznie się zwiększa. Teoria kwarków jest więc bardzo oszczędna, pomimo że liczba odkrytych kwarków wciąż wzrasta.

symbole kwarków (zapachy)

Cztery kwarki, których istnienie zostało mocno ugruntowane doświadczalnie, mają następujące symbole: *u* (ang. *up* 'górny'), *d* (ang. *down* 'dolny'), *s* (ang. *strange* 'dziwny') oraz *c* (ang. *charmed* 'powabny'). W klasyfikacji hadronów i ich oddziaływań stosuje się wciąż nieco inną symbolikę mającą dziś już raczej historyczne znaczenie. Zamiast mówić, że proton składa się z dwóch kwarków *u* oraz jednego *d*, powiada się, że proton ma liczbę barionową $B = 1$, trzecią składową izospinu $I_3 = +^1/_2$, dziwność $S = 0$ oraz powab $C = 0$. W tej terminologii np. kwark *u* nosi następujące wyróżniające go symbole: $B = ^1/_3$, $I_3 = +^1/_2$, $S = 0$, $C = 0$ przy czym wartości liczb B, I_3, S i C dla układu złożonego z kwarków równają się sumom tych liczb dla poszczególnych kwarków. Prościej chyba jednak pisać *uud* niż $B = 1$, $I_3 = ^1/_2$, $S = 0$, $C = 0$. Zamieszczona niżej tabela

Addytywne liczby kwantowe kwarków

Rodzaj kwarku	Liczba barionowa B	Trzecia składowa izospinu I_3	Dziwność S	Powab C	Ładunek elektryczny Q
u	1/3	+1/2	0	0	2/3
d	1/3	−1/2	0	0	−1/3
s	1/3	0	−1	0	−1/3
c	1/3	0	0	1	2/3

podaje liczby kwantowe wszystkich czterech kwarków. Antykwarki mają liczby kwantowe przeciwnego znaku. Łatwo się przekonać, że dla wszystkich kwarków spełniony jest związek

$$Q = I_3 + \tfrac{1}{2}(B+S+C).$$

wzór Gell-Manna–Nishijimy

Ze względu na własność addytywności wszystkich występujących w nim liczb kwantowych wzór ten obowiązuje także dla hadronów. W przypadku $C = 0$ jest to znany już od przeszło 25 lat wzór Gell-Manna–Nishijimy. Suma $Y = B+S$ zwana jest hiperładunkiem cząstki.

Znając skład kwarkowy hadronu możemy na podstawie powyższej tabeli łatwo odczytać liczby kwantowe tego hadronu. I tak np.: układ *uud* ($B = 1$, $I_3 = +^1/_2$, $S = 0$, $C = 0$, $Q = 1$) to proton, układ *udd* ($B = 1$, $I_3 = -^1/_2$, $S = 0$, $C = 0$, $Q = 0$) to neutron, układ *uuu* ($B = 1$, $I_3 = ^3/_2$, $S = 0$, $C = 0$, $Q = 2$) to rezonans Δ^{++}, układ *uus* ($B = 1$, $I_3 = 1$, $S = -1$, $C = 0$, $Q = 1$) to hiperon Σ^+, układ $u\bar{d}$ ($B = 0$, $I_3 = +1$, $S = 0$, $C = 0$, $Q = 1$) to mezon π^+, zaś układ $u\bar{s}$ ($B = 0$, $I_3 = ^1/_2$, $S = 1$, $C = 0$, $Q = 1$) to mezon K^+. Wszystko to można odczytać w Tabeli cząstek elementarnych (→ Cząstki elementarne i ich oddziaływania), na którą będziemy się wielokrotnie powoływać. Oczywiście powyższe wartości liczb kwantowych mają nie tylko wymienione stany podstawowe hadronów, ale i zamieszczona w tabeli cząstek cała rodzina ich stanów wzbudzonych (rezonansów).

ładunek elektryczny kwarków

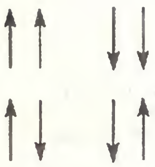
Ostatnią pozycją w tabeli liczb kwantowych kwarków jest ładunek elektryczny. Nie jest on całkowitą wielokrotnością ładunku elementarnego. Ułamkowa wartość ładunku kwarków została potwierdzona w doświadczeniach nieelastycznego rozpraszania elektronów na nukleonach. Oczywiście same hadrony — układy trzykwarkowe lub kwark-antykwark — mają ładunki całkowite.

Przejdźmy teraz do bardziej szczegółowego opisu własności hadronów z punktu widzenia ich składu kwarkowego. Jak się dalej okaże, o własnościach tych decydują różnice mas kwarków. Kwark powabny *c* jest znacznie cięższy od pozostałej trójki. Wyłączymy go więc z naszych rozważań i zajmiemy się hadronami złożonymi z kwarków *u*, *d* i *s*. Hadronów powabnych znamy zresztą stosunkowo niewiele i niewiele też wiemy o obowiązujących dla nich symetriach. Zajmiemy się dalej symetriami hadronów wynikającymi z ich składu kwarkowego.

Symetrie hadronów i ich oddziaływań

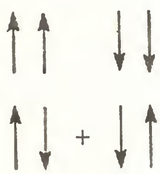
Fundamentalną własnością każdego układu cząstek jest rodzaj symetrii amplitudy falowej (tzw. amplitudy prawdopodobieństwa), której kwadrat określa prawdopodobieństwo znalezienia układu w pewnym stanie. Amplituda układu złożonego np. z dwóch cząstek może być funkcją symetryczną albo antysymetryczną ze względu na operację wzajemnego prze-

stawienia cząstek, tzn. zamiany wszystkich współrzędnych i liczb kwantowych cząstek. Symetria układu może być także mieszana. Jeśli cząstki są identyczne, to obie sytuacje (przed i po zamianie) nie mogą być w żaden sposób rozróżnione, a opisujące je amplitudy różnią się co najwyżej znakiem, przez co nie zmienia się prawdopodobieństwo. Amplituda falowa układu cząstek identycznych nie może więc mieć symetrii mieszanej, musi być albo funkcją symetryczną albo antysymetryczną. Rodzaj symetrii nie może przy tym zależeć od układu odniesienia, nie może zmieniać się przy żadnych transformacjach, jakim poddajemy układ. Jest to dosyć oczywiste, gdyż np. antysymetria amplitudy falowej powoduje, że prawdopodobieństwo znalezienia obu identycznych cząstek w tym samym stanie równa się zeru i fakt ten nie może oczywiście zależeć od tego, z jakiego układu go obserwujemy. Tak właśnie zachowują się wszystkie identyczne fermiony. Ta ich własność jest treścią słynnego zakazu Pauliego. Zagadnienie powyższe możemy nieco uogólnić. Możemy mianowicie rozważać problem przestawienia cząstek prawie identycznych, a także symetrię ze względu na zamianę tylko niektórych liczb kwantowych.



Rys. 6. Cztery możliwe ustawienia spinów dwóch kwarków

Niektóre liczby kwantowe cząstek nie ulegają zmianie przy pewnych przekształceniach układu odniesienia. Taką liczbą jest np. wewnętrzny moment pędu, czyli spin cząstki, którego wartość nie zależy od ustawienia (względem cząstki) układu mierzącego ten spin. Przy obrotach zmienia się jedynie ustawienie spinu, jego rzut na pewną oś, ale nie wartość spinu. Zupełnie podobny do poprzedniego argument przekonuje nas teraz, że cząstka elementarna złożona z dwóch innych cząstek niekoniecznie identycznych ale obdarzonych takim samym spinem musi być opisywana symetryczną albo antysymetryczną amplitudą falową ze względu na przestawienie zmiennych spinowych. Przy czym określonej cząstce złożonej przypisana jest odpowiednia symetria (spin) w sposób jednoznaczny.



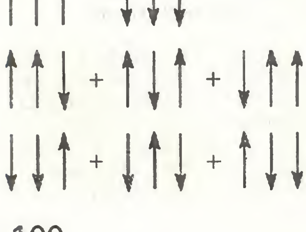
Rys. 7. Trzy symetryczne kombinacje ustawień spinów opisują cząstkę o spinie 1

Kwarki mają spin $^1/_2$. Rzut spinu na dowolną oś wyróżnioną przez warunki zewnętrzne (np. na kierunek zewnętrznego pola magnetycznego, za pomocą którego analizujemy spiny cząstek) przyjmuje jedynie wartości $\pm ^1/_2$ zgodnie z zasadami mechaniki kwantowej. Dla pary kwark-antykwark możliwe są więc 4 ustawienia spinów składników (rys. 6). Dwie pierwsze sytuacje, w których rzut spinu sumarycznego jest równy ± 1, są oczywiście symetryczne ze względu na przestawienie rzutów spinu kwarków. Pozostałe dwie sytuacje o sumarycznym rzucie równym zeru mogą tworzyć zarówno kombinację symetryczną, jak i antysymetryczną. Ostatecznie, trzy symetryczne ustawienia dają cząstkę złożoną (mezon) o spinie 1 (rys. 7), podczas gdy jedna kombinacja antysymetryczna odpowiada cząstce o spinie 0 (rys. 8). Z tego wynika, że stany podstawowe układu kwark-antykwark to mezony o spinie 0 (zwane pseudoskalarnymi) oraz mezony o spinie 1 (zwane wektorowymi). Pierwsze z nich są opisywane antysymetryczną, a drugie — symetryczną amplitudą falową. Podobną sytuację mamy dla układów trzykwarkowych (barionów). Tu jednak amplituda falowa może być np. symetryczna dla pary kwarków i antysymetryczna ze względu na rzut spinu kwarku trzeciego. Całkowita antysymetria trzech rzutów spinu o dwóch tylko możliwych wartościach nie daje się zrealizować i pozostają jedynie dwie sy-

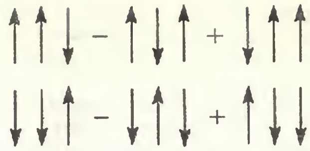


Rys. 8. Jedyna kombinacja antysymetryczna opisuje cząstkę o spinie 0

mezony pseudoskalarne i wektorowe



Rys. 9. Cztery całkowicie symetryczne kombinacje ustawień spinów opisują cząstkę o spinie $^3/_2$



Rys. 10. Dwie kombinacje o symetrii złożonej opisują cząstkę o spinie $^1/_2$

tuacje przedstawione na rys. 9 i 10. W jednej z nich są cztery możliwości opisywane amplitudą całkowicie symetryczną i przedstawiające cząstkę o spinie $^3/_2$. W drugiej amplituda ma symetrię złożoną (ale określoną, nie mieszaną) i przedstawia dwa możliwe stany cząstki o spinie $^1/_2$. Stany podstawowe układów trzykwarkowych to bariony o spinie $^1/_2$ oraz $^3/_2$.

bariony o spinie 1/2 i 3/2

Wyciągamy stąd wniosek, że zależnie od ustawienia spinów kwarków określony skład kwarkowy może odpowiadać różnym hadronom. I tak np. układ *uud* to albo proton (spin $^1/_2$) albo rezonans Δ^+ (spin $^3/_2$), układ $u\bar{d}$ to albo mezon π^+ (spin 0) albo mezon ϱ^+ (spin 1) itd. Masy tych cząstek są oczywiście różne, gdyż energia wiązania kwarków zależy od wzajemnego ustawienia ich spinów.

symetrie przybliżone

Przejdziemy teraz do symetrii przybliżonych obowiązujących wśród hadronów. Wiążą się one z pewnymi podobieństwami różnych kwarków. Kwarki mogą się różnić nie tylko ustawieniem spinu, ale również masami oraz ładunkami elektrycznymi. To, że kwarki mają różne ładunki, jest przyczyną zróżnicowania własności elektromagnetycznych hadronów. Nie ma to jednak istotnego znaczenia przy badaniu oddziaływań silnych. W oddziaływaniach tych ładunki kwarków i hadronów nie odgrywają istotnej roli i niemal o wszystkim decydują różnice mas kwarków.

Wiadomo, że proton (*uud*) oraz neutron (*udd*) mają prawie jednakowe masy. Wnioskujemy stąd, że i kwarki *u* oraz *d* niewiele różnią się masami. Różne ładunki powodują, że nie są to cząstki identyczne (mezon $\pi^+ - u\bar{d}$ to wcale nie to samo, co mezon $\pi^- - d\bar{u}$), ale w oddziaływaniach silnych nie ma to większego znaczenia. Oddziaływania silne powinny być więc symetryczne ze względu na wymianę kwarków *u* oraz *d*. Symetrię tę nazwano wiele lat temu niezmienniczością ze względu na obroty w tzw. przestrzeni izotopowej, lub też symetrią $SU(2)$. Oznacza to po prostu niezmienniczość ze względu na wspomnianą zamianę kwarków. Zwróćmy uwagę, że symetria ta jest w oczywisty sposób naruszona przez oddziaływania elektromagnetyczne hadronów.

symetria izotopowa

Jakie są konsekwencje niezmienniczości izotopowej oddziaływań silnych? Weźmy pod uwagę dwie reakcje rozpraszania: $\pi^+p \to \pi^+p$ oraz $\pi^-n \to \pi^-n$. Stan π^+p ($u\bar{d}uud$) przechodzi po wymianie $u \rightleftarrows d$ w stan π^-n ($d\bar{u}udd$). Stąd przekroje czynne obu reakcji powinny być jednakowe. I takie są rzeczywiście. Podobnie jak przekroje czynne reakcji $\pi^+n \to \pi^+n$ oraz $\pi^-p \to \pi^-p$ i wielu innych. Symetria izotopowa została sprawdzona dla wielu reakcji. Znacznie bardziej spektakularne są jednak przewidywania odnoszące się do mas hadronów tłumaczące występowanie tzw. multipletów izotopowych. Jeżeli bowiem kwarki *u* oraz *d* mają w przybliżeniu takie same masy, to zbliżone masy powinny mieć całe grupy, czyli multiplety hadronów, np.: mezony π ($\pi^+ - u\bar{d}$, $\pi^- - d\bar{u}$, π^0 — antysymetryczna kombinacja $d\bar{d}$ oraz $u\bar{u}$), mezony K ($K^+ - u\bar{s}$, $K^0 - d\bar{s}$, $\bar{K}^0 - s\bar{d}$, $K^- - s\bar{u}$), rezonanse Δ (Δ^{++} — *uuu*, Δ^+ — *uud*, Δ^0 — *udd*, Δ^- — *ddd*), hiperony Σ (Σ^+ — *uus*, Σ^0 — *uds*, Σ^- — *dds*), hiperony Ξ (Ξ^0 — *uss*, Ξ^- — *dss*). Przybliżona równość mas składników multipletów jest doskonale potwierdzona przez dane doświadczalne (Tabela cząstek elementarnych, str. 84). Zachodzi ona oczywiście także dla innych stanów spinowych (np. mezony ϱ, K^*, czy też rezonanse $\Sigma(1385)$) oraz stanów wzbudzonych cząstek.

Symetrię izotopową $SU(2)$ doskonale potwierdzają wszystkie dane doświadczalne. Natomiast nieco ogólniejsza symetria — symetria unitarna $SU(3)$ —

symetria *SU*(3)

nie jest już tak dokładnie spełniona. Byłaby ona symetrią ścisłą oddziaływań silnych, gdyby również kwark s miał taką samą masę, jak kwarki u i d. Mezon K^+ $(u\bar{s})$ ma jednak masę znacznie większą niż mezon π^+ $(u\bar{d})$ i kwark s musi być wyraźnie cięższy. Ze składu kwarkowego cząstek możemy jednak i w tym wypadku wyciągnąć pewne ciekawe wnioski.

Zwróciliśmy już uwagę na fakt, że oddziaływania silne wszystkich kwarków są takie same, a obserwowane różnice mają rzec można, naturę kinematyczną i wywołane są różnicami mas kwarków. Stąd, jeżeli istnieje cząstka, której amplituda falowa, traktowana jako amplituda układu kwarków, ma określoną symetrię, to musi też istnieć inna cząstka o tej samej symetrii, ale innym składzie kwarkowym.

Zajmijmy się najpierw barionami. Układy uuu (Δ^{++}), $ddd(\Delta^-)$ oraz sss (Ω^-) muszą mieć spiny $^3/_2$ i muszą być opisywane amplitudami falowymi całkowicie symetrycznymi ze względu na przestawienie cząstek. Może się bowiem zdarzyć sytuacja, w której trzy jednakowe kwarki mają identyczne rzuty spinów. Sytuację tę można zrealizować przez odpowiedni obrót układu współrzędnych. W wyniku tego wszystkie trzy identyczne kwarki znajdą się w tych samych stanach i muszą być opisywane całkowicie symetryczną amplitudą. Stąd też muszą istnieć następujące symetryczne kombinacje układów trzykwarkowych o spinie $^3/_2$: uud (Δ^+), udd (Δ^0), uus (rezonans Σ^+), uds (rezonans Σ^0), dds (rezonans Σ^-), uss (rezonans Ξ^0) oraz dss (rezonans Ξ^-). Razem z poprzednimi mamy więc dziesięć barionów o spinie $^3/_2$, tzw. dekuplet barionowy. Dekuplet ten składa się z kwartetu izospinowego rezonansów Δ, trypletu rezonansów Σ, dubletu rezonansów Ξ i singletu Ω^-. Biorąc pod uwagę, że każdy kolejny multiplet zawiera o jeden kwark s więcej, łatwo wyciągamy wniosek, że różnice mas między sąsiednimi multipletami są takie same:

dekuplet barionowy

$$m(\Omega^-)-m(\Xi) = m(\Xi)-m(\Sigma) = m(\Sigma)-m(\Delta),$$

co możemy sprawdzić w tabeli cząstek. Przy wyprowadzaniu tej reguły dla mas korzystaliśmy z faktu, że wszystkie spiny kwarków w dekuplecie mogą być ustawione w tę samą stronę i dlatego nie odgrywa roli ewentualna zależność energii wiązania od wzajemnego ustawienia spinów (masa cząstki nie zależy od rzutu jej spinu). Nie odnosi się to do ośmiu barionów tworzących tzw. oktet barionowy o spinie $^1/_2$ i dlatego reguła dla mas nie jest tutaj tak prosta. Nie będziemy jej wyprowadzać, wymienimy jedynie składniki oktetu barionowego. Nie są one oczywiście opisywane całkowicie symetrycznymi amplitudami falowymi, te bowiem odnoszą się do cząstek dekupletu. Nie ma więc hadronów o spinie $^1/_2$ i składzie uuu, ddd oraz sss. Pozostają nam następujące kombinacje: uud (proton), ddu (neutron), uus (hiperon Σ^+), dds (hiperon Σ^-), ssu (hiperon Ξ^0), ssd (hiperon Ξ^+) oraz układ złożony z trzech różnych kwarków uds. W sześciu pierwszych cząstkach dwa spośród kwarków (umieszczone na pierwszym miejscu) są takie same, mogą mieć spiny skierowane w tę samą stronę (wtedy trzeci kwark ma spin skierowany w stronę przeciwną) i muszą być opisywane przez symetryczną amplitudę falową. Stąd i układ uds musi mieć tę samą symetrię i może istnieć w dwóch stanach: symetrycznym ze względu na parę ud (hiperon Σ^0 uzupełniający multiplet Σ) oraz symetrycznym ze względu na parę us lub ds (hiperon Λ). Ostatecznie oktet barionowy składa się z następujących multipletów izospinowych: dubletu nukleonów, singletu Λ, trypletu Σ oraz dubletu Ξ. Z tego, co powiedzieliśmy wyżej wynika, że wszystkie bariony (również stany wzbudzone układów trzykwarkowych) mogą istnieć jedynie w oktetach i dekupletach i że zawsze odpowiednie reguły dla mas są takie same. Fakt ten jest obecnie mocno potwierdzony przez doświadczenie.

oktet barionowy

Znacznie prościej niż z barionami przedstawia się sprawa z mezonami. Komplikują ją jedynie mezony neutralne złożone z par $u\bar{u}$, $d\bar{d}$ oraz $s\bar{s}$. Układy takie mogą bowiem swobodnie przechodzić jeden w drugi (np. $u\bar{u} \to d\bar{d}$ lub $u\bar{u} \to s\bar{s}$). Przejścia $u\bar{u} \to d\bar{d}$ między stanami o bardzo zbliżonych masach mają jedynie znaczenie przy wypisywaniu jawnej postaci amplitud falowych mezonów. Ważniejsze są przejścia $s\bar{s} \to u\bar{u}$ lub $d\bar{d}$, dzięki którym mezon złożony z pary kwarków dziwnych może przechodzić w mezon złożony z pary kwarków niedziwnych i na odwrót. Ustala się jakaś równowaga, w której stanami mezonowymi o określonej masie nie są już układy $s\bar{s}$, $u\bar{u}$, czy też $d\bar{d}$, ale pewne ich kombinacje o składzie zależnym od intensywności procesu przejścia $s\bar{s} \to d\bar{d}$ lub $u\bar{u}$. Okazuje się, że dla układów o spinie 1 przejścia te są mało prawdopodobne i dlatego własności cząstek o spinie 1 (mezonów wektorowych) są prostsze. W przypadku mezonów pseudoskalarnych o spinie 0 powyższe przejścia są istotne i odpowiednie reguły masowe dosyć złożone. Mezonów pseudoskalarnych jest oczywiście dziewięć (3×3): tryplet π $(u\bar{d}$, $d\bar{u}$, $d\bar{d}-u\bar{u})$, dwa dublety $K(u\bar{s}$, $d\bar{s}$ oraz $s\bar{u}$, $s\bar{d})$ oraz dwa singlety η i η' złożone z pewnych kombinacji par $u\bar{u}$, $d\bar{d}$ oraz $s\bar{s}$. Podobnie dziewięć jest mezonów wektorowych (tzw. nonet mezonowy): tryplet ϱ $(u\bar{d}$, $d\bar{u}$, $u\bar{u}+d\bar{d})$, dwa dublety $K^*(u\bar{s}$, $d\bar{s}$ oraz $s\bar{u}$, $s\bar{d})$ oraz dwa singlety: $\omega(u\bar{u}+d\bar{d})$ i $\varphi(s\bar{s})$. W tym ostatnim wypadku możemy łatwo wyprowadzić odpowiednie zależności między masami. Po pierwsze, masa ϱ jest oczywiście równa masie ω:

nonety mezonowe

$$m(\varrho) = m(\omega).$$

Po drugie, porównując liczby kwarków dziwnych i niedziwnych w różnych mezonach dochodzimy do wniosku, że podwojona masa mezonu $K^*(u\bar{s}\ u\bar{s})$ powinna być równa sumie mas mezonu φ oraz mezonu $\varrho(s\bar{s}\ u\bar{d})$:

$$2m(K^*) = m(\varphi)+m(\varrho).$$

Reguły te zupełnie dobrze zgadzają się dla mas wyznaczonych doświadczalnie. Dodajmy wreszcie, że wszystkie mezony (także stany wzbudzone) muszą również grupować się w podobne dziewiątki (3×3). Masy i inne liczby kwantowe wszystkich znanych cząstek (oczywiście bez cząstek powabnych, na które jednak łatwo uogólnić powyższe rozważania) dają się wyjaśnić przez założenie, że wszystkie mezony grupują się w nonety, a wszystkie bariony w oktety i dekuplety.

Kolory i zapachy kwarków

Na początku artykułu stwierdziliśmy, że podstawą opisu oddziaływań cząstek jest kwantowa teoria pola. Powiedzieliśmy już chyba wystarczająco wiele o przewidywaniach i sukcesach hipotezy kwarkowej. Spróbujmy więc teraz zbudować kwantową teorię pola kwarkowego. Na wstępie zauważmy, że w dotychczasowych rozważaniach popełnialiśmy stale bardzo poważny błąd. Z jednej strony twierdziliśmy, że układy identyczne fermionów muszą być opisywane antysymetrycznymi amplitudami falowymi (stąd zakaz Pauliego). Równocześnie jednak okazało się, że barionowe układy trójkwarkowe muszą być opisywane aplitudami symetrycznymi. Np. stan podstawowy identycznych kwarków uuu jest symetryczny ze względu na ich przestawienia. I nie może być inny. Ponieważ kwarki są fermionami, więc układ taki nie mógłby po prostu istnieć. Zauważmy, że gdy wszystkie spiny skierowane są np. w górę, wszystkie trzy kwarki u znajdują się w tym samym stanie, co nie mogłoby się zdarzyć, gdyby kwarki te były rzeczywiście identyczne. Natrafiamy więc na sprzeczność, którą możemy usunąć jedynie przez wprowadzenie pewnej dodatkowej cechy, która rozróżniałaby te trzy kwarki. Wtedy zamiast układu uuu mielibyśmy układ $u_1u_2u_3$,

warunki symetrii

który już może być całkowicie antysymetryczny ze względu na wskaźniki 1, 2, 3 oraz symetryczny ze względu na pozostałe liczby kwantowe. Owe wskaźniki nazwano kolorami kwarków (rodzaje kwarków u, d, s, c nazywa się często ich zapachami) przez analogię do własności światła białego, które może z jednej strony składać się z trzech barw podstawowych (czerwonej, zielonej i niebieskiej), ale również dowolnej barwy oraz barwy do niej dopełniającej (zwykle nie jest to barwa czysta). Jak zobaczymy, hadrony są zupełnie podobnie zbudowane z kwarków kolorowych. Mówimy często, że hadrony są bezbarwne.

kolory kwarków

Wprowadzamy więc dla wszystkich kwarków trzy kolory: czerwony (c), zielony (z) oraz niebieski (n). Aby jednak nie podważyć dotychczasowych rozważań, owo potrojenie liczby kwarków nie może spowodować wzrostu liczby hadronów. Kwarki różnego koloru, lecz tego samego rodzaju (zapachu) nie mogą więc różnić się ani masą, ani ładunkiem. Poza tym jedynie pojedyncza symetria amplitudy falowej ze względu na przestawienia kolorów, może być realizowana w naturze. W przeciwnym razie mielibyśmy tyle cząstek, ile rodzajów symetrii. I wreszcie po to, by owa symetria „kolorowa" nie ulegała zmianie podczas oddziaływań hadronów, musimy mieć ścisłe prawo zachowania każdego koloru w każdym procesie. Oznacza to, że np. liczba kolorów czerwonych minus liczba kolorów antyczerwonych (antybarwy czyli barwy dopełniające mają oczywiście antykwarki) nie zmienia się w żadnej reakcji. Przy spełnieniu tych wszystkich warunków poprzednio opisane wyniki teorii kwarkowej nie ulegają żadnym zmianom. Musimy tylko odpowiedzieć na pytanie, jaka to symetria wskaźników kolorowych jest realizowana w przyrodzie w postaci hadronów. W przypadku barionów odpowiedź już znamy. Opisująca je amplituda falowa jest całkowicie antysymetryczna ze względu na wszystkie kolory. Na przykład

prawo zachowania koloru

$$uuu \rightarrow u_c u_z u_n - u_c u_n u_z + u_n u_c u_z + \\ - u_n u_z u_c + u_z u_n u_c - u_z u_c u_n .$$

Zwróciliśmy poprzednio uwagę na fakt, że trzykwarkowe układy złożone z różnych kombinacji kwarków u, d i s mogą tworzyć całkowicie symetryczne dekuplety bądź oktety o symetrii złożonej. Nie rozważaliśmy tam możliwego pojedynczego układu całkowicie antysymetrycznego: singletu $SU(3)$ uds (inne układy nie mogą być całkowicie antysymetryczne), bowiem ma on stosunkowo dużą masę i nie został dotychczas zidentyfikowany doświadczalnie. Podobnie teraz, w przypadku trzech kolorów, wyżej skonstruowany całkowicie antysymetryczny układ jest singletem kolorowym (symetria kolorowa jest również pewnego rodzaju symetrią $SU(3)$).

Spodziewamy się więc, że i mezony będą singletami kolorowymi. Jeżeli chcemy, by zachowanie koloru było ścisłym prawem przyrody i równocześnie żeby bezbarwny hadron (singlet kolorowy), np. nukleon, mógł wysyłać mezon, np. π, pozostając nadal bezbarwnym hadronem (proces ten jest źródłem występowania sił jądrowych), to mezony muszą też być bezbarwne. Podobnie jak z trzech kolorów i trzech kwarków można utworzyć tylko jeden stan całkowicie antysymetryczny (automatycznie musi to być bezbarwny singlet nie wyróżniający żadnego koloru), tak w wypadku par kolor-antykolor można utworzyć tylko jeden stan nie wyróżniający żadnego koloru, symetryczny ze względu na zamianę dowolnego koloru (wraz z odpowiadającym mu antykolorem) na inny. Na przykład

singlety kolorowe

$$u\bar{d} \rightarrow u_c \bar{d}_{\bar{c}} + u_z \bar{d}_{\bar{z}} + u_n \bar{d}_{\bar{n}} .$$

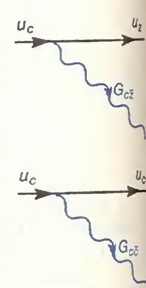
Taka kombinacja jest więc znów singletem kolorowym i w ten właśnie sposób zbudowane są wszystkie mezony.

Zauważmy też, że singletu kolorowego nie można zbudować ani z dwóch, ani z czterech czy pięciu kwarków (o trzech możliwych kolorach), ani też z kwarku i pary kwark-antykwark itp. Możemy łatwo sprawdzić, że w każdym z tych wypadków dowolnego typu symetria kolorowa amplitudy falowej albo nie może być w ogóle zrealizowana, albo ma więcej niż jedną realizację (nie odpowiada wtedy singletowi o jednej składowej). Pojedynczy kwark jest oczywiście trypletem kolorowym. Widzimy więc, że jedynie układy trzykwarkowe oraz kwark-antykwark (a także ich wielokrotności) mogą tworzyć singlety kolorowe, co prowadzi do reguły: tylko singlety kolorowe są obserwowane w postaci hadronów.

Chromodynamika kwantowa

Przypisanie kwarkom trzech różnych kolorów nie rozwiązuje problemu. Aby kwarki różniące się kolorem były naprawdę różne, musi istnieć jakieś oddziaływanie, które rozróżniałoby kolory. Oddziaływanie to musi być przenoszone przez jakieś cząstki. Cząstki te, nośniki oddziaływań międzykwarkowych, nazywa się gluonami (ang. *glue* 'klej'). Na wzór elektrodynamiki kwantowej buduje się teorię kwarkowo-gluonową zwaną chromodynamiką kwantową. Gluony, podobnie jak fotony, mają w niej zerowe masy i spin równy jedności. Oczywiście, że gluony nie zostały, jak dotychczas, wyprodukowane w żadnym doświadczeniu. Mają wprawdzie masę zero, ale dzięki ich uwięzieniu wewnątrz hadronów zasięg oddziaływań silnych jest skończony (→ Oddziaływania elektromagnetyczne). Zbadajmy, jakie kolory mają gluony. W tym celu musimy rozważyć elementarny diagram produkcji gluonu z kwarku o określonym kolorze (rys. 11). Na przykład kwark czerwony wysyła gluon i zmienia (lub nie) swoją barwę. Ponieważ obowiązuje ścisłe prawo zachowania każdej barwy, więc wysyłany gluon musi mieć dwa kolory: kolor kwarku przed aktem wysłania i antykolor (kolor dopełniający) kwarku końcowego. Każdy więc gluon ma dwa kolory, z czego łatwo możemy obliczyć całkowitą liczbę gluonów. Zwróćmy jednak uwagę na fakt, że wartości prawdopodobieństwa wysłania gluonu z kwarków różnego koloru muszą być takie same. W przeciwnym razie wymiana gluonów nie tylko powodowałaby zmianę kolorów kwarków, ale także dla różnych kolorów prowadziłaby do różnej energii oddziaływania, co musiałoby wywołać wystąpienie różnic mas w tryplecie kolorowym kwarków.

gluony



Rys. 11. Wysłanie gluonu może zmienić lub nie zmienić koloru kwarku. Nigdy jednak nie zmienia jego rodzaju (zapachu)

Tak więc wysłanie gluonu z każdego kwarku wiąże się z tą samą stałą sprzężenia, która wynosi prawdopodobnie $\alpha_c \approx 0,5$.

Ile więc jest gluonów mających kolor i antykolor? Ponieważ kolory są trzy, więc wydawałoby się, że gluonów musi być dziewięć. Jednak trzy układy kolorów, $c\bar{c}$, $z\bar{z}$, $n\bar{n}$ nie zmieniają koloru kwarku. Nie oznacza to, że gluony mające takie pary kolorów nie rozróżniają wcale kolorów kwarków, gdyż np. gluon $c\bar{c}$ może być wysłany tylko z kwarku czerwonego. Jedynie znana nam już całkowicie symetryczna kombinacja kolorów $c\bar{c}+z\bar{z}+n\bar{n}$ (singlet kolorowy) nie wyróżnia żadnego koloru kwarków. Wysłanie takiego gluonu nie zmieniłoby nawet względnych znaków amplitud falowych odpowiadających kwarkom o różnych kolorach. Taki bezbarwny gluon musi być wyeliminowany z teorii, w której obowiązuje reguła, że wszystkie singlety kolorowe nie są uwięzione w hadronach, mogą być więc obserwowane w doświadczeniach i wymieniane w reakcjach między hadronami. Teoria minimalna zawiera więc osiem gluonów o masie zero i spinie 1. Dodajmy jeszcze, że gluony nie mają ładunku elektrycznego i nie biorą udziału w oddziaływaniach słabych.

Podstawą chromodynamiki kwantowej zbudowanej na wzór elektrodynamiki kwantowej są elementarne procesy wysyłania i pochłaniania gluonów przez kwarki oraz kreacji par kwark-antykwark przez gluo-

kwark fizyczny

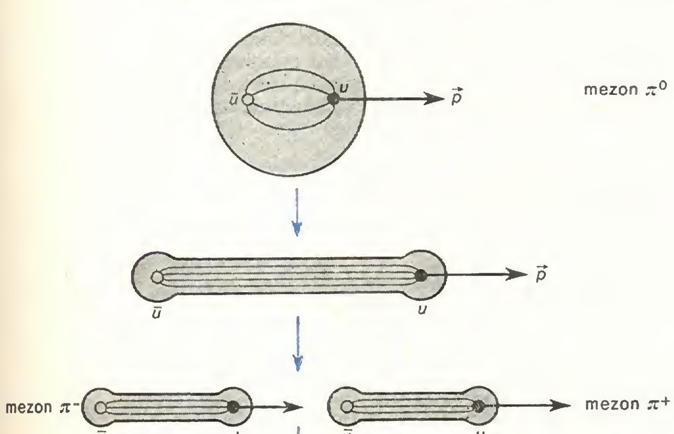
ny a także anihilacji tych par z wytworzeniem gluonów. Teoria ta różni się od elektrodynamiki jedynie liczbą nośników sił — gluonów oraz większą stałą sprzężenia, $\alpha_c \approx {}^1/_2$. Kwark fizyczny, podobnie jak poprzednio elektron, składa się, z prawdopodobieństwem bliskim jedności, z kwarku pozbawionego struktury, z prawdopodobieństwem rzędu $(\alpha_c)^2 \approx {}^1/_4$ z kwarku i gluonu, z prawdopodobieństwem rzędu $(\alpha_c)^4 \approx {}^1/_{16}$ z kwarku oraz pary kwark-antykwark itd. Podobnie proton składa się z trzech punktowych kwarków *uud* z prawdopodobieństwem niewiele różnym od jedności. To wszystko jest jednak prawdziwe jedynie w odniesieniu do wewnętrznego obszaru protonu, położonego blisko środka, i ten właśnie obszar był badany w doświadczeniach nieelastycznego rozpraszania elektronów na nukleonach. W obszarze zewnętrznym na kwarki i gluony działają ogromne siły nie pozwalające im na opuszczenie hadronów. Tak więc dla dużych odległości między kwarkami stała sprzężenia α_c z niewyjaśnionych do końca powodów silnie wzrasta i nie możemy już klasyfikować składu hadronów mówiąc o trzech kwarkach pozbawionych struktury itp. W zewnętrznych regionach hadronu kwarki mają bogatą, trudną do opisania strukturę. Jednak przedstawiony poprzednio kwarkowy obraz hadronów nadal obowiązuje, choć przeważnie nie są to kwarki bez struktury.

Przeprowadzona analiza własności i symetrii hadronów nie zależała oczywiście od tego, czy kwarki wewnątrz hadronów miały jakąś strukturę, czy też były punktowe. Ważne tam były tylko masy, rodzaje oraz liczba kwarków.

Zastanówmy się jeszcze, jak chromodynamika kwantowa opisuje reakcje produkcji dodatkowych hadronów (głównie mezonów) w procesach ich rozpraszania. Będzie to analiza czysto jakościowa, ponieważ zarówno mechanizm uwięzienia kwarków, jak i własności kwarków znajdujących się w pobliżu brzegów hadronu nie dają się łatwo opisać przy użyciu metod rachunkowych kwantowej teorii pola. Są to obszary teorii wkraczające w dziedzinę, którą można nazwać chemią oddziaływań silnych i jako takie mogą być traktowane jedynie fenomenologicznie.

Fenomenologiczny opis oddziaływań silnych

Wyobraźmy sobie, że para kwarków jest uwięziona w obszarze mezonowym o określonej z grubsza objętości. Zderzając ten mezon z innym hadronem lub też elektronem możemy jednemu z kwarków nadać duży pęd względem kwarku drugiego (rys. 12). Obszar



Rys. 12. Produkcja dwóch mezonów π^- i π^+ z mezonu π^0 następuje w wyniku kreacji dodatkowej pary $d\bar{d}$

mezonowy zaczyna się rozciągać i słabe, rozproszone początkowo statyczne pole gluonowe (o bardzo podobnych własnościach do fotonowego pola elektrostatycznego) zaczyna się wydłużać i przybiera postać pola jednorodnego. Na kwarki zaczyna wtedy działać stała siła, a energia obszaru mezonowego zaczyna być proporcjonalna do jego długości. W pewnym momencie energia jest tak duża, że wystarcza na utworzenie nowej pary kwark-antykwark. Na tych nowych cząstkach zamykają się teraz linie sił pola gluonowego i obszar mezonowy rozpada się na dwa nowe mezony. Proces ten powtarza się prowadząc do wytworzenia wielocząstkowego stanu końcowego. W ten sposób w zderzeniach cząstek z hadronami produkowane są inne hadrony. Widzimy, że zawsze polega to na wytworzeniu wzbudzonego stanu mezonowego, a następnie na jego rozpadzie. Własności tego stanu wzbudzonego nie zależą od rodzaju cząstki, która wywołała wzbudzenie (elektron, mezon π, proton), a jedynie od przekazanego kwarkom pędu. Procesy produkcji wielu cząstek powinny więc wyglądać podobnie bez względu na rodzaj zderzających się cząstek. Fakt ten bardzo dobrze zgadza się z wynikami wielu doświadczeń.

wymiana mezonów a wymiana gluonów

A dlaczego mówimy często, że w oddziaływaniach wewnątrzjądrowych i innych oddziaływaniach silnych decydującą rolę odgrywa wymiana mezonów, np. mezonu π, a nie wymiana gluonów? Otóż w krótkich przedziałach czasu nieokreśloność energii (masy) hadronu staje się na tyle duża, że hadron ten może wydłużyć się i pęknąć wysyłając np. wirtualny mezon π, który następnie może zostać pochłonięty przez inny lub ten sam hadron. Z fenomenologicznego więc punktu widzenia oddziaływania silne to wymiana wirtualnych mezonów π i innych hadronów. Podobnie można opisywać fenomenologicznie strukturę np. protonu jako cząstki złożonej częściowo (z pewnym prawdopodobieństwem) z pojedynczego protonu, częściowo z protonu i mezonu π itd. Taki obraz oddziaływań silnych i struktury hadronów mógłby być najbardziej odpowiedni do opisu większości reakcji z udziałem hadronów. W reakcjach tych bowiem przeważnie bierze udział zewnętrzna część hadronów, a wtedy nie można już stosować prostych metod rachunkowych lokalnej chromodynamiki kwantowej. Niestety, prawdopodobieństwo wysłania np. wirtualnego mezonu π z protonu jest duże i związana z tym procesem stała sprzężenia wynosi aż 14,4. Nie można więc stosować takiego rachunku zaburzeń (diagramów Feynmana), jaki stał się podstawą sukcesów elektrodynamiki kwantowej. Pozostają metody znacznie bardziej złożone, do jakich należy np. model wymiany biegunów Reggego (→ Oddziaływania silne), wykorzystanie tzw. związków dyspersyjnych itp. Metody te zupełnie poprawnie opisują wielką ilość danych doświadczalnych dotyczących rozpraszania i produkcji hadronów. Warto jednak pamiętać, że znikoma zaledwie część tych danych ma bezpośredni związek z lokalnymi oddziaływaniami kwarków i gluonów, a te przecież stanowią fundament wszystkich oddziaływań silnych. Podobnie analizując widmo promieniowania jakiejkolwiek cząsteczki chemicznej nie domyślilibyśmy się łatwo, że struktura tej cząsteczki opiera się na jednym prostym prawie Coulomba. Dopiero zbadanie szczególnego rodzaju widma — widma atomu wodoru — wyjaśniłoby szybko naturę owego fundamentalnego rodzaju sił.

chromodynamika lokalna a nielokalna

Materia jest prawie zawsze bardzo złożona i odkrycie najprostszych rządzących nią zasad wymaga przeprowadzenia bardzo szczególnych doświadczeń. Są to zwykle doświadczenia zupełnie dla danej dziedziny nietypowe. Takimi właśnie były doświadczenia nieelastycznego rozpraszania elektronów na nukleonach, dzięki którym dowiedzieliśmy się, że kwarki nie mają struktury.

G. Białkowski, R. Sosnowski *Cząstki elementarne*, Warszawa 1971; L.N. Cooper *Istota i struktura fizyki*, Warszawa 1975; R.P. Feynman i in. *Feynmana wykłady z fizyki*, t. 3, Warszawa 1974.

Atomy egzotyczne

Janusz Zakrzewski

Atom zwyczajny o liczbie atomowej Z składa się z jądra o ładunku dodatnim $+Zq_e$, wiążącego za pośrednictwem oddziaływania elektromagnetycznego (przyciągających sił kulombowskich) Z ujemnie naładowanych elektronów, z których każdy ma ładunek elementarny $-q_e$ ($|q_e| = 1{,}6 \cdot 10^{-19}$C). Jako całość atom jest więc elektrycznie obojętny. Składnikami zwyczajnego jądra są nukleony: Z dodatnio naładowanych protonów, każdy o ładunku $+q_e$, oraz $(A-Z)$ obojętnych elektrycznie neutronów (A jest liczbą masową jądra oznaczonego symbolem ^A_ZX).

Najprostszym atomem zwyczajnym jest atom wodoru, składający się z protonu, deuteronu lub trytonu jako jądra i jednego elektronu powłokowego. Jeżeli w takim atomie jądro zostanie zastąpione cząstką elementarną, naładowaną dodatnio, np. pozytonem e^+ lub mionem (lepton μ^+), powstanie układ związany cząstki dodatniej i ujemnego elektronu, zwany odpowiednio pozytonium (e^+e^-) lub mionium (μ^+e^-). Jeżeli natomiast w atomie zwyczajnym o liczbie atomowej Z jeden z elektronów powłokowych e^- zostanie zastąpiony cząstką elementarną naładowaną ujemnie, np. mionem (lepton μ^-), pionem (mezon π^-), kaonem (mezon K^-), antyprotonem $\bar{p}$ lub hiperonem Σ^-, powstanie odpowiednio atom mionowy, pionowy, kaonowy, antyprotonowy lub hiperonowy (→ Cząstki elementarne i ich oddziaływania). Składa się on z jądra o liczbie atomowej Z, cząstki ujemnej (o masie o wiele większej niż elektron) związanej w polu kulombowskim jądra oraz $(Z-1)$ elektronów powłokowych. Takie atomy będziemy zwali ogólnie atomami egzotycznymi; stosuje się też często ogólną nazwę mezoatomy lub atomy mezonowe (i to nie tylko w tych wypadkach, gdy cząstką związaną jest mezon π^- czy K^-, lecz i wtedy, gdy jest nią lepton μ^- lub barion — antyproton czy hiperon Σ^-).

atomy mionowe, pionowe, kaonowe, antyprotonowe i hiperonowe

Wszystkie wymienione atomy egzotyczne zostały zaobserwowane w doświadczeniach. Pierwszej obserwacji egzotycznych atomów mionowych dokonano w 1947 r. i w tym samym roku pojawiły się pierwsze analizy teoretyczne tego zjawiska. Atomy egzotyczne stanowią do chwili obecnej przedmiot intensywnych badań eksperymentalnych i teoretycznych, dostarczających informacji o własnościach cząstek elementarnych, ich oddziaływaniach z nukleonami oraz o strukturze jąder atomowych.

Ponieważ atomy mionowe i pionowe bada się przeszło 25 lat, w dalszym ciągu omówimy jedynie najnowsze wyniki doświadczeń nad atomami hadronowymi (hadroatomami) — kaonowymi, antyprotonowymi i hiperonowymi. Badania w tej dziedzinie rozwinęły się dopiero w latach siedemdziesiątych, gdy w kilku laboratoriach dysponujących akceleratorami protonowymi wielkich energii (Argonne, Berkeley, Brookhaven, Genewa) wytworzono intensywne wiązki cząstek ujemnych o małej energii. Umożliwiło to zastosowanie liczników półprzewodnikowych, germanowych i krzemowych, do badania mezoatomów zawierających mezony K^-, hiperony Σ^- i antyprotony.

Procesy elektromagnetyczne

Omówimy wpierw zjawiska wspólne dla wszystkich atomów egzotycznych. Historię życia cząstki naładowanej, która utworzyła atom egzotyczny w ośrodku materialnym, do momentu jej zniknięcia w końcowym procesie oddziaływania można podzielić na trzy etapy: spowolnienie, wychwyt atomowy i deekscytację utworzonego atomu egzotycznego. Wszystkie trzy procesy wywołane są oddziaływaniem elektromagnetycznym. Końcowy proces oddziaływania — rozpad lub absorpcja jądrowa — zależy od rodzaju cząstki.

spowolnienie

Poruszając się w ośrodku materialnym, cząstka naładowana ujemnie, traci swą energię kinetyczną w oddziaływaniu elektromagnetycznym z atomami ośrodka na wzbudzanie ich i jonizację. Czas potrzebny na spowolnienie cząstki od typowej energii ok. 100 MeV do ok. 2 keV wynosi w przybliżeniu 10^{-10}–10^{-9} s. Gdy energia cząstki staje się dostatecznie mała, cząstka może ulec schwytaniu przez pole kulombowskie jądra atomu ośrodka i zastąpić w atomie jeden z jego elektronów powłokowych, który ulegnie wyrzuceniu z atomu w tzw. procesie Augera:

wychwyt atomowy

cząstka + atom → atom egzotyczny + elektron.

Powstaje atom egzotyczny w stanie wysoce wzbudzonym. Podstawowe własności tego atomu można opisać poglądowo posługując się najprostszym modelem atomu — modelem Bohra, z którego wynika, że energia cząstki ujemnej związanej w polu kulombowskim jądra punktowego o liczbie atomowej Z (wpływ pozostałych elektronów powłokowych jest tu pominięty) może mieć wartości

model Bohra atomu egzotycznego

$$E_n = -\frac{\mu c^2}{2}\left(\frac{\alpha Z}{n}\right)^2, \qquad (1)$$

promień zaś orbity kołowej, odpowiadającej stanowi o głównej liczbie kwantowej n, wynosi

$$r_n = \frac{\hbar}{\mu c}\,\frac{n^2}{\alpha Z}, \qquad (2)$$

gdzie $\mu = m/(1+m/m_A)$ jest masą zredukowaną cząstki o masie spoczynkowej m i jądra o masie spoczynkowej m_A, $\alpha \approx 1/137$ jest stałą struktury subtelnej, $c = 3 \cdot 10^8$ m/s — prędkością światła w próżni, a $\hbar = h/2\pi = 1{,}05 \cdot 10^{-34}$ J·s — stałą Plancka. Taki sam wzór na energię stanów stacjonarnych otrzymuje się z nierelatywistycznego równania falowego — równania Schrödingera (→ Spektroskopia atomowa, Chemia kwantowa).

Ze wzorów (1) i (2) otrzymujemy następujące związki między energią E_n oraz promieniem bohrowskim r_n cząstki w atomie egzotycznym i odpowiednimi wielkościami atomu zwyczajnego w stanie o tej samej liczbie kwantowej n:

atom egzotyczny a zwyczajny

$$E_n = E_{ne}(\mu/\mu_e), \qquad r_n = r_{ne}(\mu_e/\mu),$$

gdzie symbole ze wskaźnikiem e odnoszą się do elektronu w atomie zwyczajnym. Zatem w atomie egzotycznym poziomy energii leżą o wiele niżej (mają większą wartość bezwzględną), a promienie bohrowskie są o wiele mniejsze niż w atomie zwyczajnym (nawet dla cząstki o najmniejszej masie — mionu — wartość $\mu/\mu_e \approx 207$). Tylko w stanie o dużych wartościach głównej liczby kwantowej n jądro jest osłaniane przez elektrony powłokowe. Przy mniejszych wartościach n wszystkie elektrony powłokowe są znacznie dalej od jądra, tak że cząstkę w atomie egzotycznym można

Własności cząstek i tworzonych przez nie atomów egzotycznych (wg modelu Bohra)

Cząstka	Masa m MeV	m/m_e	Średni czas życia τ s	Energia Bohra $E_B = E_0\left(\frac{m}{m_e}\right)Z^2$ keV	Promień Bohra $r_B = \frac{m_e}{m}\cdot\frac{a_0}{Z}$ fm	Główna l. kw. $n = \sqrt{m/m_e}$
μ^-	105,659	207	$2{,}2 \cdot 10^{-6}$	$2{,}8 \cdot Z^2$	$256/Z$	14
π^-	139,567	273	$2{,}6 \cdot 10^{-8}$	$3{,}7 \cdot Z^2$	$194/Z$	17
K^-	493,669	966	$1{,}2 \cdot 10^{-8}$	$13{,}1 \cdot Z^2$	$54{,}7/Z$	31
$\bar{p}$	938,279	1836	—	$24{,}9 \cdot Z^2$	$28{,}8/Z$	43
Σ^-	1197,34	2343	$1{,}5 \cdot 10^{-10}$	$31{,}8 \cdot Z^2$	$22{,}6/Z$	48

Wartości obliczone przy założeniu nieskończenie dużej masy jądra; $E_0 = 13{,}6$ eV, $a_0 = 5{,}3 \cdot 10^4$ fm są wartościami energii i promienia Bohra zwyczajnego atomu wodoru (wg N. Barash-Schmidta i in., 1980).

traktować z dobrym przybliżeniem jak elektron w atomie wodoropodobnym (o jednym elektronie). Stan cząstki na orbicie bohrowskiej o promieniu r_n równym promieniowi r_{1e} elektronu o najniższej energii E_{1e} ($n_e = 1$) jest nadal stanem wysoce wzbudzonym: odpowiada mu główna liczba kwantowa $n = \sqrt{(\mu/\mu_e)}$. Wartości tej liczby n oraz energii wiązania $E_B = -E_1$ i promieni bohrowskich $r_B = r_1$ cząstek w stanach podstawowych różnych atomów egzotycznych podane są w tabeli wraz z masami m i średnimi czasami życia τ tych cząstek — składników odpowiednich atomów.

ścisłe obliczenie energii

Przy analizie danych doświadczalnych energie stanów w atomie egzotycznym należy obliczać dokładniej, niż na to pozwala model Bohra. Stosuje się zatem relatywistyczne równanie falowe: Kleina–Gordona dla cząstki o spinie 0 (bozonu — np. pionu, kaonu) lub Diraca dla cząstki o spinie $^1/_2$ (fermionu — np. mionu, antyprotonu, hiperonu Σ^-). Przy założeniu, że jądro jest punktowe, wzór na energię wynikający z tych równań ma postać:

$$E_{n,j} = -\frac{\mu c^2}{2}\left(\frac{\alpha Z}{n}\right)^2\cdot\left[1+\left(\frac{\alpha Z}{n}\right)^2\left(\frac{n}{j+{}^1/_2}-\frac{3}{4}\right)-\ldots\right], \qquad (3)$$

gdzie dla bozonu $j = l$, dla fermionu zaś $j = l \pm {}^1/_2$, przy czym l jest orbitalną liczbą kwantową. (Wzór (1) stanowi przybliżenie wzoru (3), jeśli się w nawiasie kwadratowym pominie mniejsze od 1 wyrazy zawierające czynnik αZ). Widać, że dla fermionów poziomy energii $E_{n,j}$ są rozszczepione wskutek istnienia własnego momentu magnetycznego cząstki, tworzą dublety struktury subtelnej (tej samej wartości n odpowiadają dwie wartości j). Wyrażenie na energię należy uzupełnić dodając do wzoru (3) poprawki wynikające z uwzględnienia skończonych rozmiarów jądra (rozkładu ładunku w jądrze), poprawki promieniste (wynikające przede wszystkim z efektu polaryzacji próżni) oraz poprawki uwzględniające osłanianie elektronów powłokowych (do pominięcia przy małych wartościach n). Energię E_{em} stanu w atomie egzotycznym z uwzględnieniem wszystkich poprawek elektromagnetycznych ΔE_{em} można ostatecznie zapisać w postaci:

$$E_{em} = E_{KG,D} + \Delta E_{em}, \qquad (4)$$

gdzie członem dominującym ($E_{KG,D}$) jest energia wynikająca z równania Kleina–Gordona lub Diraca.

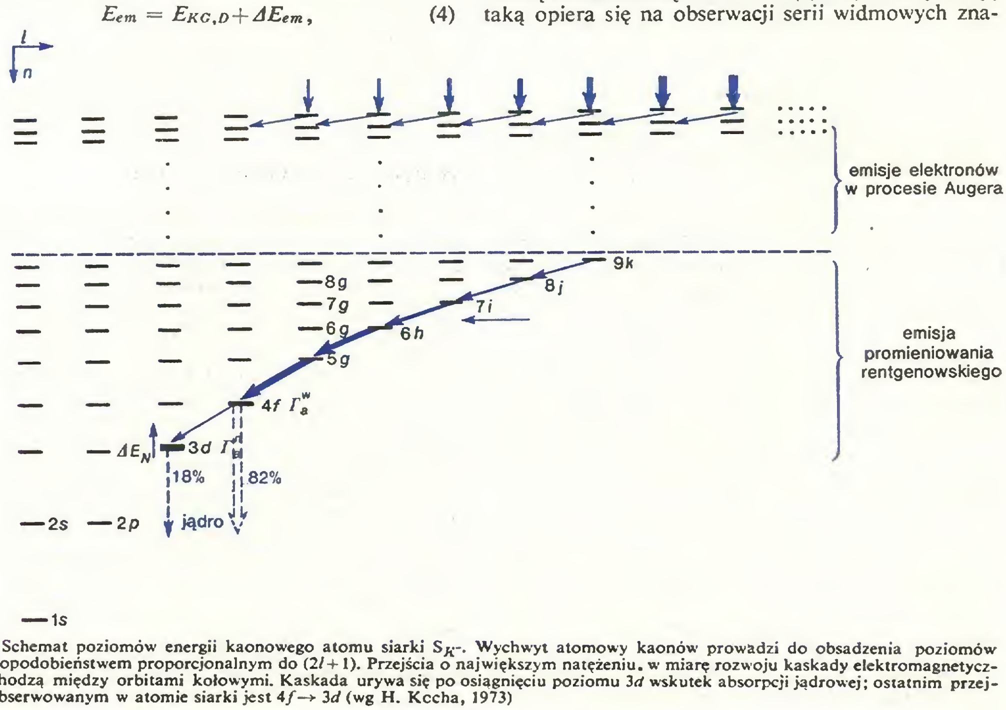
deekscytacja

Poziomy energii w atomie egzotycznym można wyznaczać badając promieniowanie wysyłane w procesie deekscytacji powstałego atomu wzbudzonego, tzn. przy przejściach elektromagnetycznych między poziomami. W procesie deekscytacji atom pozbywa się energii wzbudzenia, równej różnicy energii stanu początkowego p i końcowego k: $E^p - E^k = \hbar\omega$; pozbywa się jej w postaci energii kinetycznej elektronu powłokowego, wyrzuconego z atomu w procesie Augera, lub promieniowania rentgenowskiego o częstości ω. Proces pierwszy dominuje we wczesnych etapach rozwijającej się w atomie egzotycznym kaskady elektromagnetycznej, drugi — w stanach o mniejszych wartościach głównej liczby kwantowej n, jak to ilustruje schemat poziomów energii dla kaonowego atomu siarki (rys. 1). Reguły wyboru rządzące tymi procesami, $\Delta l = \pm 1$ (z uprzywilejowaniem $\Delta l = -1$ i z dopuszczalnie małą zmianą Δn dla procesu Augera, a dużą dla promieniowania rentgenowskiego) powodują, że cząstka zmierza do stanu o maksymalnej wartości orbitalnej liczby kwantowej $l = n-1$ (do „orbity kołowej" — w języku prostego modelu Bohra). Z chwilą osiągnięcia stanu o liczbach kwantowych (n, $l = n-1$) zachodzą tylko przejścia elektromagnetyczne typu $(n, n-1) \to (n-1, n-2) \to (n-2, n-3) \to \ldots$, tzn. przejścia między kolejnymi orbitami kołowymi. Całość procesów elektromagnetycznych, sprowadzających cząstkę do najniższego stanu, w którym następuje końcowy proces oddziaływania, trwa w sumie 10^{-12}–10^{-10} s.

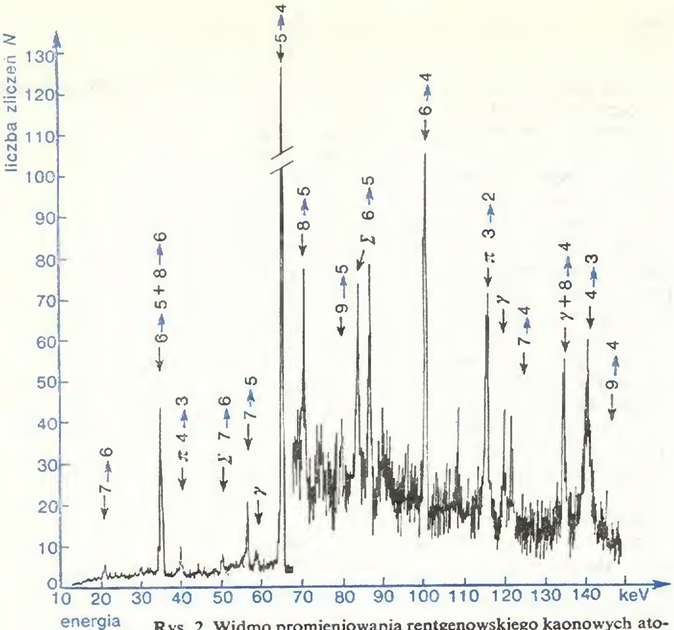
Badanie własności cząstek elementarnych

widmo rentgenowskie

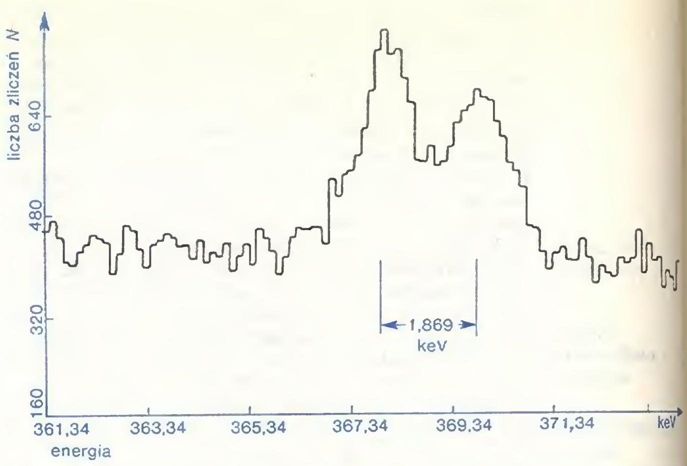
Informacje o utworzeniu i deekscytacji atomu egzotycznego pochodzą przede wszystkim z obserwacji promieniowania rentgenowskiego. Pomiary energii przeprowadza się przy użyciu detektorów półprzewodnikowych o rozdzielczości energetycznej $\Delta E/E \leqslant 10^{-2}$ w zakresie energii ok. 100 keV. Identyfikacji przejść elektromagnetycznych zachodzących z emisją promieniowania rentgenowskiego dokonuje się na podstawie porównania precyzyjnie zmierzonych wartości energii kwantu $\hbar\omega$ z energią przejścia $E^p_{em} - E^k_{em}$, obliczoną z dużą dokładnością ze wzorów (3) i (4). Identyfikację taką opiera się na obserwacji serii widmowych zna-



Rys. 1. Schemat poziomów energii kaonowego atomu siarki S_{K^-}. Wychwyt atomowy kaonów prowadzi do obsadzenia poziomów z prawdopodobieństwem proporcjonalnym do $(2l+1)$. Przejścia o największym natężeniu, w miarę rozwoju kaskady elektromagnetycznej, zachodzą między orbitami kołowymi. Kaskada urywa się po osiągnięciu poziomu $3d$ wskutek absorpcji jądrowej; ostatnim przejściem obserwowanym w atomie siarki jest $4f \to 3d$ (wg H. Kocha, 1973)



Rys. 2. Widmo promieniowania rentgenowskiego kaonowych atomów fosforu $_{15}P_{K^-}$. Linie $7 \to 6$, $6 \to 5$, $5 \to 4$, $4 \to 3$ oraz $9 \to 4$, $8 \to 4$, $7 \to 4$, $6 \to 4$ (przejście do stanu o $n_k = 4$) są liniami odpowiadającymi przejściom kaonowym. Widać też linie widma atomów hiperonowych (Σ $7 \to 6$ i Σ $6 \to 5$) i linie widma atomów pionowych (π $4 \to 3$, π $3 \to 2$) oraz linie wzorcowania γ. Linia $4 \to 3$ wykazuje poszerzenie spowodowane oddziaływaniem silnym (wg G. Backenstossa i in., 1974)



Rys. 3. Przejście $11 \to 10$ w antyprotonowych atomach uranu wykazujące rozszczepienie struktury subtelnej spowodowane przez moment magnetyczny antyprotonu (wg L. Tauschera, 1975)

nych ze zwyczajnej spektroskopii rentgenowskiej. Rysunek 2 przedstawia widmo promieniowania rentgenowskiego kaonowych atomów fosforu $_{15}P_{K^-}$ — z zaznaczonymi przejściami ($n_p - n_k$) między stanami o głównych liczbach kwantowych stanu początkowego n_p i końcowego n_k.

pomiar masy

Pomiar energii przejścia atomu egzotycznego umożliwia wyznaczenie masy cząstki schwytanej przez taki atom (masa cząstki występuje we wzorze na energię przejścia w atomie egzotycznym). Jedne z najdokładniejszych pomiarów mas hadronów ujemnych pochodzą właśnie z wyznaczenia energii atomów egzotycznych. Z precyzyjnych pomiarów energii przejścia między stanami o dostatecznie dużych wartościach liczby n w kaonowych atomach baru Ba_{K^-} ($5 < n < 11$) i złota Au_{K^-}($6 < n < 12$) oraz ołowiu Pb_{K^-}($7 < n < 13$) wyznaczono w dwóch niezależnych doświadczeniach, wykonanych w latach siedemdziesiątych masę mezonu K^- i otrzymano wartości zgodne w granicach błędu; ich wartość średnia wynosiła

$$\overline{m}_{K^-} = 493{,}667 \pm 0{,}017 \text{ MeV},$$

z błędem względnym $3{,}5 \cdot 10^{-5}$. Wyznaczono wówczas również masę hiperonu Σ^- na podstawie obserwacji promieniowania rentgenowskiego hiperonowych atomów ołowiu Pb_{Σ^-}($9 < n < 16$) i otrzymano wartość

$$m_{\Sigma^-} = 1197{,}24 \pm 0{,}14 \text{ MeV},$$

zgodną z wynikami pomiarów za pomocą emulsji jądrowej $m_{\Sigma^-} = 1197{,}24 \pm 0{,}23$ MeV oraz z wynikiem uzyskanym przy użyciu komory pęcherzykowej $m_{\Sigma^-} = 1197{,}43 \pm 0{,}08$ MeV. Dokonano też pomiaru masy antyprotonu posługując się antyprotonowymi atomami ołowiu $Pb_{\bar{p}}$ ($9 < n < 17$):

$$m_{\bar{p}} = 938{,}179 \pm 0{,}058 \text{ MeV}.$$

Odejmując tę wartość od masy protonu otrzymuje się różnicę mas protonu i antyprotonu, $m_p - m_{\bar{p}} = 100 \pm 58$ keV, równą zeru w granicach błędu, zgodnie z przewidywaniami elektrodynamiki kwantowej.

pomiar momentów magnetycznych

Rozszczepienie struktury subtelnej linii promieniowania rentgenowskiego jest w pierwszym przybliżeniu proporcjonalne do momentu magnetycznego cząstki związanej w atomie egzotycznym. Na tej podstawie zmierzono moment magnetyczny antyprotonu w dwóch niezależnych eksperymentach, w których badano przejścia $11 \to 10$ w antyprotonowych atomach ołowiu $Pb_{\bar{p}}$ i uranu $U_{\bar{p}}$ (rys. 3). Uzyskano wartości:

$$\mu_{\bar{p}} = -2{,}819 \pm 0{,}056 \mu_N$$

oraz

$$\mu_{\bar{p}} = -2{,}790 \pm 0{,}021 \mu_N,$$

zgodne ze sobą oraz bliskie wartości momentu magnetycznego protonu $\mu_p = 2{,}7928456 \mu_N$ (jednostką momentu magnetycznego w fizyce jądrowej jest magneton jądrowy $\mu_N = \hbar q_e / 2m_p = 5{,}05 \cdot 10^{-27}$ J·T^{-1}).

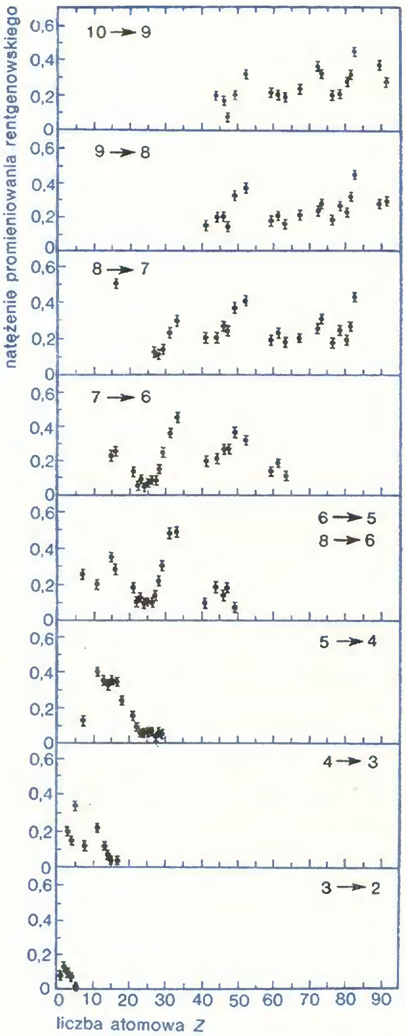
Pomiar momentu magnetycznego hiperonu Σ^- jest znacznie trudniejszy ze względu na małe natężenie promieniowania rentgenowskiego atomów hiperonowych i duże tło pochodzące z innych procesów. Około 8–10% aktów oddziaływania mezonów K^- z jądrami tarczy prowadzi do powstania hiperonów Σ^-. Tylko te hiperony, które przeżyją procesy spowalniania, mogą ulec schwytaniu w polu kulombowskim jądra-tarczy i utworzyć atomy hiperonowe. W tym samym widmie pojawiają się więc linie promieniowania atomów kaonowych i pionowych stanowiące tło (zob. rys. 2). Z powyższych względów wyznaczenie momentu magnetycznego hiperonu Σ^- jest mniej pewne i opiera się na subtelnej analizie statystycznej widma.

Wpływ oddziaływania silnego

Na elektromagnetyczne stany związane hadronu w atomie egzotycznym wpływa oddziaływanie silne hadronu z jądrem. Gdy hadron jest związany w stanie o dużej wartości liczby kwantowej n, tzn. w dużej odległości od jądra, wpływ tego oddziaływania ze względu na jego krótki zasięg jest pomijalnie mały (z tego względu pomiary mas i momentów magnetycznych hadronów wykonuje się dla przejść między stanami o dużych wartościach liczby n). Jednakże w miarę deekscytacji promienistej atomu egzotycznego cząstka osiąga stan o mniejszej wartości n, w którym jej orbita bohrowską wnika w obszar jądra. W takim stanie wpływ oddziaływania silnego zaczyna odgrywać istotną rolę i powoduje odstępstwa od schematu elektromagnetycznych poziomów energii (zob. wzór 4). Wpływ ten można zaobserwować za pośrednictwem trzech zjawisk: przesunięcia poziomu energii ΔE_N, jego poszerzenia Γ_a oraz zmniejszenia wydajności Y przejść z emisją promieniowania rentgenowskiego wskutek absorpcji jądrowej, stanowiącej najbardziej prawdopodobny końcowy proces oddziaływania hadronu.

absorpcja jądrowa

Absorpcja jądrowa staje się dominująca po przejściu do następnego, niższego poziomu, o liczbie kwantowej $n-1$; hadron ulega unicestwieniu w oddziaływaniu silnym, o czym świadczy urywanie się serii widmowej. Widać to z rys. 4, przedstawiającego zależność wydajności przejścia rentgenowskiego od liczby atomowej Z atomów kaonowych, zmierzonej w zakresie od $Z = 2$ do $Z = 92$. Rysunek ten ilustruje najbardziej charakterystyczną cechę atomów hadronowych świadczącą o jądrowej absorpcji hadronów — zmniejszanie się natężenia określonego przejścia $(n, n-1) \to (n-1, n-2)$ w miarę zwiększania się liczby atomowej pierwiastka, z ostatecznym zniknięciem tego



Rys. 4. Wydajność przejść rentgenowskich między orbitami kołowymi w atomach kaonowych. Kończenie się danego przejścia przy określonej wartości liczby atomowej Z stanowi najbardziej charakterystyczną cechę hadronowych atomów egzotycznych (wg E. Wieganda i G. L. Godfreya, 1974)

przejścia przy pewnej wartości Z. Zmiany wydajności przejścia, w tym minima przy wartościach Z ok. 28, 38, 60 i 75, wiążą się z procesem atomowego wychwytu kaonu i wynikającym stąd obsadzeniem początkowych stanów w atomie egzotycznym.

Przesunięcie i poszerzenie poziomów energii

Wskutek oddziaływania silnego poziom energii ulega przesunięciu ΔE_N, które się definiuje jako różnicę między wartością energii zmierzonej E_{eksp} i obliczonej E_{obl}, z uwzględnieniem wszelkich poprawek elektromagnetycznych (zob. wzór 4): $\Delta E_N = E_{eksp} - E_{obl}$. Szerokość poziomu Γ_a, spowodowana oddziaływaniem silnym, jest związana z szybkością W_a absorpcji jądrowej z tego poziomu: $\Gamma_a = \hbar W_a$. Można je wyznaczyć bezpośrednio, obserwując poszerzenie linii promieniowania rentgenowskiego, co widać na przykładzie przejścia $4 \to 3$ w atomie P_{K^-} (rys. 2). Otrzymuje się w ten sposób szerokość Γ_a^n $(n-1,\ l = n-2)$ poziomu niższego, z którego hadron jest absorbowany z prawdopodobieństwem 100% (zob. rys. 1); jest ona o 2-3 rzędy wielkości większa od szerokości Γ_a^w $(n,\ l = n-1)$ poziomu wyższego. $\Gamma_a^w = \Gamma_x({}^P/_Y - 1) - \Gamma_A$ można obliczyć mierząc wydajność Y przejść z emisją promieniowania rentgenowskiego. Jest ona związaną z prawdopodobieństwem obsadzenia poziomu wyższego, które albo oblicza się uwzględniając rozwój kaskady elektromagnetycznej w atomie egzotycznym albo wyznacza się doświadczalnie, mierząc wydajność wszystkich przejść elektromagnetycznych zasilających ten poziom. W tabeli przytoczono niektóre dane dla atomów kaonowych, ilustrujące wartości omawianych wyżej wielkości. Tego rodzaju dane stanowią ważny sprawdzian każdej teorii usiłującej opisać oddziaływanie silne hadronu z jądrem.

Atomy kaonowe

Pierwiastek	Przejście	ΔE_N, eV	Γ_a^n, eV	Γ_a^w, eV
$^{10}_{5}B$	3→2	−208±35	810±100	–
$^{11}_{5}B$	3→2	−167±35	700±80	–
$^{12}_{6}C$	3→2	−590±80	1730±150	0,98+0,19
$^{31}_{15}P$	4→3	−330±80	1440±120	1,94±0,33

Procesy absorpcji jądrowej

absorpcja K

Absorpcja jądrowa hadronu w atomie egzotycznym zachodzi w elementarnym procesie silnego oddziaływania hadronu z nukleonami. Dla mezonów K^- dominuje proces oddziaływania jednonukleonowego z produkcją mezonu π (oddziaływanie mezonowe):

$$K^- + N \to \Sigma \text{ (lub } \Lambda) + \pi$$
$$\hookrightarrow \Sigma + N \to \Lambda + N \text{ (w materii jądrowej),}$$

gdzie N oznacza nukleon (proton lub neutron), a Σ i Λ — odpowiednie hiperony (w stanie końcowym muszą się pojawiać hiperony Σ lub Λ ze względu na zasadę zachowania dziwności w silnym oddziaływaniu). W 20–30% przypadków mezon K^- jest absorbowany w oddziaływaniu dwunukleonowym (wielonukleonowym), bez produkcji mezonu π (oddziaływanie niemezonowe):

$$K^- + 2N \to \Sigma \text{ (lub } \Lambda) + N$$
$$\hookrightarrow \Sigma + N \to \Lambda + N \text{ w (materii jądrowej).}$$

W procesie takim bierze udział skorelowana para nukleonów, może więc on służyć do badania korelacji nukleonowych w jądrach atomowych.

absorpcja Σ^-

Hiperony Σ^- ulegają absorpcji w oddziaływaniu silnym z protonami,

$$\Sigma^- + p \to \Lambda + n,$$

co prowadzi, podobnie jak w wypadku kaonów, do powstania hiperonów Λ jako produktu końcowego. Wskutek przyciągającego oddziaływania silnego hiperonu Λ z nukleonami hiperon Λ może zostać związany w materii jądrowej i utworzyć hiperjądro (→ Hiperjądra). Absorpcja jądrowa antyprotonów zachodzi w procesie anihilacji z produkcją wielu mezonów, głównie pionów:

absorpcja $\bar{p}$

$$\bar{p} + N \to \pi + \pi + \pi.$$

potencjał oddziaływania kaonu z jądrem

We wszystkich dyskutowanych wyżej wypadkach wpływ oddziaływań silnych uwzględnia się przez wprowadzenie do falowego równania jądrowego potencjału optycznego, opisującego oddziaływanie hadronu z jądrem (→ Siły jądrowe). Najprostszą postać dla kaonów ma potencjał proporcjonalny do gęstości materii jądrowej $\varrho(r)$:

$$V(r) = -\frac{2\pi}{\mu}\left(1+\frac{m_{K^-}}{m_N}\right)\bar{A}\varrho(r),$$

gdzie m_{K^-} i m_N oznaczają masy mezonu K^- i nukleonu, μ — masę zredukowaną układu kaon-jądro, $\bar{A}$ zaś jest parametrem zespolonym. Przyjmuje się przy tym, dla uproszczenia, że rozkład neutronów i protonów jest taki sam $\varrho_n(r) = \varrho_p(r) = \varrho(r)$ i że jest równy gęstości ładunku w jądrze. Opis oddziaływania kaonów z jądrem jest jednakże utrudniony wskutek występowania w układzie kaon-proton podprogowego stanu rezonansowego Λ (1405) (→ Cząstki elementarne i ich oddziaływania). Prawdopodobnie z tego względu prosta ekstrapolacja danych dotyczących rozpraszania kaonów na nukleonach, przy niskich energiach, do obszaru energii odpowiadającego kaonom związanym w atomach egzotycznch nie prowadzi do wyników zgodnych z doświadczeniem.

sondowanie jądra

Poprawny opis oddziaływania kaon-nukleon w atomie egzotycznym umożliwiłby wykorzystanie K^- (a także hiperonów Σ^- i antyprotonów) jako cząstek sondujących strukturę jądrową. Ponieważ mezony K^- absorbowane są głównie w obszarze powierzchniowym, gdzie gęstość materii jądrowej nie przekracza ok. 10% wartości centralnej, można by użyć ich do badania powierzchni jądra, w tym — rozkładu neutronów i protonów oraz korelacji nukleonowych na powierzchni jądra.

Absorpcja jądrowa mezonów — pionów i kaonów — często prowadzi do emisji ugrupowań nukleonowych, w szczególności takich jak cząstki α. Informacje o tym zjawisku uzyskuje się bądź bezpośrednio, obserwując końcowe fragmenty jądrowe, bądź pośrednio, wykrywając fotony γ emitowane z jąder końcowych. Rysunek 5 przedstawia wychwyt niemezonowy kaonu przez jądro węgla, zarejestrowany w emulsji jądrowej. Zaobserwowany proces dowodzi, że mezon K^- może oddziaływać z dwoma nukleonami (lub ich większą liczbą) w procesie niemezonowym, powodując rozpad jądra końcowego na cząstki α. Częstą emisję cząstek α w oddziaływaniach mezonów można przypisywać albo własnościom jądra, albo mechanizmowi reakcji, np. wychwytowi wielonukleonowemu.

Dalsze badania atomów egzotycznych, a zwłaszcza badanie koincydencji między promieniowaniem rentgenowskim a cząstkami wtórnymi, powstałymi w wyniku absorpcji hadronu przez jądro, przyniosą niewątpliwie odpowiedź na wiele problemów otwartych, jeszcze nie rozstrzygniętych w dotychczasowych doświadczeniach.

oddziaływanie mezonu K^- z jądrem ^{12}C



Rys. 5. Mikrofotografia oddziaływania w emulsji jądrowej mezonu K^- z jądrem węgla ^{12}C (punkt *a*) z produkcją hiperonu Σ^- (*1*), protonu (*2*), dwóch cząstek α (*3*, *4*) i deuteronu (*5*). Cząstki α pochodzą prawdopodobnie z rozpadu jądra końcowego (^{8}Be w stanie podstawowym). Hiperon Σ^- w punkcie *b* rozpada się w locie na mezon π^- (wg D. Evansa i in., 1961)

G. Backenstoss Ann.Rev.Nuclear Sci. 20, 467 (1970); G. Backenstoss i J. Zakrzewski Contemp. Phys. 15, 197 (1974); E.H.S. Burhop High-Energy Physics 3, 109 (1969) oraz Contemp. Phys. 11, 335 (1970); S. Devons i I. Duerdoth Advances Nuclear Phys. 2, 295 (1969); R. Seki, C.E. Wiegand *Kaonic and other Exotic Atoms*, w: Ann. Rev. Nucl. Sci. 25, 241 (1975); C.S.Wu i L. Wilets Ann. Rev. Nuclear Sci. 19, 527 (1969).

Detekcja cząstek

Tomasz Hofmokl

proces widzenia

Czy można zobaczyć cząstkę, której rozmiary są mniejsze niż 10^{-15} m, a masa jest rzędu 10^{-30} kg? Na to pytanie odpowiada się często przecząco argumentując tym, że zdolność rozdzielcza układu optycznego nie może być mniejsza od długości fali padającego światła, która dla zakresu promieniowania widzialnego zawiera się w granicach $4-7\cdot10^{-7}$ m, a rozmiary cząstki są o kilka rzędów wielkości mniejsze. Taka odpowiedź zakłada jednak potoczne, bardzo ograniczone znaczenie słowa widzieć. Zastanówmy się nad fizyczną stroną procesu widzenia. Jak przebiega z punktu widzenia fizyka oglądanie czerwonego przedmiotu w słoneczny dzień? Mamy do dyspozycji źródło kwantów promieniowania elektromagnetycznego (fotonów) o energiach $3\text{-}5\cdot10^{-19}$ J (zakres światła widzialnego), co odpowiada w częściej używanych w tej dziedzinie jednostkach 1,8–3,1 eV. Doprowadzamy fotony do zderzenia z tarczą (przedmiot) i rejestrujemy energię i częstości kwantów wpadających do urządzenia pomiarowego (oko). Stwierdzamy, że wśród cząstek dochodzących po zderzeniu do detektora przeważają kwanty o energii 2 eV, a pozostałe cząstki zostały pochłonięte przez tarczę. Wnioskujemy stąd, że tarcza odbija promieniowanie elektromagnetyczne o długości fali $6{,}2\cdot10^{-7}$ m. Potocznie mówimy, że przedmiot jest czerwony. W rzeczywistości to powiedzenie charakteryzuje nie sam przedmiot (w świetle zielonym przedmiot będzie czarny), a proces oddziaływania z przedmiotem fal elektromagnetycznych o określonym rozkładzie długości fali. Oglądanie przedmiotu to badanie oddziaływania z nim światła.

Rozszerzmy to określenie dopuszczając dowolne oddziaływanie czegokolwiek z badanym obiektem. Wtedy obmacywanie po ciemku ściany będzie rów-

oglądanie cząstek

nież jej oglądaniem. Takie rozszerzenie pojęcia widzenia pozwoli nam obejrzeć cząstki elementarne. Postępowanie będzie analogiczne do oglądania przedmiotu. Należy w tym celu dysponować źródłem cząstek, na przykład akceleratorem. Powinny mieć one odpowiednio dużą energię, aby długość fali de Broglie'a była rozmiarów porównywalnych lub mniejszych niż obszar oddziaływania. Przykładem oglądania protonów jest eksperyment, w którym korzystano z wiązki mezonów pi plus (π^+) o energii 16 GeV. Energii tej odpowiada długość fali de Broglie'a $7{,}7 \cdot 10^{-17}$ m, co jest długością znacznie mniejszą niż rozmiary protonu. Zderzano mezony (piony) z tarczą z ciekłego wodoru, a więc praktycznie z protonami i obserwowano cząstki rozproszone.

Oczywiście na podstawie jednej obserwacji nie można wyciągać żadnych wniosków. Ze zderzenia jednego fotonu z przedmiotem też byśmy nic nie powiedzieli o kształcie i kolorze przedmiotu. Należy zbadać wiele, od kilkuset do kilkudziesięciu tysięcy zarejestrowanych zderzeń, aby móc coś powiedzieć o oddziaływaniu.

Nie wchodząc na razie w zagadnienie co chcemy wiedzieć globalnie o badanym oddziaływaniu, możemy od razu wyliczyć co należałoby wiedzieć o każdej cząstce biorącej udział w każdym pojedynczym zdarzeniu, wchodzącej lub opuszczającej obszar oddziaływania. Najczęściej wystarcza znajomość energii i trzech składowych pędu każdej cząstki. Ideałem byłaby znajomość również ustawienia w przestrzeni spinów pojedynczych cząstek (o ile spin jest różny od zera), ale o tym można jeszcze tylko marzyć. W praktyce zagadnienie sprowadza się do wyznaczenia pędu oraz masy lub energii każdej interesującej w danym doświadczeniu cząstki.

Bardzo często musimy się jednak ograniczać ze względów technicznych do niepełnej informacji, zredukowanej niekiedy tylko do sygnału czy cząstka przeszła przez detektor czy też nie.

detektor

Nazwa „detektor" pochodzi od łacińskiego czasownika *detegere* — odkryć, odsłonić, wyjawić. Detektor cząstek pozwala wykryć przejście cząstki i dostarcza różnych informacji umożliwiających identyfikację i określenie stanu jej ruchu. Podstawą detekcji cząstek jest ich elektromagnetyczne oddziaływanie z materią (→ Oddziaływania elektromagnetyczne). Cząstki nie oddziałujące elektromagnetycznie (neutrina) można zarejestrować obserwując produkty wywoływanych przez nie reakcji zdolne do oddziaływań elektromagnetycznych. Celem detekcji jest znalezienie odpowiedzi na następujące pytania: Jaka cząstka została zarejestrowana? Jaka jest energia (pęd) cząstki? Jaka jest lokalizacja toru? W jakiej chwili pojawiła się cząstka?

Rzadko udaje się, używając jednego typu detektora, uzyskać odpowiedź na wszystkie pytania. Toteż buduje się złożone układy detekcyjne. W tabeli przedstawionej poniżej zestawione są zjawiska fizyczne mogące służyć do rejestracji i identyfikacji cząstek oraz stanu ich ruchu.

Podstawowe zjawiska wykorzystywane przy rejestracji cząstek

Rejestracja cząstki polega na zaobserwowaniu skutków oddziaływania cząstki z otaczającą materią. Poza bardzo słabymi oddziaływaniami grawitacyjnymi rozróżnia się oddziaływania słabe, silne i elektromagnetyczne. Te ostatnie są najczęściej wykorzystywane do rejestracji. Omówimy pokrótce te zjawiska fizyczne, które mogą służyć do rejestracji i identyfikacji cząstek (tabela poniżej).

rejestracja cząstki

Ruch naładowanej cząstki w polu elektromagnetycznym

Na poruszającą się naładowaną cząstkę o ładunku e w polu elektromagnetycznym działa siła

$$\vec{F} = e(\vec{E} + \vec{v} \times \vec{B}),$$

gdzie $\vec{E}$ — wektor natężenia pola elektrycznego, $\vec{B}$ — wektor indukcji pola magnetycznego, $\vec{v}$ — wektor prędkości cząstki. Siła pochodząca od pola elektrycznego jest skierowana wzdłuż wektora tego pola, a przyczynek od pola magnetycznego (proporcjonalny do iloczynu wektorowego $\vec{v}$ i $\vec{B}$) jest prostopadły zarówno do prędkości, jak i do wektora indukcji pola. Spostrzeżenie to jest bardzo ważne, albowiem w jednorodnym, stałym polu magnetycznym cząstka porusza się po okręgu (jeżeli nie traci energii po drodze), którego promień krzywizny r zależy od pędu, ładunku i indukcji pola:

$$r = \frac{p \cdot 0{,}3333 \cdot 10^{-2}}{ZB},$$

gdzie r jest wyrażone w m, pęd p w MeV/c, indukcja B w teslach, a ładunek Z w jednostkach ładunku elementarnego.

Znajomość krzywizny toru cząstki w znanym polu magnetycznym pozwala wyznaczyć stosunek pędu do ładunku, co w większości przypadków jest równoznaczne z wyznaczeniem pędu, ponieważ praktycznie zawsze mamy do czynienia z cząstkami o ładunku jednostkowym (wyjątek stanowią oddziaływania z jądrami atomowymi). Wielkość odchylenia w polu elektrycznym prostopadłym do ruchu cząstki zależy od prędkości cząstki, a nie od pędu. Układ, który wykorzystuje oba rodzaje pól, pozwala na wyznaczenie pędu i energii cząstki, a zatem — jej masy.

wyznaczanie pędu

Rozpraszanie kulombowskie na jądrach

Niezależnie od stosowania do identyfikacji cząstek silnych pól zewnętrznych można do tego celu wykorzystać silne pola elektryczne w pobliżu jądra. Naładowana cząstka przechodząca przez materię napotyka

Zestawienie zjawisk lub wielkości pomocnych przy detekcji i identyfikacji cząstek

Wielkość mierzona lub zjawisko	Od czego zależy	Uwagi
Krzywizna toru w polu magnetycznym	pęd/ładunek	bardzo często stosowane w pomiarach
Krzywizna toru w polu elektrycznym	pęd × prędkość/ładunek	ograniczone zastosowanie ze względu na trudności uzyskania dużych pól
Jonizacja i zjawiska pokrewne	ładunek × prędkość	pomiar często wykonywany
Zasięg	ładunek × prędkość	pomiar stosowany do powolnych cząstek
Wielokrotne rozpraszanie	ładunek; pęd × prędkość	pomiar stosowany, gdy brak pola magnetycznego, a ślad jest dokładnie zarejestrowany
Promieniowanie Czerenkowa	ładunek × prędkość	szerokie zastosowanie w licznikach Czerenkowa
Zjawisko Comptona i tworzenie par e^+e^-	energia fotonu	stosowane w pomiarach kwantów γ
Kaskada elektronowo-fotonowa	energia fotonu	stosowane w pomiarach kwantów γ i elektronów
Czas przelotu między dwoma licznikami	prędkość	metoda elektroniczna szeroko stosowana
Zderzenia sprężyste i rozpady cząstek nietrwałych	masa i energia	metoda szczególnie przydatna do badania cząstek neutralnych; stosuje się ją również do naładowanych

obszary o natężeniu pola elektrycznego rzędu 10^{10} V/m. Istnienie tak dużych pól zawdzięczamy ładunkom elektrycznym jądra. Te silne pola elektryczne powodują tzw. rozpraszanie kulombowskie. Przechodząc przez warstwę substancji, cząstka wielokrotnie ulega rozpraszaniu kulombowskiemu. Małe pojedyncze rozproszenia w różnych kierunkach dodają się i obserwujemy wypadkowe średnie wielokrotne rozproszenie. W nieobecności zewnętrznego pola elektromagnetycznego tor cząstki naładowanej powinien być prosty. Jednak wskutek wielokrotnego rozpraszania kulombowskiego odchyla się on od prostej, przy czym średni kwadrat kąta rozproszenia zależy od pędu cząstki i od właściwości ośrodka, przez który cząstka przechodzi. Średni kwadrat kąta rozproszenia $\langle\theta^2\rangle$ (w radianach) można w przybliżeniu wyrazić w zależności od pędu cząstki p (podanego w MeV/c), jej prędkości β (wyrażonej w stosunku do prędkości światła) i stałej X charakteryzującej materiał ośrodka, przez który cząstka przechodzi, zwanej długością radiacyjną (podanej w centymetrach):

długość radiacyjna

$$\langle\theta^2\rangle \approx \left(\frac{21}{p\beta}\right)^2 \Big/ X.$$

W tabeli zestawiono wartości długości radiacyjnej kilku częściej używanych substancji. Z porównania danych widać, że średnie kulombowskie rozproszenie

Minimalne straty energii na jednostkę drogi cząstki w różnych substancjach

Substancja	Gęstość g/cm³	Długość radiacyjna X, cm	$\left(\frac{dE}{dx}\right)_{min}$, $\frac{MeV}{cm}$
Ciekły wodór	0,063	1000	0,26
Propan C_3H_8	0,41	111	0,98
Ksenon	2,3	3,9	2,8
Emulsja jądrowa	3,82	2,94	5,49
Ołów	11,35	0,56	12,8

jest praktycznie pomijalne dla cząstek poruszających się w wodorze (w dokładnych pomiarach uwzględnia się jednak to zjawisko), ale może znacznie zniekształcić tory w ksenonie. Ilustracją są zdjęcia torów z komór wypełnionych wodorem, propanem i ksenonem (tabl. 9, il. 28). W ksenonie wielokrotne kulombowskie rozpraszanie odgrywa tak doniosłą rolę, że praktycznie uniemożliwia stosowanie zewnętrznych pól magnetycznych. Musiałyby one mieć ogromną indukcję, aby dominować nad wielokrotnym rozpraszaniem.

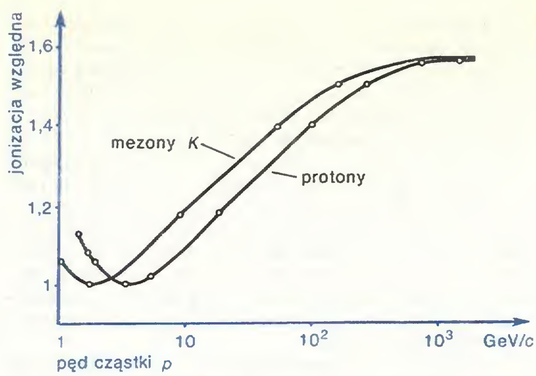
Straty energii na jonizację

Przechodząc przez materię, naładowana cząstka oddaje część swojej energii napotkanym atomom powodując ich wzbudzenie lub jonizację. Dostarczona energia zostaje pochłonięta przez któryś z elektronów powłoki atomowej przeniesiony do wyższego stanu energetycznego. Jeżeli energia ta jest dostatecznie duża, to elektron może się całkowicie uwolnić z atomu. Mówimy wtedy o jonizacji atomu. Przechodząca cząstka pozostawia więc ślad w postaci wzbudzonych lub zjonizowanych atomów. Ilość takich wytrąconych z równowagi atomów jest ważną informacją o przejściu i o samej naturze cząstki.

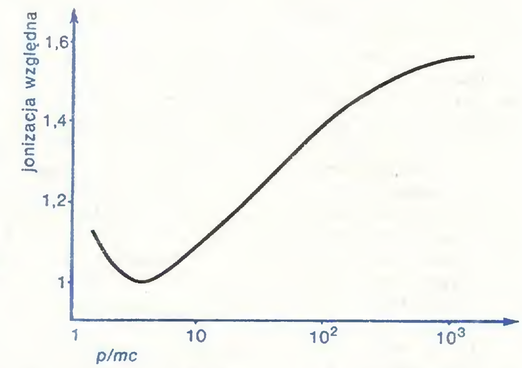
Straty energii na jonizację lub wzbudzenie są spowodowane niesprężystym oddziaływaniem kulombowskim przelatującej cząstki z elektronami. Rysunek 1 ukazuje zależność względnej jonizacji (stosunku straty energii na jednostkę długości do minimalnej możliwej straty energii) od pędu dla mezonów K i protonów. Umiejąc ocenić jonizację cząstki i znając jej pęd, możemy ją zidentyfikować. Rysunek ten można uprościć wykreślając względną jonizację w zależności od pędu podzielonego przez masę cząstki razy prędkość światła (rys. 2). Pokazana na nim krzywa

zależność względnej jonizacji od pędu

względna jonizacja



Rys. 1. Zależność względnej jonizacji gazu używanego w urządzeniu identyfikującym cząstki od pędu protonów i mezonów K



Rys. 2. Zależność względnej jonizacji od zmiennej p/mc. Krzywa jest uniwersalna dla cząstek o dowolnej masie i ładunku jednostkowym

jest uniwersalna, tzn. że można ją stosować do cząstek o dowolnej masie, a ładunku — jednostkowym.

Początkowo krzywa jonizacji względnej maleje bardzo szybko, odwrotnie proporcjonalnie do kwadratu prędkości cząstki, osiąga wartość minimalną (minimum jonizacji) i wzrasta do wartości maksymalnej (maksimum jonizacji), przy której się utrzymuje niezależnie od energii. Obszar malenia jonizacji z energią i obszar wzrostu są wykorzystywane do identyfikacji cząstek.

rozróżnianie cząstek wg zasięgu

Znając stratę energii na jednostkę przebytej drogi, można obliczyć zasięg cząstki, czyli drogę, na której straci całą swoją energię na jonizację i zatrzyma się w ośrodku. Można wykazać, że stosunek zasięgu R do masy cząstki m zależy tylko od stosunku energii cząstki E do jej masy:

$$R/m = f(E/m).$$

Jeżeli więc znana jest krzywa zasięg–energia jakiejkolwiek cząstki w danej substancji, to można łatwo wykreślić również krzywą dla innej cząstki o tym samym ładunku. Dla stosunkowo powolnych cząstek, a tylko takie mają szanse zatrzymać się w urządzeniach detekcyjnych, zasięg R można wyrazić zależnością przybliżoną:

$$R \sim m^{-2,3} p^{3,3}.$$

Tak więc proton o tym samym pędzie co mezon π zatrzyma się po przebyciu drogi prawie 80 razy krótszej niż pion, ponieważ ma 6,7 razy większą masę.

Mechanizm scyntylacji

Atomy wzbudzone pozbywają się energii emitując promieniowanie elektromagnetyczne, czyli po prostu świecą. Średni czas życia stanu wzbudzonego jest

rzędu 10^{-9}–10^{-8} s, a więc świecenie jest praktycznie natychmiastowe. W omawianym wypadku interesuje nas zjawisko fluorescencji lub radioluminescencji, czyli fluorescencji pod działaniem cząstek jonizujących. Substancję, w której można obserwować radioluminescencję, nazywamy scyntylatorem. Dobrymi materiałami scyntylacyjnymi są niektóre gazy, przede wszystkim szlachetne, związki organiczne oraz kryształy nieorganiczne. Niektóre typy scyntylatorów omówimy opisując detektory scyntylacyjne.

radioluminescencja

Zjawisko Czerenkowa

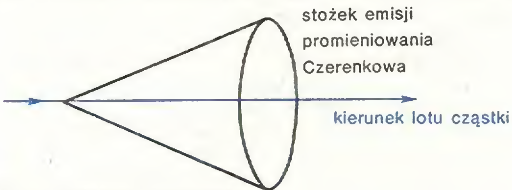
Badając w 1934 r. zjawisko emisji światła przy przechodzeniu przez substancję cząstek jonizujących, Czerenkow zauważył, że istnieje silna korelacja między kierunkiem padania cząstek i kierunkiem emisji światła. Jeżeli cząstka jonizująca porusza się w ośrodku o współczynniku załamania n_ν dla promieniowania o częstości ν, a jej prędkość v jest większa niż prędkość światła w tym ośrodku c/n_ν:

$$v > \frac{c}{n_\nu},$$

czyli

$$\beta = \frac{v}{c} > \frac{1}{n_\nu},$$

to cząstka wysyła promieniowanie elektromagnetyczne w stożku skierowanym w kierunku lotu cząstki



Rys. 3. Zjawisko Czerenkowa

(rys. 3). Fale o różnej częstości będą wysyłane pod różnymi kątami:

$$\cos\theta_\nu = \frac{1}{\beta n_\nu}.$$

Im większa prędkość cząstki padającej, tym większe rozwarcie stożka. Maksymalny kąt rozwarcia

$$\theta_{\nu\,max} = \arccos\frac{1}{n_\nu}.$$

Typowe wartości parametrów charakteryzujących zjawisko Czerenkowa zestawiono w tabeli.

Wartości parametrów charakteryzujących zjawisko Czerenkowa

Parametr	Materiał		
	tworzywo perspex	woda	powietrze
Współczynnik załamania, n	1,5	1,33	1,00029
$\theta_{max} = \arccos 1/n$	48°	41°	1,3°
$\beta_{min} = 1/n$	0,667	0,751	0,9997
Minimalny pęd cząstek odpowiadający β_{min}:			
elektronu	0,46 MeV/c	0,58 MeV/c	21 MeV/c
mezonu π	125 MeV/c	159 MeV/c	5,7 GeV/c
protonu	840 MeV/c	1,067 GeV/c	38 GeV/c

Promieniowanie Czerenkowa pojawia się dopiero wtedy, gdy pęd cząstki przekroczy pewną progową wartość zależną od jej masy. Zjawisko to znalazło szerokie zastosowanie w identyfikacji cząstek.

Promieniowanie hamowania

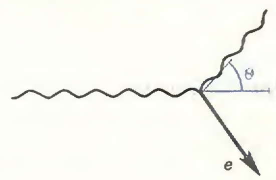
Jeżeli poruszająca się cząstka naładowana doznaje opóźnienia w jakimś polu zewnętrznym (np. jeżeli jest przyhamowana w polu jądra, albo tylko zmienia kierunek), to emituje kwant promieniowania elektromagnetycznego. Promieniowanie to nazywamy promieniowaniem hamowania. Jest ono szczególnie ważne dla elektronów. Jeżeli foton o dużej energii przechodzi przez substancję, to jest duże prawdopodobieństwo, że wytworzy parę e^+e^-. Każdy z elektronów spowalniany w substancji wysyła kwanty promieniowania hamowania, te z kolei mogą znowu tworzyć pary e^+e^- i proces powtarza się, w wyniku czego tworzy się kaskada elektronowo-fotonowa. Ilustracja 29 (tabl. 10) przedstawia rozwój kaskady elektronowo-fotonowej zarejestrowanej w ksenonowej komorze pęcherzykowej.

Oddziaływanie elektromagnetyczne kwantów γ

zjawisko Comptona

Kwanty γ, czyli kwanty promieniowania elektromagnetycznego o dużej energii, są obojętne elektrycznie. O ich przejściu przez substancję możemy wnioskować na podstawie charakterystycznych zjawisk oddziaływania elektromagnetycznego z materią.

Kwant γ o dużej energii, przechodząc przez materię, może się zderzyć z elektronem. W wyniku zderzenia



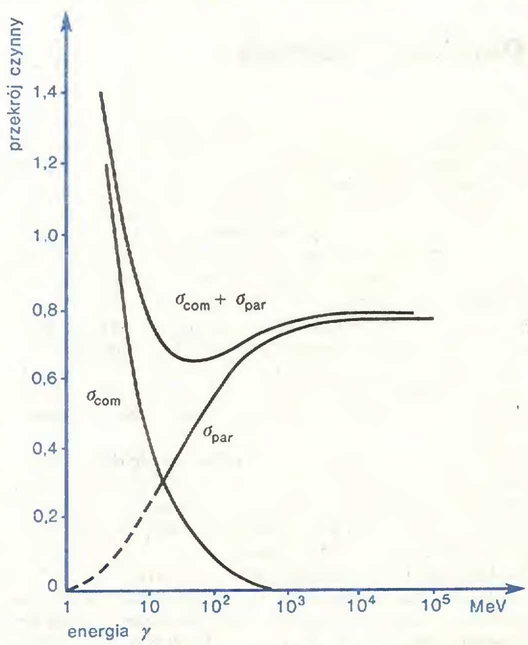
Rys. 4. Zjawisko Comptona

następuje przekazanie elektronowi części energii, kwant γ ze zmniejszoną energią zmienia kierunek lotu, a elektron zostaje odrzucony z pewną prędkością (rys. 4). Energia, a więc częstość ν' rozproszonego kwantu, $E' = h\nu'$, zależy od kąta rozproszenia θ i energii pierwotnej $E = h\nu$:

$$E' = \frac{E m_e c^2}{m_e c^2 + E(1-\cos\theta)}.$$

Energia rozproszonego kwantu jest maksymalna ($E' = E$), gdy $\theta = 0$, i minimalna, gdy $\theta = \pi$. Dla

przekrój czynny na zjawisko Comptona



Rys. 5. Przekrój czynny na tworzenie się par e^+e^- i zjawisko Comptona w powietrzu wyrażony w jednostkach $4{,}5 \cdot 10^{-26}$ cm^2 na nukleon

dużych energii $E \gg m_e c^2$, $E'_{max} = m_e c^2/2$. Rysunek 5 przedstawia wykres przekroju czynnego na rozpraszanie komptonowskie w powietrzu kwantów γ w zależności od ich energii. Na zdjęciu (tabl. 10, il. 32) pokazano elektron zarejestrowany w komorze freonowej wybity w zjawisku Comptona. W procesie tym kwant γ znika, a jego energia przekształca się całkowicie w energię pary elektron-pozyton. Zdjęcie ilustruje powstanie pary elektronowo-pozytonowej w komorze wodorowej w obecności pola magnetycznego. Z kierunku zakrzywienia torów wnioskujemy, że musiały to być cząstki o przeciwnych znakach. Ze względu na zasadę zachowania pędu produkcja par musi następować w pobliżu jądra, które przejmuje część pędu fotonu. Przekazanie energii jądru można pominąć, ponieważ ma ono masę znacznie większą od elektronu. Prawo zachowania energii może być zapisane w postaci

zjawisko tworzenia par

$$E = E' + E'' + 2m_e c^2,$$

gdzie E' i E'' są energiami kinetycznymi pozytonu i elektronu, a E — energią fotonu. Na rys. 5 przedstawiono przekrój czynny na tworzenie par w zależności od energii. Porównanie z zależnością przekroju czynnego na rozpraszanie komptonowskie wskazuje, że przy dużych energiach najważniejszym zjawiskiem pozwalającym wnioskować, że przez materię przeleciał foton, jest tworzenie się par.

Inne zjawiska

Z praktycznego punktu widzenia należy wymienić jeszcze dwa zjawiska stanowiące oddziaływanie elektromagnetyczne fotonu z całym atomem: zjawisko fotoelektryczne i zjawisko wzbudzenia atomu. Wzbudzenie atomu omawialiśmy przy okazji przechodzenia promieniowania jonizującego przez materię. Powrotowi atomu do niższego stanu energetycznego może towarzyszyć emisja promieniowania, które łatwo zarejestrować w detektorze.

Zjawisko fotoelektryczne, polegające na wybijaniu z substancji elektronów, ma duże zastosowanie w fotopowielaczach, pozwalających rejestrować słabe sygnały świetlne.

Detektory cząstek

Cząstki — i to zarówno naładowane elektrycznie, jak i neutralne — mają zawsze różne od zera prawdopodobieństwo oddziaływania z materią, przez którą przechodzą. Niekiedy to prawdopodobieństwo oddziaływania jest niezwykle małe, jak np. neutrina, które bierze udział tylko w oddziaływaniach słabych. Wtedy o przejściu takiej cząstki możemy wnioskować tylko na podstawie zaobserwowanych produktów oddziaływania słabego. W znacznie lepszej sytuacji jesteśmy, badając cząstki, które mogą brać udział w procesach elektromagnetycznych. Oddziałując elektromagnetycznie z materią, zostawiają ślad np. w postaci zjonizowanych atomów (tym samym sygnalizują swoje przejście), przy czym tracą na to bardzo mało energii.

oddziaływanie cząstek z materią

Problem detekcji cząstek sprowadza się do zaobserwowania i zinterpretowania pozostawionych przez cząstkę śladów. Ze względów praktycznych detektory dzieli się na takie, które możliwie wiernie rejestrują ślad cząstki (detektory śladowe), oraz takie, w których istotne jest tylko zarejestrowanie faktu, że cząstka przeszła, i ewentualnie określenie jej natury. Granica podziału zaciera się obecnie coraz bardziej, buduje się bowiem układy hybrydowe, w których możliwie dokładnie określa się i tor cząstki i jej naturę za pomocą wielu typów współdziałających detektorów.

Emulsja jądrowa

Historia zastosowania emulsji fotograficznej do rejestracji cząstek sięga badań A. Becquerela w 1896 r. nad fluorescencją związków uranu. Zauważył on, że związki te zadymiały kliszę fotograficzną. Chociaż Becquerel szybko się przekonał, że zadymienie nie miało nic wspólnego z fluorescencją, minęło wiele lat, zanim stwierdzono, że jest ono związane z przechodzeniem przez emulsję cząstek jonizujących. Dopiero w 1911 r. M. Reinganum zarejestrował fotograficznie ślad cząstki α, ale przez wiele następnych lat nie umiano wyprodukować takiej emulsji fotograficznej, która byłaby czuła na dowolne cząstki naładowane. W 1948 r. firma Kodak dokonała tego, a odtąd udoskonalano emulsje wielokrotnie i stosowano w wielu eksperymentach.

Emulsja fotograficzna składa się z halogenków srebra, głównie AgBr w zawiesinie żelatynowej. Reakcja na światło lub na cząstki zależy od jonizacji wytworzonej w aktywnych ziarnach. Jonizacja ta, czyli oderwanie elektronu z powłoki atomowej, powstaje w wyniku przekazania atomowi przez przechodzącą cząstkę naładowaną lub kwant promieniowania elektromagnetycznego odpowiedniej energii. Jonizacja atomów nie powoduje żadnych widocznych zmian w ziarnach halogenków srebra, lecz zmienia ich strukturę w ten sposób, że mogą reagować z czynnikiem wywołującym. Ziarna, w których nie nastąpiła jonizacja, nie dają się wywołać. Mechanizm zmian wewnętrznych w ziarnie kryształu bromku srebra można w przybliżeniu przedstawić następująco: Na powierzchni kryształu AgBr są małe centra czułości składające się z cząsteczek Ag_2S, które mogą wychwytywać swobodne elektrony poruszające się w krysztale. Elektrony uwolnione z atomów wskutek ich jonizacji dyfundują aż na powierzchnię kryształu i tam są wyłapywane przez centra czułości, które ładują się ujemnie. W krysztale AgBr są zawsze swobodne dodatnio naładowane jony srebra, które wskutek dyfuzji również mogą dotrzeć do ujemnych już teraz centrów czułości. Tam następuje zobojętnienie ładunków i na powierzchni kryształu osadzają się kryształy srebra — powstaje obraz utajony. W procesie wywoływania srebro to katalizuje redukcję jonów srebra z wnętrza kryształu do srebra metalicznego. Tak powstaje obraz wywołany. Na str. 216 (rys. 1) przedstawione jest historyczne już dzisiaj zdjęcie pierwszego zarejestrowanego przypadku produkcji i rozpadu najprawdopodobniej hiperjądra $^{11}_{\Lambda}B$, czyli jądra boru 11, w którym jeden z nukleonów został zastąpiony przez hiperon Λ^0. Zdarzenie to zaobserwowali w 1953 r. Marian Danysz i Jerzy Pniewski z Uniwersytetu Warszawskiego, zapoczątkowując tym odkryciem nową dziedzinę fizyki — fizykę hiperjąder.

mechanizm rejestracji

Spójrzmy uważnie na zdjęcie z punktu widzenia charakteru torów. Niektóre z nich są ciemne, „tłuste" — to te pozostawione przez cząstki o dużym ładunku, np. przez hiperfragment $^{11}_{\Lambda}B$, lub cząstki bardzo powolne. Gęstość liniowa ziaren srebra jest miarą jonizacji. Z odstępstw śladu od linii prostej można wyznaczyć średnią wartość kwadratu wielokrotnego rozproszenia. Te dwa pomiary pozwalają na określenie pędu i prędkości cząstki, a co za tym idzie — jej masy.

wyznaczanie pędu i prędkości

Technika emulsji jądrowych przeżyła już okres swojej świetności. Jest jednak używana jako technika pomocnicza, szczególnie gdy wymagana jest duża zdolność rozdzielcza, długi czas czułości oraz małe rozmiary detektora.

Komora Wilsona (komora mgłowa)

Zjonizowane atomy w przesyconej (przechłodzonej) parze odgrywają rolę zarodzi kondensacji dla kropelek cieczy powstających ze skraplającej się pary. Komora Wilsona składa się zasadniczo z naczynia zawie-

rającego gaz w stanie dalekim od skroplenia oraz parę jakiejś cieczy nasyconą lub bliską nasycenia. Objętość naczynia można nagle powiększać. Temperatura mieszaniny przy tym maleje i para przechodzi w stan przesycenia. Jeżeli w gazie jest kurz, nawet niewielkie rozprężenie powoduje powstanie w całym naczyniu mgły, ponieważ kondensacja rozpoczyna się na cząsteczkach pyłu. W komorze wolnej od pyłu nie obserwuje się kondensacji, dopóki stosunek objętości komory po i przed rozprężeniem (stosunek rozprężenia) nie osiągnie tzw. granicy jonowej. Po jej przekroczeniu kropelki cieczy pojawiają się tam, gdzie są jony. W silnym oświetleniu kropelki te widać jako jaskrawe punkty na czarnym tle. Jony powstałe wskutek przejścia przez komorę naładowanej cząstki są zarodziami kropelek układających się w ślad cząstki. Ilustracja 27 (tabl. 9) przedstawia wykonaną przy użyciu komory Wilsona pierwszą rejestrację hiperonu Ξ^- (ksi), który się rozpada na hiperon Λ^0 i mezon π^-:

granica jonowa

$$\Xi^- \rightarrow \Lambda^0 + \pi^- .$$

Hiperon Λ^0 z kolei rozpada się na proton i π^-. Stąd załamanie śladu cząstki pierwotnej jest miejscem rozpadu hiperonu Ξ^- na hiperon neutralny Λ^0, który nie zostawił śladu w komorze, i ujemnie naładowany π^-. Hiperon Λ^0 przeleciał pewien odcinek drogi i rozpadł się na dwie cząstki naładowane, tworząc charakterystyczny ślad w kształcie widełek.

Komory Wilsona nie są obecnie stosowane. Są one mało wydajne, ponieważ nie można zbyt często powtarzać rozprężenia. Gaz wypełniający ma niewielką gęstość, czyli małą liczbę jąder w porównaniu z emulsją jądrową lub komorą pęcherzykową, i dlatego prawdopodobieństwo zaobserwowania oddziaływania jest stosunkowo małe. Komora Wilsona nadaje się natomiast do badania cząstek promieniowania kosmicznego. Jony w tej komorze żyją dostatecznie długo, aby można było wywoływać rozprężenie w jakiś czas po przejściu cząstki. Pozwala to na współpracę komory z licznikowym urządzeniem wyzwalającym rozprężanie wtedy, gdy zdarzenie spełnia założone kryteria.

Komora pęcherzykowa

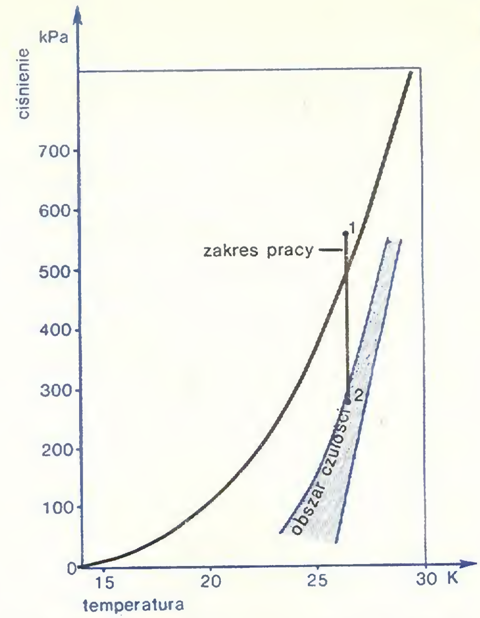
W komorze tej pęcherzyki pary tworzą się w cieczy przegrzanej. Istnienie stanu cieczy przegrzanej jest zjawiskiem bardzo dobrze znanym. Wiadomo również, że naruszenie równowagi przez dostarczenie niewielkich nawet ilości energii zapoczątkowuje proces wrzenia. Przejście z jednego do drugiego stanu skupienia jest przejściem fazowym I rodzaju (→ Przejścia fazowe). Jeżeli nowym stanem skupienia jest ciecz (albo gaz), powierzchnia rozdziału faz przybiera kształt kuli, aby zminimalizować energię powierzchniową. Można wykazać, że istnieje promień krytyczny takiej kuli, poniżej którego nowa faza jest niestabilna, powyżej zaś kula rośnie, a metatrwały stan starej fazy zanika. Aby kula — pęcherzyk osiągnęła promień krytyczny, należy wykonać pewną pracę. Jeżeli nie ma mechanizmu dostarczającego energię, krytyczny promień nigdy nie zostaje osiągnięty i ciecz pozostaje w początkowym stanie skupienia nieskończenie długo.

promień krytyczny

Cząstka naładowana, przelatując przez przegrzaną ciecz, jonizuje atomy, czyli dostarcza im energii. Jeżeli energia ta zostanie przekazana pęcherzykom pary, będą one mogły osiągnąć krytyczny promień i wzdłuż śladu cząstki rozpocznie się wrzenie cieczy.

Pomijając szczegółowe rozważania teoretyczne co do mechanizmu przekazywania energii cząstki pęcherzykowi pary, stwierdzamy, że podwyższenie temperatury w pęcherzyku powoduje wzrost jego rozmiaru powyżej promienia krytycznego, a tym samym umożliwia dalszy jego rozwój. Duży pęcherzyk można zobaczyć widząc tym samym ślad cząstki. Rysunek 6 przedstawia wykres zależności ciśnienia od temperatu-

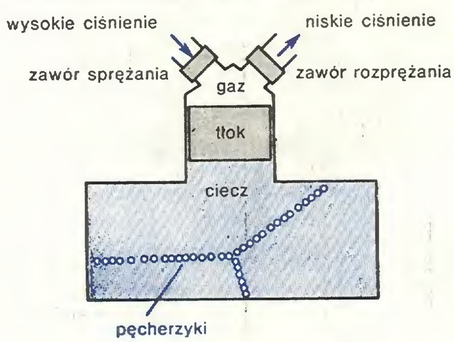
warunki pracy komory wodorowej



Rys. 6. Krzywa ciśnienie-temperatura dla ciekłego wodoru. Pionowa linia wskazuje możliwe warunki pracy komory

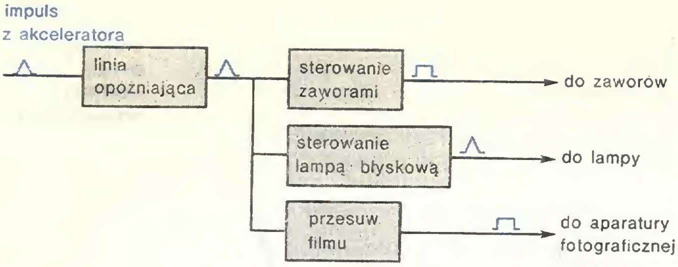
ry dla ciekłego wodoru. Obszar zacieniowany odpowiada takim wartościom ciśnienia i temperatury, w których się nie rozpoczyna jeszcze samorzutne wrzenie, lecz energia jonizacji jest już wystarczająca do wzrostu pęcherzyków. Jest to obszar czułości. Pionowa linia odpowiada pracy komory w temperaturze $T = 27$ K. Zasada działania komory jest prosta. Załóżmy, że ciecz komory jest we właściwej temperaturze. Jedynym problemem jest sterowanie ciśnieniem. W pierwszej fazie ciśnienie jest większe niż ciśnienie pary nasyconej (*1*). W drugiej fazie następuje rozprężenie do ciśnienia odpowiadającego obszarowi czułości (*2*). Jeżeli teraz przez komorę przejdzie cząstka naładowana, to zjonizuje atomy na swej drodze i dostarczy energii na wzrost pęcherzyków do rozmiarów krytycznych. Pęcherzyki rosną przez kilka milisekund i są fotografowane. Ostatnią fazą jest zwiększanie ciśnienia do stanu początkowego. Rysunek 7 pokazuje schematyczny szkic komory pęcherzykowej.

obszar czułości

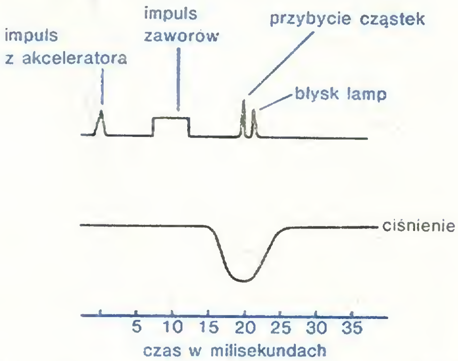


Rys. 7. Schemat komory pęcherzykowej

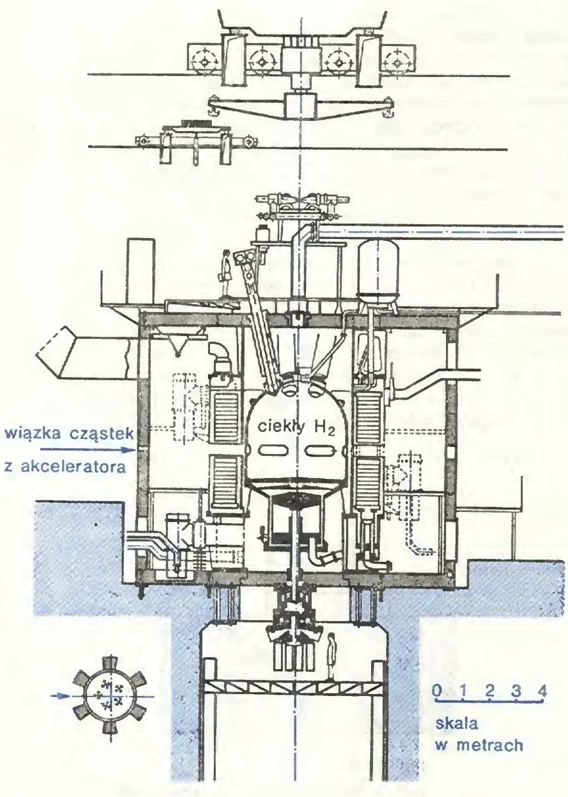
Widać stąd, jak bardzo ważną rzeczą jest synchronizacja czasu nadejścia cząstki, chwili rozprężenia komory i zapłonu lamp błyskowych, umożliwiających fotografowanie. Dlatego komora pęcherzykowa może współpracować tylko z akceleratorem, wtedy dokładnie wiadomo, w której chwili nadleci wiązka przyspieszanych cząstek. Schemat blokowy sterowania komorą podaje rys. 8. Na rys. 9 przedstawiono możliwy zapis oscyloskopowy impulsów informujących o przebiegu pracy układu akcelerator–komora.



Rys. 8. Schemat blokowy sterowania komorą



Rys. 9. Przykład zapisu oscyloskopowego impulsów sterujących komorą i zmian ciśnienia



Rys. 10. Przekrój pionowy przez Wielką Europejską Komorę Pęcherzykową

Pierwszą komorę pęcherzykową zbudował D.A. Glaser w 1953 r. Była ona bardzo mała. Zdjęcie 33 (tabl. 10) przedstawia ślad elektronu w szklanej komorze pęcherzykowej o rozmiarach: $^1/_2$ cala (1,27 cm) średnicy i 1 cal (2,54 cm) długości, wypełnionej izopentanem, w temperaturze 130°. W następnych latach rozwinięto technikę budowy komór wypełnionych różnymi cieczami, w szczególności ciekłym wodorem.

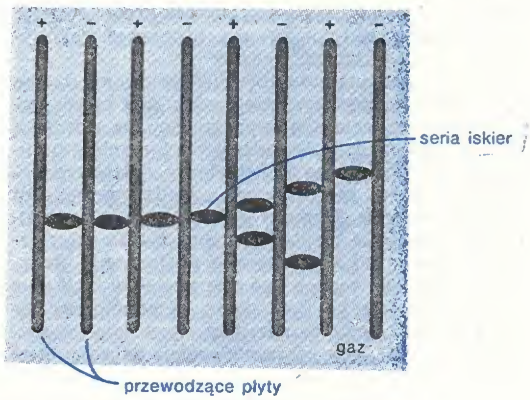
BEBC

Największa istniejąca obecnie komora wodorowa BEBC (*Big European Bubble Chamber*) mieści 33,5 m³ wodoru. Na rys. 10 jest schematyczny przekrój tej komory. Całość jest umieszczona w polu magnetycznym 3,5 T nadprzewodzącego magnesu. Celem uzmysłowienia skali urządzenia na dolnym podeście narysowana jest postać ludzka. Na zdjęciu 35 (tabl. 11), pokazano komorę wraz z urządzeniem do identyfikacji cząstek.

Badania z użyciem komór wodorowych ogromnie rozszerzyły zakres wiedzy o cząstkach i ich oddziaływaniach.

Komory iskrowe

Podstawową rolę w komorach iskrowych odgrywa wyładowanie w gazie w przerwie między dwiema elektrodami (rys. 11). Między elektrodami wytwarza się różnicę potencjałów dostatecznie dużą, aby spowodować wyładowanie elektryczne, jeżeli w przerwie między nimi pojawią się elektrony lub jony. Przebieg zjawiska jest złożony. W uproszczeniu można powiedzieć, że jony i elektrony w zewnętrznym polu



Rys. 11. Zasada działania komory iskrowej

elektrycznym uzyskują dostateczną energię, by jonizować napotkane atomy, wywołując tym samym lawinowe narastanie wyładowania. Przeskok iskry następuje najłatwiej w miejscu, gdzie się znajdują zjonizowane atomy. Przestrzeń między płytami odległymi o 1-2 cm wypełnia się gazem. Stosując szereg równoległych płyt pod napięciem, można uzyskać ślad toru w postaci szeregu iskier. Zdjęcie 30 (tabl. 10) zostało wykonane w dużym układzie komór iskrowych zbudowanym w Zjednoczonym Instytucie Badań Jądrowych w Dubnej. Widać na nim oddziaływanie pionu o energii 40 GeV z jądrem żelaza, w wyniku czego powstało 6 cząstek naładowanych. Zdjęcie to było przykładem optycznego (fotograficznego) odczytu informacji. W praktyce stosuje się jeszcze inne metody, np.:

metody odczytu informacji

metoda akustyczna, polegająca na detekcji fali dźwiękowej towarzyszącej iskrze:

metoda rdzeni ferrytowych, które ulegają przemagnesowaniu przez prąd iskry; informację o przemagnesowaniu rdzeni odczytuje się elektrycznie;

metoda magnetostrykcji; w pobliżu drutów zasilających umieszcza się metalowy pręt; gdy przeskakuje iskra, w pręcie generuje się mechaniczny impuls; opóźnienie w dotarciu fali sprężystej do końca pręta określa pozycję drutu, a więc i płyty, która brała udział w wyładowaniu;

metoda pojemnościowa; przeskok iskry powoduje zmianę pojemności w układzie kondensatorów sprzężonych z płytami; rejestrację przeprowadza układ elektroniczny.

Bardzo dokładne informacje o współrzędnych toru cząstki (z dokładnością do 100 μm) można uzyskać

z komory iskrowej z przerwą między płytami rzędu 4–10 cm. Iskra biegnie wtedy prawie dokładnie wzdłuż śladu jonizacji. Zjawisko rozwoju wyładowania w gazie, będące podstawą działania komory iskrowej, jest wykorzystywane w pokrewnych typach detektorów, jak np. komora dryfowa, strimerowa, proporcjonalna wielodrutowa. Komory dryfowa i proporcjonalna wielodrutowa wykorzystują zjawisko wzmocnienia gazowego. Chmura elektronów wytworzona przez jonizującą cząstkę porusza się (dryfuje) w kierunku bardzo cienkiego przewodnika (np. pozłacany drut wolframowy o średnicy 20 μm). W pobliżu (ok. 50 μm) drutu istnieje pole elektryczne o bardzo dużym natężeniu, tak przyspieszające elektrony, że następuje jonizacja lawinowa atomów. Wypadkowy impuls potencjału na drucie może być wzmocniony 10^6 razy w stosunku do impulsu, jaki by wywołała pierwotna chmura elektronów.

komora dryfowa i proporcjonalna wielodrutowa

W komorze dryfowej mierzy się opóźnienie sygnału z drutu względem chwili przejścia cząstki przez komorę wyznaczonej za pomocą np. licznika scyntylacyjnego. Znając prędkość dryfu (rzędu $5 \cdot 10^6$ cm/s przy natężeniu pola 1 kV/cm), można ocenić położenie miejsca, w którym nastąpiła jonizacja, z dokładnością do 100 μm.

komora drutowa

W komorze drutowej, złożonej z dużej liczby cienkich drutów odległych od siebie o 1–5 mm, sygnał z drutu określa równocześnie jego pozycję. Należy stosować układy wzajemnie prostopadłych przewodników, jeżeli się chce odczytać trzy współrzędne miejsca, w którym cząstka zjonizowała gaz. Komora strimerowa jest komorą iskrową o dużej przerwie. Napięcie przykłada się do elektrod na bardzo krótko ($2 \cdot 10^{-8}$ s), tak że iskra dopiero się zaczyna formować. Na filmie fotografuje się zaczątki iskier (strimery), które pokazują przebieg cząstki. Wszystkie detektory oparte na zasadzie rozwoju wyładowania w gazie odznaczają się dużą szybkością działania, tzn. mogą wykonać kilka milionów rejestracji w ciągu sekundy. Ma to ogromne znaczenie przy poszukiwaniu i badaniu rzadkich procesów.

komora strimerowa

Liczniki Czerenkowa

W działaniu licznika wykorzystuje się zjawisko Czerenkowa. Cząstka jonizująca przechodzi przez odpowiednio dobrany ośrodek, zwany konwerterem. Promieniowanie elektromagnetyczne zjawiska Czerenkowa, w większości w zakresie długości fal widzialnych, jest rejestrowane przez fotopowielacze i daje informację o przejściu cząstki o prędkości większej niż prędkość graniczna (patrz tabela str. 111), a więc licznik Czerenkowa jest licznikiem progowym. Ponieważ emisja promieniowania elektromagnetycznego odbywa się pod ściśle określonym kątem do kierunku lotu cząstki, zależnym od jej prędkości i współczynnika załamania ośrodka konwentera, można budować liczniki czułe tylko na wybrane cząstki. Światło emitowane w stożku o określonym kącie rozwarcia jest ogniskowane na pierścieniu z fotopowielaczy. Fotony nie spełniające danych kryteriów są ogniskowane poza obszarem czułego pierścienia i nie są rejestrowane. Prędkość β wszystkich cząstek o dużej energii — niezależnie od ich masy — zdąża do jedności ($\beta = v/c$). W takim wypadku konwerter musi mieć współczynnik załamania bliski jedności,

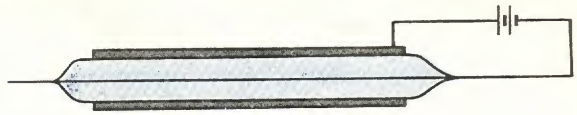
$$n < 1/\beta,$$

i dlatego trzeba używać gazu pod małym ciśnieniem. To z kolei zmniejsza ilość fotonów emitowanych na jednostkę drogi w konwerterze i zmusza do stosowania bardzo długich konwerterów.

Licznik ogniskujący fotony promieniowania Czerenkowa na wybranym pierścieniu pozwala określić równocześnie kierunek lotu i prędkość. Ilustracja 36 na tablicy 11 przedstawia wygląd zewnętrzny licznika Czerenkowa.

Licznik proporcjonalny

Zasadę działania tego typu detektora ilustruje rys. 12. W przewodzącej rurce wypełnionej gazem przeciągnięty jest wzdłuż osi cienki drut. Między rurką a osłoną panuje różnica potencjałów ok. 1 kV. Przejściu



Rys. 12. Zasada działania licznika proporcjonalnego

cząstki naładowanej przez licznik towarzyszy jonizacja. Swobodne ładunki elektryczne wędrują do elektrod. W pobliżu środkowej elektrody następuje zjawisko wzmocnienia gazowego, tak więc ładunek elektryczny powstający na elektrodzie zbierającej licznika jest N razy większy od ładunku jonizacji pierwotnej. Współczynnik N nazywa się współczynnikiem wzmocnienia gazowego; jego wartość zależy od napięcia zasilania oraz rodzaju i ciśnienia gazu wypełniającego licznik. Liczniki proporcjonalne znalazły bardzo szerokie zastosowanie w detekcji cząstek, w pomiarze jonizacji, a w szczególności do pomiarów aktywności źródeł promieniotwórczości o małej energii wysyłanych cząstek.

współczynnik wzmocnienia gazowego

Urządzenie do pomiaru jonizacji

Il. 35 (tabl. 11) przedstawia 52-tonowe urządzenie do identyfikacji cząstek, umieszczone za (patrząc od strony nadlatujących cząstek) Wielką Europejską Komorą Pęcherzykową. Do identyfikacji cząstek wykorzystuje się zjawisko wzrostu jonizacji przy bardzo dużych prędkościach cząstek (rys. 1). Detektor składa się z 4096 liczników proporcjonalnych, ułożonych w ciśnieniowym pojemniku w 128 warstwach, po 32 liczniki w każdej warstwie. Długość detektora wynosi 8,5 m, a efektywna powierzchnia detekcji 1 m × 2 m. Energia jonizacji, wydzielona w każdym liczniku przez cząstkę jonizującą, jest automatycznie mierzona przez układ elektroniczny. Dla każdego toru odbywa się więc 128 pomiarów wydzielonej energii. Stąd można wyznaczyć jonizację z dokładnością do ok. 3%, a tym samym można określić prędkość cząstki. Jeżeli znany jest pęd z pomiaru krzywizny toru w polu magnetycznym, to można zidentyfikować cząstkę.

Układy eksperymentalne

Dwuramienny spektrometr mionów

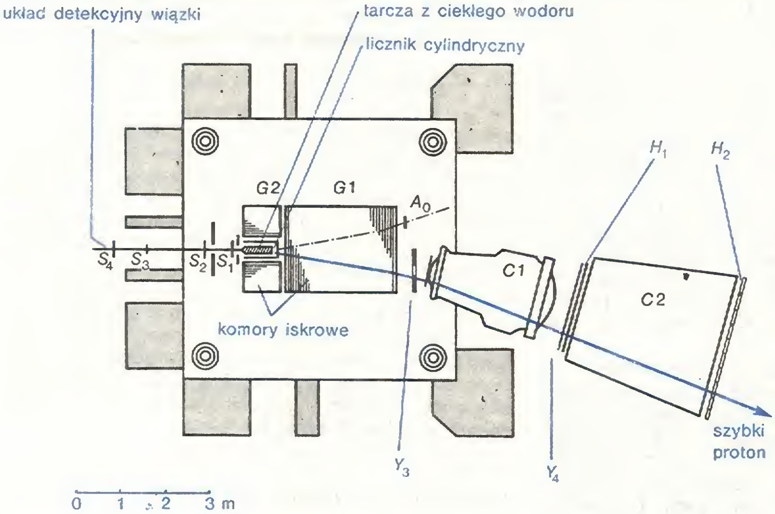
W laboratorium imienia Fermiego grupa fizyków pod kierunkiem Leona Ledermana odkryła w 1977 r., badając reakcje p + tarcza $\to \mu^+\mu^-$ + cokolwiek, nową cząstkę — ypsilon. Pęd protonu dobrze znano, należało zidentyfikować, zarejestrować i wyznaczyć pędy obu mionów μ. W tym celu użyto dwuramiennego spektrometru, przedstawionego na zdjęciu (tabl. 11, il. 37) wykonanym z osłony tarczy wzdłuż wiązki. Na pierwszym planie widać dwa duże analizujące magnesy, za nimi umieszczone są układy liczników wyzwalających układ, następnie wielodrutowe komory proporcjonalne, liczniki Czerenkowa i absorbery stalowe. Układ ten mógł zarejestrować zdarzenia, podczas których zarówno w jednym, jak i drugim ramieniu przeleciał mion, oraz określić pęd i kierunek lotu każdego mionu. Pozwoliło to na poszukiwanie cząstek, które rozpadają się na dwa miony. Eksperyment zakończył się sukcesem. Odkryto cząstkę ypsilon, Υ.

odkrycie cząstki Υ

Spektrometr Ω

Zdjęcie 38 (tabl. 11) przedstawia widok ogólny spektrometru Ω. Urządzenie to łączy zalety komory pęcherzykowej — umożliwiającej rejestrację cząstek emitowanych w dowolnym kierunku z punktu oddziaływania — z szybkością pracy komór iskrowych. Szybkość rejestracji jest tu bardzo istotna. Można bowiem dokonywać bezpośredniej selekcji zdarzeń i rejestrować tylko określone procesy. Rysunek 13 ukazuje schematycznie budowę spektrometru. W 1400-tonowym jarzmie elektromagnesu umieszczono 2 nadprzewodzące cewki, pozwalające przy natężeniu prądu 4800 A wytwarzać pole magnetyczne o indukcji 1,8 T w obszarze 14 m³. Wewnątrz tego obszaru znajdują się tarcza wodorowa oraz dwa układy komór iskrowych *G*2 i *G*1. Komory są „widziane" przez specjalne kamery telewizyjne, tzw. plumbikony, które automatycznie rejestrują współrzędne iskry. Informacja ta może być zapisana na taśmie magnetycznej maszyny cyfrowej, jeżeli przypadek spełnia dane kryteria eksperymentu. O tym decyduje układ liczników sterujących systemem wyboru danych i zmieniających się w zależności od potrzeb eksperymentu. Opiszemy układ wybierający tylko zdarzenia z bardzo szybkim protonem emitowanym do przodu. W komorze pęcherzykowej, w której się rejestruje wszystko po kolei, trzeba by wykonać kilkadziesiąt lub kilkaset tysięcy zdjęć (zależnie od reakcji), aby zaobserwować odpowiednią do analizy liczbę zdarzeń z szybkim protonem. Identyfikacja szybkiego protonu jest w komorze pęcherzykowej trudna, więc nawet zebrana w ten sposób próbka zdarzeń byłaby „zanieczyszczona". W spektrometrze Ω zadanie jest łatwiejsze.

możliwość selekcji zdarzeń



Rys. 13. Spektrometr Ω — schemat budowy. *G*1 i *G*2 układy komór iskrowych, S_1, S_2, S_3 i S_4 liczniki scyntylacyjne, *C*1 i *C*2 liczniki Czerenkowa, Y_3 i Y_4 drutowe komory proporcjonalne, H_1, H_2 i A_0 liczniki dodatkowe wspomagające układ sterowania (zob. też il. 38, tabl. 11)

układ wybierający

Wiązka cząstek ujemnych o określonym pędzie wchodzi do obszaru tarczy z lewej strony, przy czym nie pokazane na rys. 13 liczniki Czerenkowa pozwalają określić, czy jest to mezon π^-, mezon K^- czy antyproton. Ponieważ (tabela str. 111) przy tej samej prędkości minimalnej β_{min} proton musi mieć znacznie większy pęd niż mezon π, aby wywołać zjawisko Czerenkowa wystarczy więc przy określonym pędzie cząstek wiązki umieścić dwa liczniki Czerenkowa o tak dobranych substancjach, aby pierwszy (C_π) reagował tylko na mezony π^-, drugi (C_K) na mezony K^- i π^-, a był nieczuły na antyprotony. Wtedy przejściu mezonu π^- będzie odpowiadał sygnał z obu liczników, przejściu K^- sygnał tylko z jednego licznika, a przy przejściu antyprotonu żaden z liczników nie zareaguje. Oznaczając kreską poziomą nad symbolem licznika brak sygnału, możemy zapisać symbolicznie sygnały związane z przejściem cząstki:

$$\pi^- : C_\pi \cdot C_K, \quad K^- : \bar{C}_\pi \cdot C_K, \quad \bar{p} : \bar{C}_\pi \cdot \bar{C}_K.$$

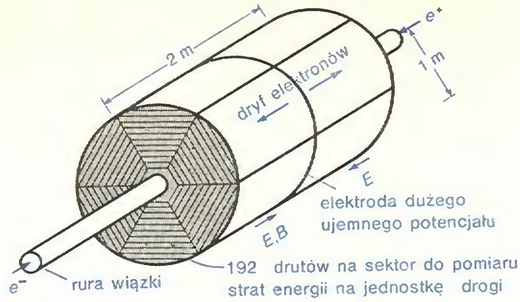
Dodatkowe liczniki scyntylacyjne *S* pozwalają stwierdzić przejście cząstki i określić natężenie wiązki padającej. Za komorą umieszczono dwa duże liczniki Czerenkowa, *C*1 i *C*2, pozwalające zidentyfikować szybką cząstkę opuszczającą obszar oddziaływania, jeżeli znany jest jej pęd. Pęd ten wyznaczają dwie drutowe komory proporcjonalne Y_3 i Y_4. Znając położenie tarczy, a więc tym samym przybliżone położenie punktu oddziaływania, oraz współrzędne toru w komorach Y_3, Y_4 i konfigurację pola magnetycznego, można określić pęd cząstki. Sygnały z dwóch liczników Czerenkowa o odpowiednio dobranych parametrach wraz ze znajomością pędu pozwalają zidentyfikować cząstkę. Układ zawiera jeszcze dodatkowe liczniki H_1, H_2, A_0, wspomagające układ sterowania. W chwili, gdy ze współpracujących liczników nadejdą sygnały stwierdzające, że w obszarze tarczy nastąpiło oddziaływanie i że wśród cząstek wtórnych był proton o pędzie powyżej danej wartości, układ elektroniczny sterowania podejmie decyzję zebrania informacji ze wszystkich komór iskrowych i liczników pomocniczych i zapisania jej na taśmie magnetycznej. Decyzja taka jest podejmowana w $7{,}5 \cdot 10^{-7}$ s po przejściu cząstki. Dalsze opracowywanie danych odbywa się już znacznie później, z wykorzystaniem dużych komputerów. Jest to proces żmudny, wymagający wielu obliczeń, prowadzący do rekonstrukcji w przestrzeni zaszłego zdarzenia. Komputer, znając współrzędne wszystkich iskier, może na żądanie operatora odtworzyć na ekranie telewizyjnym obraz zdarzenia.

Urządzenie TPC

TPC — komora z projekcją czasu

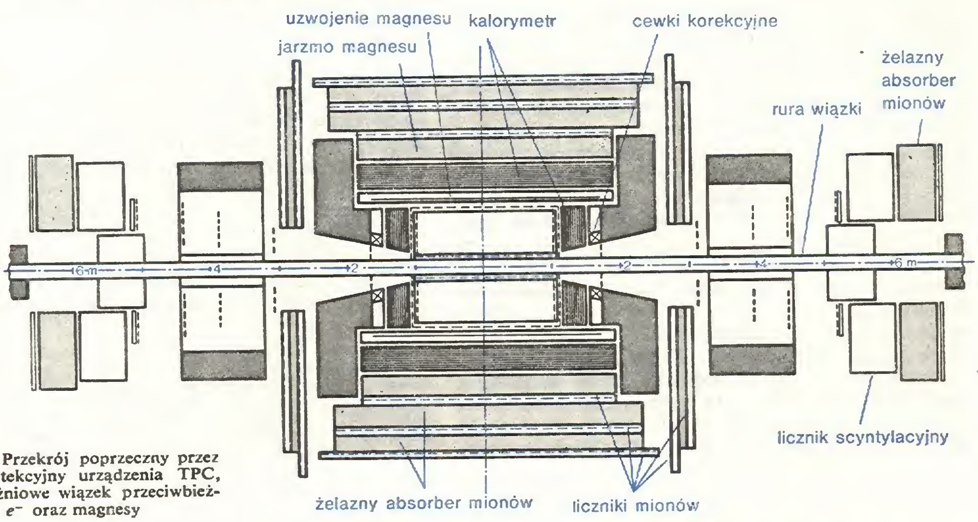
Pod tym skrótem kryje się angielska nazwa *Time Projection Chamber*, czyli komora z projekcją czasu. Urządzenie jest dopiero w budowie w Berkeley i będzie gotowe kosztem 10 mln dolarów w pierwszej połowie lat osiemdziesiątych. W założeniu konstruktorów komora zapewni rejestrację torów poszczególnych cząstek oraz identyfikację elektronów, pionów, kaonów i protonów w bardzo szerokim zakresie kąta bryłowego, a nawet dla zdarzeń z wieloma cząstkami wtórnymi. Komora pozwoli wyznaczyć również energię kwantów γ oraz zidentyfikować miony. Jednym słowem ma to być niemal idealny detektor, o jakim śnili fizycy. Ma on współdziałać z przeciwbieżnymi wiązkami e^+ i e^- w eksperymentach mających na celu poszukiwanie nowych cząstek.

Urządzenie to składa się z dużej cylindrycznej komory dryfowej o średnicy i długości 2 m (rys. 14). Współrzędne prostopadłe do kierunku dryfu będą odczytywane za pomocą układu drutów. Informację



Rys. 14. Szkic budowy urządzenia TPC

dotyczącą współrzędnej wzdłuż kierunku dryfu uzyska się na podstawie pomiaru opóźnienia sygnału w drutach w stosunku do czasu przejścia cząstki. Prędkość dryfu elektronów ma wynosić 7 cm/μs. Na torze cząstki można będzie mieć do 192 odczytów współrzędnych (tyle bowiem jest drutów). Pozwoli to zrekonstruować nawet oddziaływania o dużej liczbie cząstek wtórnych. Z wielkości sygnału na każdym drucie można będzie ocenić pierwotny ładunek jonizacji, a więc wyznaczyć straty energii przechodzącej cząstki na jonizację. Informacja ta wraz ze znajomością pędu cząstki pozwoli na jej pełną identyfikację. Oczekuje się, że elektrony, piony, kaony, protony będzie można identyfikować w zakresie pędów 100 MeV/c–15 GeV/c. Cała komora będzie umieszczona w polu magnetycznym nadprzewodzących cewek o indukcji 1,5 T. Energia i kierunek kwantów γ ma być wyznaczony przez kalorymetry wypełnione ciekłym argonem, w których się będzie mierzyć całkowitą energię kwantu γ. Identyfikację powolnych mionów uzyska się z pomiaru jonizacji, a mionów o większej energii — na podstawie zdolności przenikania przez warstwy żelaza. Cały układ będzie sterowany przez komputer wybierający według danych kryteriów określone zdarzenia. Rysunek 15 ilustruje przekrój poprzeczny przez układ detekcyjny, rury próżniowe wiązek przeciwbieżnych e^+e^- oraz magnesy.



Rys. 15. Przekrój poprzeczny przez układ detekcyjny urządzenia TPC, rury próżniowe wiązek przeciwbieżnych e^+ i e^- oraz magnesy

Badanie krótkożyjących cząstek

Dotychczas mówiliśmy o cząstkach, które mogą przejść przez detektor, tzn. ich droga od punktu powstania do punktu rozpadu jest tak długa, że można ją zmierzyć. Przykładem może być powstanie i rozpad hiperonu Λ^0 (il. 31, tabl. 10). Hiperon Λ^0 ma średni czas życia $2{,}6 \cdot 10^{-10}$ s, a więc żyje stosunkowo długo. Mezon π^0 rozpada się średnio po czasie $0{,}8 \cdot 10^{-16}$ s i tylko z najwyższym trudem można ocenić drogę, którą przebywa w emulsji jądrowej od powstania do rozpadu. Znamy cząstki, które żyją jeszcze parę rzędów wielkości krócej niż π^0. Odległość między punktem produkcji i punktem rozpadu jest rzędu 1 fm $= 10^{-15}$ m i bezpośrednia obserwacja toru takiej cząstki, zwanej rezonansem, jest obecnie niemożliwa; o jej istnieniu wnioskujemy na podstawie analizy produktów rozpadu.

doświadczenie Ledermana

Omówimy zagadnienie na przykładzie doświadczenia prowadzonego przez Leona Ledermana i zakończonego odkryciem nowej cząstki Υ (ypsilon) o masie 9,5 GeV. Omawialiśmy już spektrometr dwuramienny, który ustawiono w laboratorium im. Fermiego przy akceleratorze protonów. Protony o energii 400 GeV padały na tarczę, za którą stał spektrometr rejestrujący tylko miony μ^+ i μ^-. Badana reakcja przebiegała następująco:

$$p + \text{jądro tarczy} \rightarrow \mu^+ + \mu^- + \text{cokolwiek, o czym nic nie wiemy.}$$

Mamy natomiast precyzyjną informację, że zostały wyprodukowane dwa miony, wiemy, pod jakimi kątami zostały emitowane i z jaką energią. Sytuację przedstawia rys. 16a. Możemy jednak dopuścić, że naprawdę proces przebiegał inaczej (rys. 16b). W wyniku oddziaływania proton–jądro powstała jakaś cząstka o nieznanej masie m_Υ, która po bardzo krótkim czasie rozpadła się na $\mu^+\mu^-$. Jeżeli czas życia był tak krótki, że punkt rozpadu był doświadczalnie nie do odróżnienia od punktu produkcji, to obraz oddziaływania będzie taki jak na rys. 16a. O hipotetycznej cząstce nie mamy żadnych informacji, znamy natomiast wektory pędu ewentualnych produktów rozpadu $\vec{p}_1$ i $\vec{p}_2$. Piszemy „ewentualnych", ponieważ nie wiemy, czy taka cząstka istnieje, a jeżeli istnieje, czy rozpada się tylko na dwa miony μ^+ i μ^-. Ponieważ nie wiemy, robimy założenie: a) cząstka istnieje, b) rozpada się na parę $\mu^+\mu^-$. Teraz można obliczyć masę m_Υ, czyli masę, jaką miałaby cząstka rozpadająca się na dwa konkretnie zarejestrowane miony. Masę tę nazywamy masą niezmienniczą układu (w tym wypadku układu dwóch mionów). W mechanice relatywistycznej energia cząstki wyraża się wzorem

obliczenie masy niezmienniczej układu

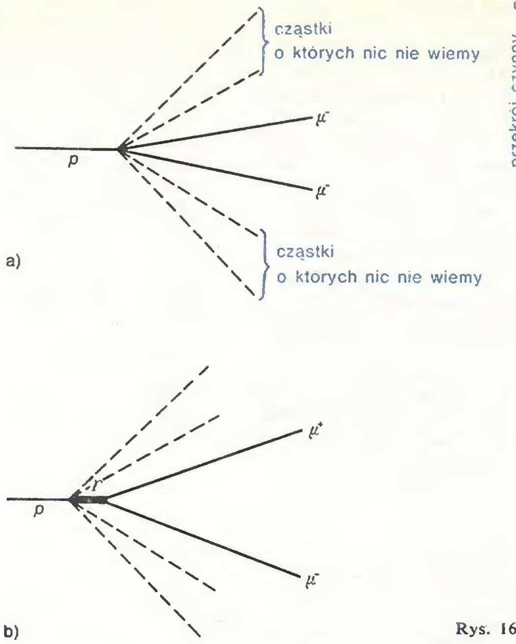
$$E_\Upsilon = \sqrt{p_\Upsilon^2 c^2 + m_\Upsilon^2 c^4},$$

z zasady zachowania pędu otrzymujemy

$$p_\Upsilon = \sqrt{(p_{1x}+p_{2x})^2+(p_{1y}+p_{2y})^2+(p_{1z}+p_{2z})^2},$$

gdzie p_x, p_y, p_z są składowymi pędu $\vec{p}$ wzdłuż odpo-

hipoteza istnienia cząstki Υ



Rys. 16

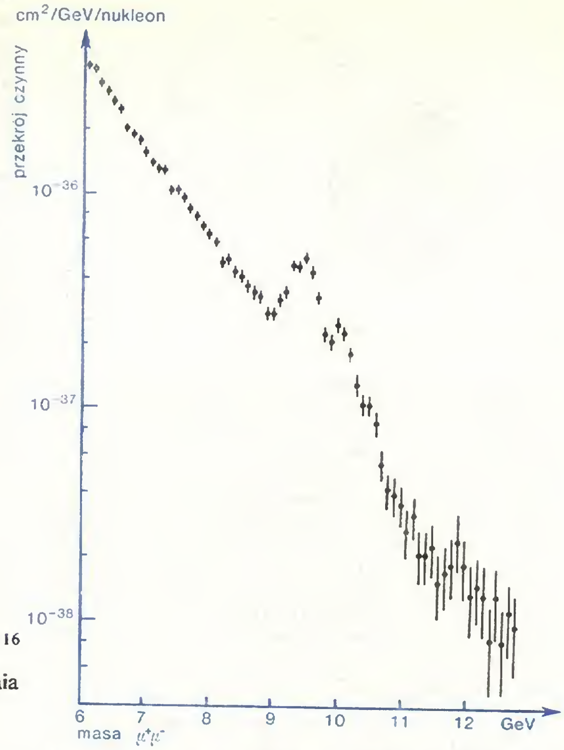
wiednich osi współrzędnych, a z zasady zachowania energii wynika:

$$E_{\Upsilon}=\sqrt{p_1^2c^2+m_\mu^2c^4}+\sqrt{p_2^2c^2+m_\mu^2c^4}\,.$$

Znając z pomiarów całkowite pędy obu mionów i ich składowe, możemy z powyższych wzorów obliczyć m_{Υ}.

Mając do dyspozycji tylko jedną rejestrację, nie moglibyśmy wyciągnąć żadnego wniosku. Każde dwie cząstki dają jakąś masę niezmienniczą, choćby każda z nich powstała w innym procesie. Dopiero dysponując dużą liczbą rejestracji, można sprawdzić, czy obliczone masy określonej kombinacji cząstek grupują się wokół jakiejś wartości. Jeżeli tak, to może znaczyć, że w oddziaływaniu powstała cząstka (mówimy: rezonans) o określonej masie (im węższe jest maksimum, tym dłużej cząstka żyje). Rysunek 17 przedstawia wykres zależności częstości pojawiania się określonej masy pary $\mu^+\mu^-$ od wielkości masy, wykazujący, że w obszarze ok. 10 GeV jest wyraźne zgrupowanie przypadków. Bliższe przyjrzenie się rysunkowi wskazuje na istnienie dwóch maksimów przy 9,4 GeV i 10 GeV, co z kolei wskazuje na istnienie dwóch rezonansów Υ i Υ' o powyższych masach. W podobny sposób wykryto bardzo wiele cząstek.

potwierdzenie istnienia cząstki Υ



Rys. 17. Rozkład masy układu $\mu^+\mu^-$ otrzymany przez zespół Ledermana

Metoda badania tak krótkożyjących tworów jest statystyczna. Nie umiemy na razie wskazać zdarzenia i powiedzieć o nim z całą pewnością, że w tym właśnie wypadku został wytworzony ypsilon. Możemy tylko powiedzieć, że w badanej reakcji produkuje się ypsilon z takim to a takim przekrojem czynnym. Nie zmienia to faktu, że cząstka taka istnieje w takim samym sensie, w jakim istnieje proton. Wszędzie tam, gdzie się będzie produkował ypsilon, stwierdzimy nadmiar kombinacji mas μ^+ i μ^- w obszarze masy 9,4 GeV, jeśli tylko warunki nie będą zabraniały rozpadu na μ^+ i μ^-.

I. Bartke *Wielka Europejska Komora Pęcherzykowa*, Post. Fiz. 28, 433 (1977); D.A. Bromlej (ed.) *Dedectors in Nuclear Science, Nuclear Instruments and Methods*, t. 162, Nr 1-3, p. 1, 2 (1979); G. Charpak *Wielodrutowe i dryfowe komory proporcjonalne*, Post. Fiz. 30, 579 (1979); W.J. Willis *Wielkie spektrometry*, Post. Fiz. 31, 279 (1980); J.W. Zaniewskij *Prowołocznyje dietiektory elementarnych czastic*, Moskwa 1978.

Akceleratory

Ryszard Sosnowski

Akceleratorami nazywamy urządzenia, które pozwalają wytwarzać strumienie cząstek obdarzonych znaczną energią kinetyczną. Strumienie rozpędzonych cząstek służą fizykom do badania bardzo małych obiektów, takich jak jądra atomów lub cząstki elementarne. Obecnie coraz częściej wykorzystuje się akceleratory na potrzeby techniki oraz terapii medycznej.

akcelerator — narzędzie badań jądrowych

Pierwsze akceleratory cząstek zbudowano ok. 1930 r. Zapoczątkowały one nowy okres w fizyce jądrowej. Z ich pomocą stała się możliwa „alchemia XX wieku" — przeprowadzanie reakcji jądrowych. Rozpędzone cząstki, uderzając w jądra ustawionych na ich drodze tarcz, zamieniały je w jądra innych pierwiastków. Procesy te badano już wcześniej, wykorzystując cząstki emitowane przez izotopy radioaktywne. Dopiero jednak zbudowanie akceleratorów pozwoliło na systematyczne badanie reakcji jądrowych oraz na wytwarzanie i zbadanie własności jąder wielu nowych izotopów.

akcelerator — narzędzie badania cząstek elementarnych

Budowa alceleratorów dostarczających wiązek cząstek o coraz większej energii pozwoliła fizykom badać w laboratoriach obiekty jeszcze mniejsze niż jądra atomów — cząstki elementarne (→ Cząstki elementarne i ich oddziaływania). W latach pięćdziesiątych powstały akceleratory nadające protonom tak duże energie, że mogły one wytworzyć cząstki obserwowane uprzednio tylko w zderzeniach promieni kosmicznych (→ Promieniowanie kosmiczne): mezony i hiperony. Ponadto umożliwiły odkrycie nowych, nieobserwowanych do tego czasu antycząstek, a mianowicie: antyprotonów, antyneutronów i antyhiperonów.

Zwiększenie intensywności wiązek przyspieszonych cząstek oraz ich energii pozwoliło w latach sześćdziesiątych na znaczne powiększenie rodziny cząstek elementarnych o tzw. stany rezonansowe. Są to cząstki o tak krótkim czasie życia, że rozpadają się niemal natychmiast (tzn. po czasie ok. 10^{-23} s) po swym powstaniu. Dalszy rozwój techniki akceleracji doprowadził do odkrycia w latach siedemdziesiątych nowej rodziny cząstek elementarnych obdarzonych tzw. powabem. Nie jest przesadą twierdzić, że bez akceleratorów nie potrafilibyśmy zbadać ani własności jąder atomowych, ani poznać świata najmniejszych znanych obecnie obiektów fizycznych — cząstek elementarnych.

Akceleratory elektrostatyczne

Najprostszym akceleratorem cząstek naładowanych jest układ dwóch elektrod, między którymi istnieje różnica potencjału elektrostatycznego. Cząstka naładowana, która znajduje się w obszarze między elektrodami jest odpychana przez elektrodę o tym samym znaku ładunku a przyciągana przez elektrodę naładowaną przeciwnie. Aby cząstka została bez przeszkód przyspieszona między elektrodami musi istnieć próżnia taka, by średnia droga na zderzenie przyspieszanej cząstki z cząsteczkami gazu była znacznie większa od odległości między elektrodami. Wędrując od jednej elektrody do drugiej cząstka uzyskuje energię kinetyczną, E, proporcjonalną do różnicy potencjałów między elektrodami, V, oraz do wielkości ładunku elektrycznego tej cząstki, e,

$$E = eV.$$

Z tego właśnie względu przyjęto określać energię przyspieszonych cząstek przez podanie różnicy potencjałów, które należałoby zastosować do uzyskania tej energii, gdyby cząstka miała ładunek jednostkowy. Różnicę potencjałów wyrażamy zazwyczaj w woltach, a ładunek jest zawsze wielokrotnością ładunku elementarnego, tzn. takiego, którym jest obdarzony np. elektron. Stąd powszechnie używaną jednostką energii jest elektronowolt (eV), a jednostki pochodne to — kiloelektronowolt (keV), megaelektronowolt (MeV), gigaelektronowolt (GeV) i teraelektronowolt (TeV),

elektronowolt

$$1\,\text{TeV} = 10^3\,\text{GeV} = 10^6\,\text{MeV} = 10^9\,\text{keV} = 10^{12}\,\text{eV}.$$

Energię jednego elektronowolta uzyskuje cząstka o ładunku takim jaki ma elektron, przyspieszona różnicą potencjału równą 1 wolt,

$$1\,\text{eV} \approx 1{,}6 \cdot 10^{-19}\,\text{J}.$$

rurki Crooksa

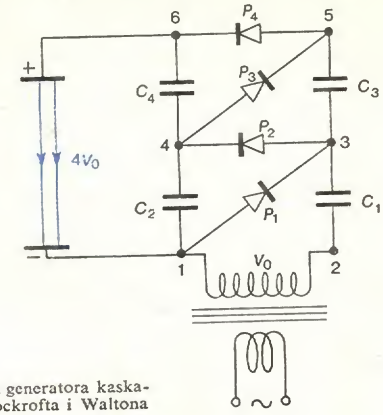
Prototypem pierwszych akceleratorów elektrostatycznych były rurki Crooksa, w których — w drugiej połowie ubiegłego wieku — wytworzono promienie katodowe, będące wiązkami przyspieszonych elektronów. Również lampy rentgenowskie i wszelkie lampy elektronowe, są małymi akceleratorami elektronów. Jednakże, aby badać własności jąder atomowych należało użyć cząstek o energiach znacznie przewyższających energie uzyskiwane w lampach rentgenowskich. Ponadto jako pocisków należało użyć nie elektronów lecz protonów.

Akceleratory typu Cockrofta–Waltona

J.D. Cockroft i E.T.S. Walton w 1932 r. opublikowali opis zbudowanego przez nich akceleratora protonów. Protony były przyspieszane różnicą potencjałów 600 kV.

W celu uzyskania tak wysokiego napięcia Cockroft i Walton zastosowali układ nazywany generatorem kaskadowym, którego schemat pokazany jest na rys. 1. Transformator prądu zmiennego wytwarza na końcówkach uzwojenia wtórnego zmienne napięcie o amplitudzie V_0. Gdy końcówka 1 jest biegunem

generator kaskadowy



Rys. 1. Schemat generatora kaskadowego typu Cockrofta i Waltona

dodatnim a końcówka 2 biegunem ujemnym, wówczas kondensator C_1 ładuje się do napięcia V_0, gdyż punkty 1 i 3 są zwarte przez prostownik P_1. W drugiej połówce okresu końcówka 1 jest biegunem ujemnym, lecz prostownik P_1 nie pozwala dodatniemu ładunkowi odpłynąć z puntu 3 do punktu 1. Transformator i kondensator C_1 są wtedy połączone szeregowo, wytwarzając na okładkach kondensatora C_2 napięcie $2V_0$. W następnej połówce okresu znowu końcówka 1 jest biegunem dodatnim i wtórne uzwojenie transformatora jest połączone szeregowo z kondensatorem C_2, wytwarzając między punktem 5 i punktem 2 napięcie $3V_0$. W kolejnej połówce półokresu połączone szeregowo kondensatory C_1 i C_3 oraz wtórne uzwojenie transformatora wytwarza między punktem 1 i punktem 6 napięcie $4V_0$.

Zastosowanie odpowiedniej liczby prostowników i kondensatorów pozwala na uzyskanie większego zwielokrotnienia napięcia. Nie możemy jednak zwiększać napięcia nieograniczenie. Ze względu na wyładowania, które powodują ucieczkę ładunku elektrycznego z wyższych stopni kaskady, generatory wysokiego napięcia typu Cockrofta–Waltona dostarczają napięcia nie większego niż 3 MV. Ucieczkę ładunku można zmniejszyć, zamykając całe urządzenie w pojemniku, w którym zwiększa się ciśnienie gazu. W ten sposób uzyskiwano napięcie do 6 MV.

zastosowanie akceleratorów Cockrofta–Waltona

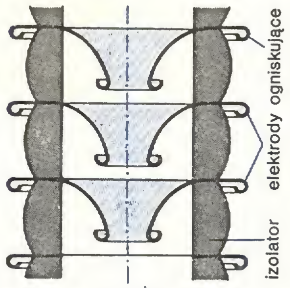
Wadą akceleratorów Cockrofta–Waltona jest mała stabilność uzyskiwanego napięcia i dlatego nie znalazły one większego zastosowania do badania jąder atomowych. Używa się ich natomiast jako pierwszego stopnia przyspieszenia cząstek, które są następnie wprowadzane do akceleratorów cyklicznych wielkich energii. Stosunkowo duża moc prądu wiązki uzyskiwana za pomocą akceleratorów kaskadowych pozwala na uzyskanie intensywnych wiązek. Na il. 39 (tabl. 11) jest pokazany generator typu Cockrofta–Waltona, w którym protony są przyspieszane do energii 10 MeV. Następnie są one wprowadzane do największego obecnie akceleratora, synchrotronu Narodowego Laboratorium Akceleratorowego im. H. Fermiego w Batavii i uzyskują energię do 500 GeV.

Wytworzona przez generator wysokiego napięcia różnica potencjałów wykorzystywana jest do przyspieszenia jonów. Następuje to w tzw. rurze przyspieszającej, w której panuje wysoka próżnia. Możliwie dokładne usunięcie z rury gazu jest niezbędne z dwóch powodów. Po pierwsze, gaz nie powinien stwarzać przeszkód na drodze przyspieszanych jonów. Po drugie, cząsteczki zjonizowanego gazu w rurze akceleracyjnej powodowałyby odpływ ładunku z elektrody wysokiego napięcia. Zmniejszenie ucieczki ładunku przez gaz jest możliwe albo za pomocą zwiększenia ciśnienia gazu, albo przez możliwie dokładne jego usunięcie.

źródło cząstek

Ważnym elementem każdego akceleratora jest źródło cząstek, które mają być przyspieszane. Na ogół są to jony określonego typu. Najczęściej stosowano protony, deuterony i cząstki α. Ostatnio jednak przyspiesza się również ciężkie jony (→ Fizyka ciężkich jonów). Zadaniem źródła jonów jest wytworzenie i dostarczenie tych jonów — żądanego rodzaju i ilości — w pobliże elektrody wysokiego napięcia. Bardzo często ważne jest, aby energia jonów po przyspieszeniu była dobrze określona, tzn. aby wszystkie jony z wiązki wychodzącej z akceleratora miały tę samą energię. W tym wypadku oczywiście rozrzut energii jonów przed przyspieszeniem musi być mały.

Jony dostarczone przez źródło są przyspieszane wzdłuż rury akceleracyjnej, którą wykonuje się z dobrych izolatorów, aby nie powodować upływu ładunku, co zmniejszałoby wartość napięcia. Wzdłuż rury umieszcza się wiele elektrod pośrednich. Ich



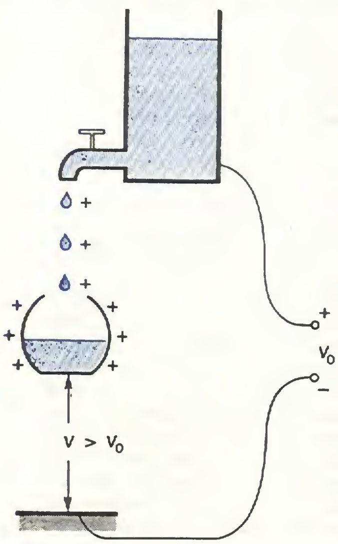
Rys. 2. Fragment rury akceleratora elektrostatycznego. Elektrody pośrednie zapewniają równomierny rozkład pola wzdłuż rury oraz ogniskują cząstki wzdłuż osi symetrii rury

zadaniem jest ogniskowanie wiązki przyspieszonych cząstek wokół osi rury. Zmniejsza to straty wiązki, zabezpiecza przed gromadzeniem się ładunków na ściankach rury (co mogłoby zakłócić proces przyspieszania) oraz zapewnia mały wymiar poprzeczny wiązki na wyjściu z akceleratora, ułatwiający wykorzystanie przyspieszonych cząstek (rys. 2).

Akceleratory z generatorami Van de Graaffa

generator Van de Graaffa

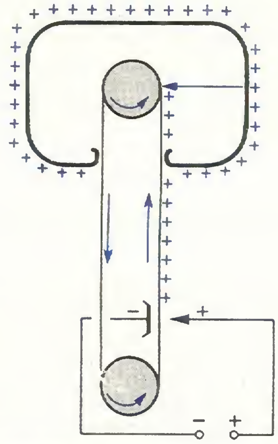
Nieomal w tym samym czasie, gdy Cockroft i Walton zbudowali swój pierwszy akcelerator, R.J. Van de Graaff skonstruował generator wysokiego napięcia działający na odmiennej zasadzie. Prototypem tego generatora było urządzenie, którego pomysł poddał lord Kelvin. Pozwalało ono na uzyskanie wysokiego napięcia za pomocą źródła o znacznie niższym napięciu. Zasada działania tego urządzenia pokazana jest na rys. 3. Spadające do wnętrza puszki Faradaya krople wody przenoszą do niej ładunek elektryczny,



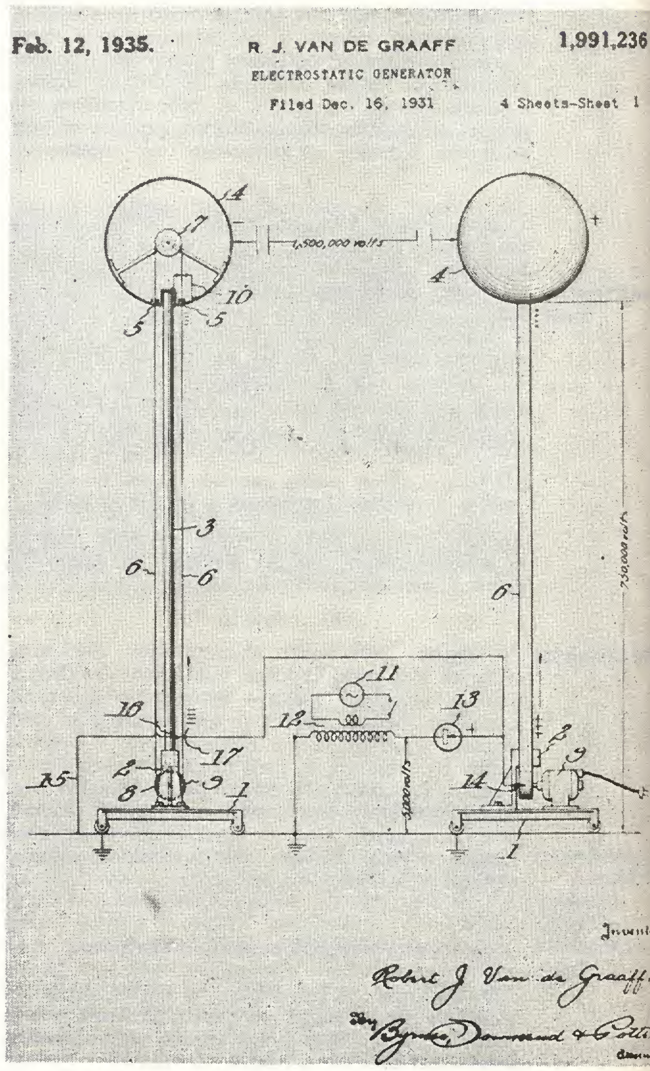
Rys. 3. Zasada generatora Van de Graaffa. Przekazywany do puszki Faradaya ładunek przenoszony przez krople wody pozwala uzyskać różnicę potencjałów znacznie wyższą niż wytwarzana przez generator połączony ze zbiornikiem wody

jeżeli zbiornik z wodą ma potencjał elektryczny względem ziemi. W rezultacie puszka może uzyskać taki ładunek, że jej potencjał znacznie przewyższy potencjał zbiornika z wodą.

Generator wysokiego napięcia skonstruowany przez Van de Graaffa działał na podobnej zasadzie. W urządzeniu tym generator dający stosunkowo



Rys. 4. Schemat generatora Van de Graaffa



Rys. 5. Rysunek ze zgłoszenia patentowego generatora Van de Graaffa

niskie napięcie ładował powierzchnię pasa jedwabnego, który przenosił ładunek do wnętrza metalowej kuli, spełniającej rolę puszki Faradaya (rys. 4). Tam ładunek był zbierany z pasa za pomocą szczotki metalowej.

podwójny generator Van de Graaffa

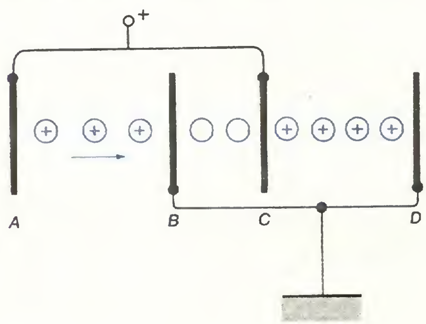
W 1931 r. Van de Graaff zbudował podwójny generator składający się z dwóch identycznych urządzeń, z których jedno generowało potencjał dodatni, a drugie ujemny. Każda z elektrod osiągała napięcie 750 kV względem ziemi, a więc napięcie między elektrodami wynosiło 1,5 MV. Na rys. 5 jest pokazane zgłoszenie patentowe tego urządzenia.

Pierwszy akcelerator, w którym generator Van de Graaffa wyposażony był w rurę akceleracyjną i źródło jonów przedstawiono na il. 34 (tabl. 10). Dostarczał on wiązki protonów i deuteronów o energii 0,6 MeV.

Współcześnie działające generatory Van de Graaffa przekraczają napięcie 20 MV, a budowany w Daresbury generator ma osiągnąć napięcie 30 MV. Przy tak dużych napięciach urządzenia te muszą być zamknięte w pojemnikach, w których panuje wysokie ciśnienie gazu (ok. 500 hP). Do tego celu używa się gazów o szczególnie dobrych własnościach izolacyjnych, np. SF_6. Wielkość uzyskiwanego napięcia zależy od tego, jak szybko można uzupełniać uciekający ładunek. Do transportu ładunku używa się więc pasów poruszających się ze znaczną prędkością — ok. 1 km/min. Przenoszony przez te pasy prąd ładowania wynosi 100–500 μA.

Akcelaratory typu tandem

Już w 1932 r. A.J. Dempster, stosując różnice potencjałów 22,5 kV, otrzymał protony o energii 45 keV. Schemat urządzenia Dempstera jest pokazany na



Rys. 6. Zasada działania akceleratora typu tandem zastosowana przez Dempstera

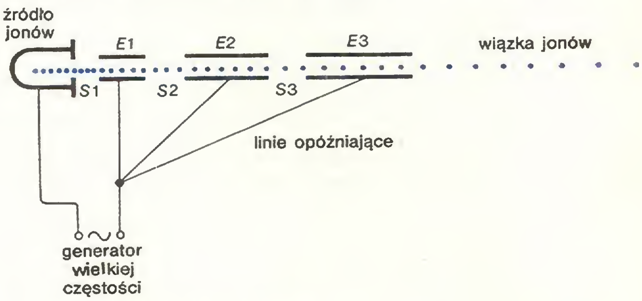
rys. 6. Protony są przyspieszane różnicą potencjałów między elektrodą dodatnią A i uziemioną B. W pobliżu elektrody B część protonów wychwytuje elektrony i już jako neutralne atomy wodoru przebywają przestrzeń BC. Po wyminięciu elektrody C znaczna część obojętnych atomów w zderzeniu z resztkami wodoru traci elektrony i jako jony dodatnie jest przyspieszana różnicą potencjałów panującą między C *i* D.

Zasada zmiany ładunku przyspieszanych jonów została wykorzystana we współczesnych akceleratorach typu tandem. Powstały one w końcu lat pięćdziesiątych mimo, że zasada ich działania była sformułowana znacznie wcześniej. W.H. Bennett przedstawił ją jako zgłoszenie patentowe w 1935 r. Dopiero jednak opracowanie źródeł jonów ujemnych oraz opanowanie techniki zamiany ich na jony dodatnie pozwoliło budować akceleratory typu tandem. W akceleratorach tych jony są przyspieszane w dwóch etapach. Ujemnie naładowane jony są przyspieszane w polu między uziemionym źródłem jonów i elektrodą naładowaną dodatnio. Po osiągnięciu elektrody jony przechodzą przez obszar o nieco wyższym ciśnieniu gazu lub przez cienką folię. W wyniku zderzeń część jonów traci pewną ilość elektronów i staje się jonami dodatnimi. Wiązka dodatnich jonów jest następnie przyspieszana potencjałem dodatnim elektrody.

Akceleratory typu tandem znalazły powszechne zastosowanie w fizyce jądrowej. Pozwalają one uzyskać wiązki różnorodnych jonów o energiach do paru dziesiątków MeV. Ponadto wiązki te mają bardzo dobrze określoną energię. Względny rozrzut energii jonów wynosi 10^{-5} energii średniej.

Akceleratory liniowe

W omawianych dotychczas akceleratorach wykorzystywano stałe, niezmieniające się w czasie pole elektrostatyczne. Już jednak w 1924 r. Ising opublikował projekt, w którym pole przyspieszające pojawiało się



Rys. 7. Schemat akceleratora liniowego

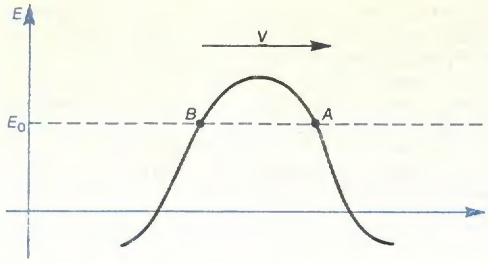
wykorzystanie zmiennego napięcia

w poszczególnych przerwach między elektrodami tylko wtedy, gdy docierała do tej przerwy grupa przyspieszanych jonów. Na rys. 7 pokazano schemat takiego akceleratora. Generator zmiennego napięcia połączony jest z elektrodami E_1, E_2 i E_3 przez linie opóźniające. W obszarach S_1, S_2 i S_3 powstaje zmienne pole elektryczne. Długość elektrod E jest tak dobrana, aby w czasie, gdy w obszarach S występuje pole o znaku przeciwnym niż potrzebny do przyspieszenia, jony znajdowały się wewnątrz cylindrycznych elektrod E. Ze względu na uzyskiwaną przez jony coraz większą prędkość każda następna elektroda jest dłuższa niż poprzednia. W akceleratorze pokazanym na schemacie jony są przyspieszane trzykrotnie. Akcelerator ten dostarcza cząstek nie w postaci nieprzerwanego strumienia, lecz jako krótkotrwałe impulsy. Akcelerator pracujący na tej zasadzie został po raz pierwszy zbudowany przez Wideröe. Za pomocą generatora o napięciu 20 kV uzyskano jony potasu o energii 40 keV. Jednakże rozwój akceleratorów liniowych nastąpił dopiero z chwilą opanowania techniki mikrofal i generatorów wysokiej częstości. Technika ta przeżyła gwałtowny rozwój w czasach II wojny światowej dzięki wprowadzeniu radaru.

zasada samoogniskowania się cząstek

Drugim ważnym czynnikiem umożliwiającym budowę akceleratorów liniowych było odkrycie zasady samoogniskowania się cząstek. Współczesny akcelerator liniowy działa w ten sposób, że elektryczne pole przyspieszające, równoległe do osi rury akceleracyjnej, jest wytworzone w niewielkim stosunkowo obszarze, który przesuwa się wzdłuż akceleratora z tą samą prędkością z jaką biegną przyspieszone cząstki. Na pozór realizacja takiego procesu wydaje się wymagać niesłychanie precyzyjnego uregulowania prędkości początkowej cząstek, wartości natężenia pola oraz prędkości przesuwania się tego pola. W praktyce okazuje się jednak, że dochodzi do automatycznego dopasowania się tych wielkości, ułatwiającego uzyskanie efektu przyspieszenia.

Na rys. 8 przedstawiono rozkład natężenia pola elektrycznego przyspieszającego cząstki. Pole to przesuwa się z pewną prędkością wzdłuż rury akceleratora wraz z cząstkami, które w tym polu uzyskują



Rys. 8. Zasada samodopasowania wartości pola przyspieszającego w akceleratorze liniowym — E_0 i prędkości jego przesuwania się wzdłuż rury akceleracyjnej — V

przyspieszenie. Oznaczmy przez E_0 tę wartość pola, która zapewnia takie przyspieszenie cząstek, że ich prędkość jest zawsze identyczna z prędkością przesuwania się pola. Wartość E_0 występuje w dwóch punktach A i B. Załóżmy, że jakaś cząstka zamiast znaleźć się w punkcie A opóźni się i pole zacznie ją wyprzedzać. Wówczas podlega ona działaniu nieco silniejszego pola, prędkość jej wzrośnie szybciej niż innych cząstek i wkrótce opóźniająca się cząstka dogoni lub nawet przegoni pole przyspieszające. W wypadku uzyskania zbyt dużego przyspieszenia i wyprzedzenia pola cząstka znajdzie się w obszarze pola słabszego i będzie przyspieszana mniej efektywnie. Pozwoli to z kolei przesuwającemu się polu elektrycznemu dopędzić uciekającą cząstkę. Widzimy więc, że w pobliżu punktu A występuje mechanizm samoczynnego dopasowania się prędkości cząstek do wartości pola przyspieszającego i do prędkości przemieszczania się tego pola. Warto zauważyć, że w pobliżu punktu B mechanizm samodopasowania nie występuje.

zestosowanie akceleratorów liniowych

Akceleratory liniowe są używane zarówno do przyspieszania protonów jak i elektronów. Największy akcelerator protonów pracujący w Los Alamos osiąga energię 800 MeV. Największym akceleratorem liniowym przyspieszającym elektrony jest liniowy akcelerator w Stanford, który przyspiesza elektrony do energii 20 GeV.

Akceleratory cykliczne

Poza akceleratorami, w których tory przyspieszanych cząstek biegną w przybliżeniu po liniach prostych, istnieje wiele typów akceleratorów o torach zbliżonych do okręgów. W akceleratorach tych przyspieszane cząstki biegną w polu magnetycznym zakrzywiającym ich tory. Dzięki temu można znacznie wydłużyć drogę, na której cząstki doznają przyspieszenia, bez konieczności budowy bardzo długich rur akceleracyjnych. W czasie pokonywania tej drogi cząstki doznają drobnych, lecz wielokrotnych przyrostów energii, co powoduje że całkowity przyrost energii jest znaczny. Akceleratory tego typu nazwano akceleratorami cyklicznymi.

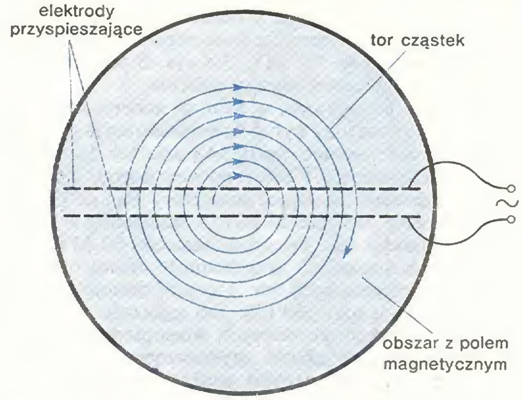
Cyklotron

wykorzystanie pola magnetycznego

Cyklotron jest najstarszym akceleratorem cyklicznym. Zasada jego działania polega na tym, że cząstka naładowana poruszająca się w jednorodnym polu magnetycznym prostopadle do linii pola biegnie po torze kołowym, a czas obiegu wynosi

$$T = \frac{2r}{v} = \frac{2mc}{eH},$$

gdzie m — masa cząstki, c — prędkość światła, e — ładunek elektryczny cząstki, H — natężenie pola magnetycznego, r — promień toru, v — prędkość cząstki. Jak widzimy okres obiegu cząstki nie zależy od jej energii, jeśli pominiemy relatywistyczny wzrost masy. Zasadę działania cyklotronu pokazano na rys. 9. Wzdłuż średnicy toru cząstki ustawione są dwie elektrody, między którymi jest przyłożone zmienne pole elektryczne o częstości równej $1/T$. Naładowana cząstka, która w pewnej chwili znajduje



Rys. 9. Schemat działania cyklotronu

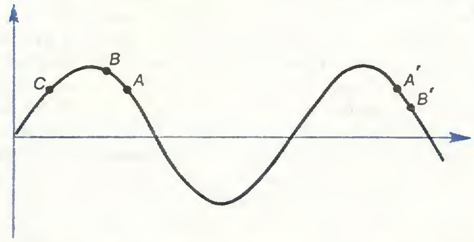
się między elektrodami, zostanie przyspieszona i zatoczywszy półokrąg po czasie $T/2$, znajdzie się ponownie w obszarze między elektrodami. Natrafi wówczas na pole, które jest skierowane zgodnie z jej ruchem i w związku z tym cząstka ta ulegnie dalszemu przyspieszeniu. Promień jej toru zwiększy się. Po czasie $T/2$ cząstka zatoczy następne pół okręgu i ulegnie dalszemu przyspieszeniu. W ten sposób, dzięki wielokrotnemu przechodzeniu przez obszar przyspieszający, tor cząstki dojdzie do granicy pola jednorodnego i dalszy wzrost energii cząstki przestanie być możliwy.

Synchrocyklotron

Możliwość kontynuowania przyspieszania cząstek nie jest ograniczona jedynie skończonymi rozmiarami pola magnetycznego. Ze względu na relatywistyczny wzrost masy cząstki, w miarę wzrostu jej energii okres obiegu po orbicie kołowej wydłuża się i cząstka zaczyna się pojawiać w obszarze między elektrodami zbyt późno, aby trafić na pole przyspieszające.

wykorzystanie zmiany częstości pola

W celu uzyskania cząstek o energiach większych niż to jest możliwe w cyklotronach, zaczęto budować tzw. synchrocyklotrony. Są to akceleratory działające w podobny sposób jak cyklotrony, lecz w tym wypadku częstość pola przyspieszającego maleje w miarę, jak przyspieszone cząstki uzyskują energię i okres ich obiegu po torze kołowym wzrasta. Oczywiście w odróżnieniu od cyklotronów, w których możliwe jest jednoczesne przyspieszanie cząstek w całym obszarze pola magnetycznego, synchrocyklotrony przyspieszają w danej chwili tylko te cząstki, których energia odpowiada aktualnej częstości pola przyspieszającego. Wykorzystując zasadę synchrocyklotronu, uzyskano wiązki protonów o energii do 700 MeV.



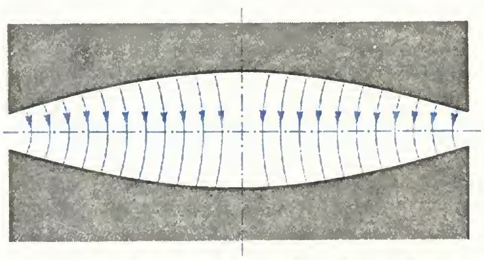
Rys. 10. Zasada samodopasowania w synchrocyklotronie

samoogniskowanie fazy cząstek

W konstrukcji synchrocyklotronów ważną rolę odgrywa zasada samoogniskowania fazy cząstek omawiana już przy opisie akceleratorów liniowych. Zasada ta działa tutaj następująco. Niech między elektrodami przyspieszającymi synchrocyklotronu panuje zmienne pole elektryczne o wolno malejącej częstości. Przebieg jego amplitudy przedstawiono na rys. 10. Punktem A jest oznaczona wartość pola konieczna do uzyskania takiego wzrostu energii cząstki, by po czasie T natrafiła ona na pole A' w celu dalszego przyspieszenia. Jeżeli jednak zamiast w punkcie A cząstka znajdzie się w polu przyspieszającym wcześniej (punkt B), to uzyska zbyt dużo energii. Ze względu na relatywistyczny wzrost masy okres jej obiegu wydłuży się i w następnym cyklu przyspieszania zostanie ona poddana działaniu słabszego pola (punkt B'). W ten sposób zbyt silne przyspieszenie w jednym cyklu zostaje skompensowane słabszym przyspieszeniem w cyklu następnym i odwrotnie. Warunkiem wystąpienia takiego samoregulującego się mechanizmu jest pojawienie się cząstek w obszarze pola, gdy pole to maleje (obszar wokół punktu A), a nie gdy rośnie (obszar wokół punktu C). Należy zauważyć, że sytuacja w synchrocyklotronie jest odmienna niż w akceleratorze liniowym. Tam cząstki opóźniające się, a więc wolniejsze, powinny podlegać działaniu silniejszego pola przyspieszającego. W synchrocyklotronie natomiast spóźniają się cząstki szybsze o zbyt dużej energii.

Cyklotrony izochroniczne

Jak wynika ze wzoru na okres obiegu cząstki po orbicie kołowej, okres ten można przy wzroście masy utrzymać stały, gdy jednocześnie zwiększy się wartość natężenia pola magnetycznego. Ponieważ cząstki

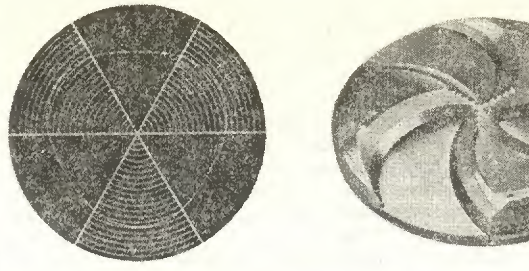


Rys. 11. Magnes dający pole magnetyczne o natężeniu rosnącym w kierunku zewnętrznym. Pole takie nie może być użyte w cyklotronie, gdyż nie ogniskuje cząstek w pobliżu poziomej płaszczyzny symetrii szczeliny magnesu

o większej energii poruszają się po torach o większym promieniu, mogłoby się wydawać, że pole magnetyczne magnesu (pokazanego na rys. 11) może zapewnić możliwość przyspieszania cząstek przy stałej częstości zmiennego pola elektrycznego. W miarę uzyskiwania coraz większej energii, a więc i wzrostu relatywistycznego masy cząstki, porusza się ona po torach biegnących w obszarze silniejszego pola magnetycznego. Niestety, pola magnetycznego pokazanego na rys. 11 nie można wykorzystać w cyklotronie, ponieważ charakteryzuje się ono tym, że tylko cząstki poruszające się w idealnie poziomej płaszczyźnie symetrii szczeliny między nabiegunnikami pozostaną w tej płaszczyźnie. Każde najdrobniejsze odchylenie od tej płaszczyzny spowoduje, że cząstki będą się od niej stopniowo oddalać i uderzą w nabiegunnik magnesu.

Aby temu przeciwdziałać, w cyklotronach izochronicznych stosuje się pole magnetyczne, którego natężenie uśrednione po całej orbicie kołowej wzrasta wraz z odległością od środka cyklotronu. Nabiegunniki nie mają jednak symetrii cylindrycznej jak w cyklotronach i synchrocyklotronach. Biegnące cząstki natrafią kolejno na obszary słabszego i silniejszego

wpływ kształtu nabiegunników



Rys. 12. Przykłady kształtu nabiegunników magnesów stosowanych w cyklotronach izochronicznych

pola magnetycznego, co uniemożliwia im oddalanie się od płaszczyzny symetrii szczeliny między nabiegunnikami. Na rys. 12 pokazane są nabiegunniki stosowane w cyklotronach izochronicznych.

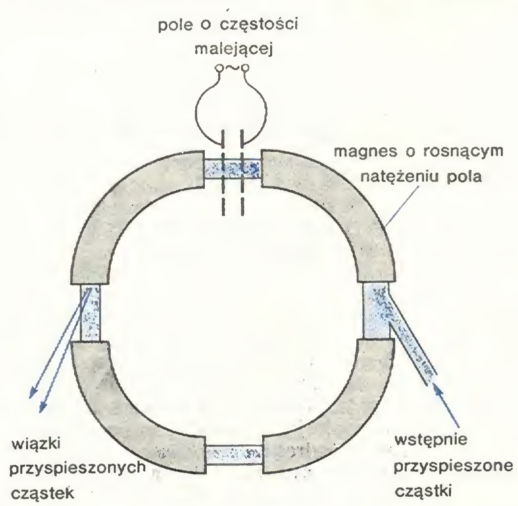
Za pomocą cyklotronów izochronicznych można uzyskiwać cząstki o energiach znacznie większych niż w cyklotronach klasycznych. W ten sposób uzyskano protony o energii kilkuset MeV. Ograniczenie wielkości otrzymywanych energii za pomocą cyklotronów izochronicznych wynika z faktu, że dla dużych energii wzrostowi pędu cząstki towarzyszy bardzo mały wzrost jej prędkości. Należy więc poza cyklotronem utrzymywać pole rosnące bardzo szybko wraz z promieniem. W odległości

$$r_\infty = \frac{m_0 c^2}{eH_0}$$

natężenie pola magnetycznego powinno osiągnąć wartość nieskończoną. We wzorze tym m_0 oznacza masę spoczynkową cząstki, a H_0 — natężenie pola magnetycznego w pobliżu osi nabiegunnika. Uzyskanie pola magnetycznego zmieniającego się gwałtownie z promieniem orbity jest prawie niemożliwe.

Synchrotron

W odróżnieniu od cyklotronu klasycznego, czy nawet izochronicznego, synchrocyklotron może teoretycznie dostarczać cząstki o nieograniczenie dużej energii. Przeszkodą jest jednak zbyt duży koszt jego budowy. Najkosztowniejszym jego elementem jest magnes, który musi wytwarzać prawie jednorodne pole magnetyczne w bardzo dużym obszarze, obejmującym zarówno tory cząstek powolnych jak i tych w końcowej fazie przyspieszenia. Aby uniknąć wysokich kosztów magnesu, buduje się akceleratory, w których przy-



Rys. 13. Schemat synchrotronu ze „słabym ogniskowaniem"

spieszane cząstki krążą po stałych orbitach. Dopóki cząstki mają małą energię, elektromagnesy wytwarzają pole magnetyczne o niewielkim natężeniu. Natężenie to rośnie w miarę uzyskiwania przez cząstki coraz wyższej energii. W ten sposób czastki biegną w akceleratorze po stałych, w przybliżeniu, torach i tylko w tym obszarze potrzebne jest pole magnetyczne. Przyspieszanie cząstek odbywa się podobnie jak w cyklotronach za pomocą zmiennego pola elektrycznego. Częstość tego pola rośnie w miarę, jak rośnie prędkość cząstek.

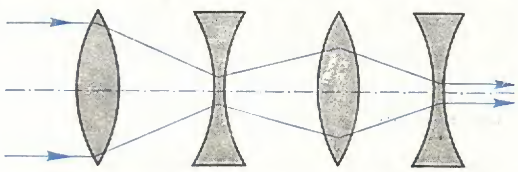
Schemat synchrotronu pokazany jest na rys. 13. Wstępnie przyspieszone cząstki wprowadzane są do synchrotronu, gdzie ich tor składa się z czterech ćwiartek okręgu połączonych odcinkami prostoliniowymi. Rosnące w procesie akceleracji pole zlokalizowane jest tylko w obszarze, gdzie tory się zakrzywiają. Przyspieszanie cząstek odbywa się na jednym z prostych odcinków. Po uzyskaniu żądanej energii cząstki są na ogół wyprowadzane z akceleratora. W tym celu na pewnym odcinku ich orbity pojawia się pole magnetyczne skierowujące wiązkę cząstek w pożądanym kierunku.

ogniskowanie słabe

Schemat pokazany na rys. 13 przedstawia synchrotron starszego typu z tzw. „słabym" ogniskowaniem. W akceleratorach tego typu tory przyspieszanych cząstek mogły się znacznie odchylać od średniego toru. Zmuszało to konstruktorów akceleratora do stosowania dużych magnesów, a próżniowy kanał akceleratora musiał mieć dużą średnicę. Odkrycie zasady silnego ogniskowania umożliwiło stosowanie kanałów o średnicy paru centymetrów i tylko w tym obszarze powinno się wytworzyć pole magnetyczne.

ogniskowanie silne

Zasada silnego ogniskowania polega na zastosowaniu szeregu magnesów, które na przemian skupiają i rozpraszają wiązkę przyspieszanych cząstek. W rezultacie, średnica wiązki zostaje zmniejszona, podobnie jak



Rys. 14. Analogia optyczna zasady „silnego ogniskowania" cząstek

zmniejszona jest średnica wiązki świetlnej (pokazanej na rys. 14) na skutek przechodzenia jej przez soczewki skupiające i rozpraszające. Synchrotrony ze słabym ogniskowaniem pozwalały na przyspieszanie protonów do energii ok. 10 GeV. Zastosowanie magnesów silnie ogniskujących pozwoliło osiągnąć znacznie wyższe energie. Supersynchrotron protonowy w Europejskim Centrum Badań Jądrowych w Genewie przyspiesza protony do energii 450 GeV, a zbudowany wcześniej w Batavii pod Chicago akcelerator protonów dostarcza wiązek o energii do 500 GeV (il. 40, tabl. 11).

Układy zderzających się wiązek

Przy badaniu zderzeń o wysokich energiach ważne jest, aby uzyskać jak największą energię w układzie środka masy dwóch zderzających się cząstek. Jeżeli cząstka o masie m_1 spoczywa w laboratorium a uderzająca w nią cząstka o masie m_2 ma energię E, to energia zderzenia w układzie środka masy tych cząstek E^* wyniesie

$$E^* = \sqrt{m_1^2+m_2^2+2Em_1}.$$

Jeżeliby jednak obu cząstkom nadać energię E i skierować jedną przeciwko drugiej, to energia E^* wyniosłaby w tym przypadku

$$E^* = 2E.$$

protonowe wiązki przeciwbieżne

W ten sposób np. dwie wiązki protonów, każda o energii 20 GeV będą wywoływać zderzenia o energii E^* wynoszącej 40 GeV. Ażeby tę energię uzyskać w warunkach, gdy przyspieszony zostaje tylko jeden proton a drugi spoczywa, należałoby dysponować akceleratorem na energię 800 GeV. Zasadę zderzania ze sobą dwóch protonowych wiązek przeciwbieżnych zastosowano w Europejskim Centrum Badań Jądrowych w Genewie w celu badania zderzeń o najwyższych wytworzonych przez człowieka energiach. Dwie intensywne wiązki protonów o natężeniu 20 amperów każda krążą po przeciwbieżnych torach w wysokiej próżni. „Czas życia" takiej wiązki wynosi kilkadziesiąt godzin. W miejscu, gdzie te wiązki się spotykają, zachodzą zderzenia dwóch biegnących naprzeciw siebie protonów. Energia każdej z wiązek może osiągnąć do 30 GeV, co daje $E^* = 60$ GeV. Dla osiągnięcia takiej energii w klasycznych akceleratorach należałoby przyspieszyć protony do energii 1800 GeV.

elektronowe wiązki przeciwbieżne

Dwie wiązki przeciwbieżne stosuje się także przy badaniu zderzeń elektronów z pozytonami. Jest to zresztą jedyny sposób badania tych zderzeń. Istnieje w tej chwili na świecie sześć akceleratorów połączonych z układami zderzających się wiązek pozytonowo-elektronowych. Dwa ostatnio zbudowane układy umożliwiają badanie zderzeń elektronowo-pozytonowych przy energii każdej z wiązek dochodzącej do 20 GeV.

W.S. Scharf *Akceleratory cząstek naładowanych i ich zastosowania*, Warszawa 1978.

Oddziaływania silne

Grzegorz Białkowski

Oddziaływania silne cząstek trwałych i nietrwałych

Oddziaływania silne są jednym z typów oddziaływań między cząstkami elementarnymi (→ Cząstki elementarne i ich oddziaływania). Biorą w nich udział tylko hadrony. Istnienie tych oddziaływań zostało po raz pierwszy dostrzeżone przy próbach wyjaśnienia budowy i własności jąder atomowych. Początkowo zjawiska wewnątrzjądrowe próbowano zrozumieć stosując wyłącznie prawa elektrodynamiki. Po wykryciu neutronu przez Jamesa Chadwicka w 1932 r. stało się oczywiste, że są one do tego niewystarczające.

hipoteza istnienia mezonu

Na właściwą myśl wpadł Hideki Yukawa, który zaproponował, aby do opisu sił jądrowych wprowadzić nową cząstkę elementarną, w owym czasie jeszcze nie zaobserwowaną w doświadczeniu, którą nazwano „mezonem".

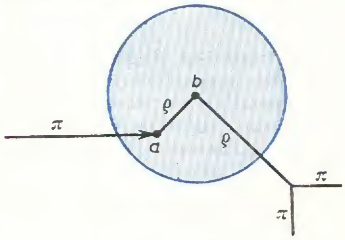
W kilkanaście lat po wypowiedzeniu przez Yukawę tej hipotezy mezony π, oddziałujące z nukleonami, zostały istotnie wykryte. Był to jednak dopiero początek lawiny odkryć, trwającej zresztą do dziś, która z fizyki oddziaływań silnych czyni dziedzinę szczególnie pociągającą, ale też i szczególnie trudną. Fakty, które zostały stwierdzone doświadczalnie, dają się wyjaśnić teoretycznie na razie tylko częściowo. Mimo to można powiedzieć śmiało, że nie-

wiele jest takich dziedzin fizyki, w których postęp w ostatnich latach byłby równie szybki, i w których zarazem tak zupełna byłaby świadomość stąpania po ziemi całkowicie nieznanej. Jest to więc dziedzina istotnie potrzebująca nowych idei, a w której tak niewiele da się przewidzieć na podstawie znanych już faktów.

Informacje eksperymentalne dotyczące oddziaływań silnych pochodzą z dwu nierównej wartości źródeł: z analizy stanów związanych (jąder atomowych i hiperjąder) oraz z analizy procesów zderzeń zachodzących między cząstkami elementarnymi. Znacznie bogatszy materiał pochodzi z tego drugiego źródła, przy czym okazuje się, że charakter danych zależy bardzo silnie od energii zderzających się obiektów.

akceleratory cząstek

Istotne znaczenie dla rozwoju badań miała więc — poza udoskonaleniem techniki analizy otrzymywanych danych — także budowa coraz to potężniejszych akceleratorów cząstek (→ Akceleratory). Urządzenia te, w których dochodzi do zderzeń między hadronami, można podzielić na dwie grupy. Do pierwszej można zaliczyć urządzenia, w których przyspieszona wiązka pada na nieruchomą tarczę, a do drugiej te, w których dochodzi do czołowych zderzeń wiązek przeciwbieżnych. Każdy z tych dwóch typów akceleratorów ma swoje zalety: w drugim wykorzystuje się znacznie wydajniej moc akceleratora, ponieważ nie traci się energii na (niemal) bezproduktywny fizycznie ruch środka masy zderzających się obiektów. Za to w pierwszym akceleratorze uzyskuje się znacznie większą wydajność przy danej gęstości wiązki, a przede wszystkim istnieje możliwość wytworzenia wiązek wtórnych, zawierających obiekty nietrwałe, których nie można przyspieszać ze względu na ich krótki czas życia. Ostatecznie, jeśli odliczyć zderzenia, w których biorą udział złożone jądra atomowe, mamy do dyspozycji dane pochodzące z oddziaływań na tarczach protonowych wiązek złożonych z protonów, neutronów, niektórych hiperonów, mezonów $\pi^{\pm}$, mezonów K^{+} i K^{-} i wreszcie mezonów K_L^0. Badanie zderzeń z nukleonami innych cząstek, jeszcze mniej trwałych, z których nie można sformować nawet wiązki wtórnej, jest umożliwione (w ograniczonym zakresie i z większą niepewnością) przez badanie kaskad wewnątrzjądrowych. Także analiza oddziaływań w stanie końcowym dostarcza pewnych informacji o oddziaływaniach cząstek nietrwałych.



Rys. 1. Oddziaływanie cząstek nietrwałych wewnątrz jąder atomowych

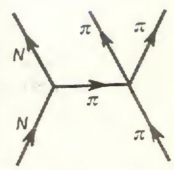
oddziaływania cząstek nietrwałych

Zasada badania oddziaływań cząstek nietrwałych w zderzeniach wewnątrzjądrowych jest przedstawiona schematycznie na rys. 1. Wiązka cząstek padających (np. mezonów π) uderza w jądro. Jedna z tych cząstek wytwarza w wyniku oddziaływania z jednym z nukleonów znajdujących się w jądrze (w punkcie a) inną cząstkę, bardzo nietrwałą (np. mezon ϱ). Ze względu na to, że odległości między nukleonami w jądrze są bardzo małe, ok. 1 fm, istnieje duże prawdopodobieństwo, że nowo wytworzona cząstka zdąży oddziałać z kolejnym nukleonem (w punkcie b) zanim się rozpadnie. Jeżeli energia tej cząstki nie jest zbyt duża, to w oddziaływaniu tym nie dojdzie do zmiany rodzaju cząstek (oddziaływanie elastyczne). Po jednym lub kilku takich oddziaływaniach cząstka ta opuści jądro i rozpadnie się. Oczywiście padający na jądro mezon π może wytworzyć mezon ϱ nie w swym pierwszym zderzeniu, ale w drugim, lub jeszcze dalszym. Na ogół jednak, znając wielkości opisujące procesy rozpraszania mezonu π na nukleonie oraz produkcji mezonu ϱ w zderzeniu mezonu π z nukleonem, możemy przystąpić do analizy danych dotyczących produkcji mezonów ϱ na jądrach mając tylko jedną wielkość nieznaną, a mianowicie amplitudę rozpraszania mezonów ϱ na nukleonach. Tę wielkość więc w zasadzie można wyznaczyć. Oczywiście postępowanie takie zawiera wiele niezbyt pewnych elementów, jednakże można się nim posłużyć do uzyskania co najmniej przybliżonych informacji o procesie, którego innymi metodami nie można byłoby w ogóle badać.

Jeśli chodzi o oddziaływanie w stanie końcowym, to dostarcza nam ono informacji o zderzeniach cząstek, które zostały właśnie wytworzone w danym procesie i które jeszcze przez jakiś czas po akcie produkcji pozostają w zasięgu sił. Takie cząstki, zanim się od siebie dostatecznie oddalą, oddziałują ze sobą, co modyfikuje np. ich rozkłady kątowe i to w sposób wyraźnie zależny od energii ruchu względnego tych obiektów. Badając owe rozkłady możemy się dowiedzieć, jak oddziaływałyby cząstki, gdybyśmy z nich umieli formować wiązki i tarcze, stosując standardowe metody badania zderzeń.

przybliżenie biegunowe

W pewnych wypadkach analizę oddziaływania wzajemnego cząstek nietrwałych można jeszcze bardziej udoskonalić, stosując metodę opartą na przybliżeniu biegunowym (znaczenie tego pojęcia zostanie wyjaśnione nieco dalej). Zasada postępowania jest przedstawiona na rys. 2. Przykładem może być proces produkcji jednego mezonu π w zderzeniu mezonu π z nukleonem. Sądzimy w tym wypadku, że główny przyczynek do amplitudy produkcji przy bardzo małych przekazach czteropędu od nukleonu początkowego do końcowego wnosi wymiana pojedynczego mezonu π, jak to przedstawia rys. 2. Gdy bowiem przekaz czteropędu jest mały, to (zgodnie z zasadą nieokreśloności) parametr zderzenia jest duży. Z drugiej strony, średni promień chmury cząstek wirtualnych otaczających nukleon jest odwrotnie proporcjonalny do masy tych cząstek. Skoro zaś najlżejszym hadronem jest mezon π, możemy sądzić, że najbardziej zewnętrzna część chmury jest tworzona przez te właśnie cząstki. Zatem w zderzeniu peryferycznym nadlatujący mezon π uderza w wirtualny mezon π znajdujący się w chmurze. Taka interpretacja umożliwia przeprowadzenie analizy procesów produkcji dodatkowego mezonu π w zderzeniu π-N oraz w zderzeniu K-N w taki sposób, aby z niej uzyskać informacje dotyczące odpowiednio oddziaływań π-π i K-π. Takie właśnie analizy są wykonywane i dostarczają nam niezwykle cennych danych, choć nie całkiem pewnych i teoretycznie uzasadnionych, głównie ze względu na istnienie poprawki odpowiadającej wymianie więcej niż jednej cząstki.



Rys. 2. Procesy produkcji cząstek w przybliżeniu biegunowym

Obszary energii dla zderzeń hadron-hadron

Jak już wspomniano, charakterystyka zderzeń zależy w dużym stopniu od energii, przy której zderzenia te zachodzą. Można wyróżnić kilka obszarów energetycznych, w których zderzenia hadron-hadron mają specyficzne cechy. (Opisany tu przebieg zależności energetycznej jest wyidealizowany; w rzeczywistości bowiem nie wszystkie cząstki zachowują się jednakowo, a poszczególne obszary energetyczne nie są od siebie wyraźnie oddzielone).

obszar skrajnie niskich energii

W obszarze skrajnie niskich energii oddziaływanie kulombowskie nawet przy stosunkowo dużych kątach rozpraszania konkuruje z oddziaływaniem silnym; w tym wypadku trudno uzyskać informacje o oddziaływaniach silnych hadronów naładowanych. Istnienie oddziaływań kulombowskich przejawia się też w tym, że bardzo powolne naładowane ujemnie hadrony są chwytane przez protony, co prowadzi od powstania specyficznych układów atomowych, zw. hadroatomami (jeśli

hadroatomy

mezoatomy

schwytaną cząstką jest mezon, taki obiekt nazywamy mezoatomem, → Atomy egzotyczne). Atomy te zachowują się podobnie do zwykłych atomów, w których na orbicie znajdują się elektrony, z tą tylko różnicą, że orbity te mają znacznie mniejsze promienie (w stosunku równym stosunkowi mas danego hadronu i elektronu). Z tych orbit hadrony naładowane ujemnie są chwytane przez jądra, przy czym dochodzi do anihilacji ładunku, a czasem i innych liczb kwantowych (np. liczby barionowej, jeśli schwytaną cząstką jest antyproton). Badanie tych procesów ma zasadnicze znaczenie dla fizyki jądra atomowego, pozwala bowiem badać jego strukturę; dla fizyki oddziaływań silnych ma ono jednak mniejsze znaczenie.

obszar bardzo niskich energii

W kolejnym obszarze, energii bardzo niskich, możemy odróżnić dwa zasadnicze przypadki. Jeżeli zderzające się cząstki nie anihilują ze sobą, to znaczy, jeśli nie istnieje stan dwu lub więcej cząstek o niższej masie lecz tych samych wszystkich liczbach kwantowych, jakie ma układ cząstek zderzających się (do układów nie anihilujacych należą układy p-p, $\pi^{\pm}$-p, K^{+}-p i Λ-p), to w obszarze bardzo niskich energii w oddziaływaniu nie powstają nowe cząstki i oddziaływanie to można bardzo łatwo opisać, wprowadzając niewielką liczbę parametrów fenomenologicznych, takich jak długość rozpraszania czy też zasięg efektywny (por. tekst dalej). Jeżeli jednak zderzające się cząstki mogą anihilować (np. dotyczy to układu p-$\bar{p}$ lub też K^{-}-p, który nawet już przy energii progowej może przejść np. w układ Λ-π^{0}), to oddziaływanie już od energii progowej ma charakter nieelastyczny, co zwykle prowadzi do istotnych komplikacji w opisie i interpretacji danych.

obszar niskich energii

Kolejny obszar można by nazwać obszarem niskich energii. W obszarze tym procesy nieelastyczne występują już nawet w tych zderzeniach, które w obszarze poprzednim miały charakter czysto elastyczny, ale zjawiska nieelastyczne odgrywają w nim jeszcze niewielką rolę. Można tu znów rozróżnić dwie grupy procesów. W jednej z nich w obszarze niskich energii dochodzi do powstania jednocząstkowych bardzo nietrwałych obiektów, które nazywa się stanami rezonansowymi układu zderzających się cząstek. Do grupy tych procesów należy m.in. rozpraszanie $\pi^{\pm}$-p, czy π^{+}-π^{-} lub K^{-}-p. Do drugiej grupy procesów należy rozpraszanie K^{+}-p czy π^{+}-π^{+} lub też p-p. Zdania co do istnienia stanów rezonansowych także i w tych układach są podzielone; w każdym razie można uznać, że jeśli one istnieją, to do ich powstania dochodzi trudniej i nie są one tak bezpośrednio widoczne jak w pierwszej grupie procesów.

obszar średnich energii

W następnym obszarze energii zwanym obszarem średnich energii, struktury rezonansowe stopniowo zanikają, rośnie natomiast nieelastyczność zderzeń. Rezonanse mają coraz to krótsze czasy życia, co utrudnia ich identyfikację. W rezultacie są one coraz mniej dostrzegalne. Nie sprzyja temu również wzrastająca liczba stanów orbitalnych, w których mogą się znaleźć cząstki silnie oddziałujące (por. dalej). Jest to obszar może najtrudniejszy z punktu widzenia analizy teoretycznej, ale też chyba i najmniej interesujący.

obszar wysokich energii

W obszarze wysokich energii obraz nieco się upraszcza. Rezonanse przestają być widoczne, natomiast można wprowadzić pewne metody specjalnie przystosowane do tego obszaru energii (np. metoda biegunów Reggego; zob. dalej). W obszarze tym silnie rośnie krotność wyprodukowanych cząstek i wkład procesów nieelastycznych znacznie przewyższa wkład procesów elastycznych.

obszar skrajnie wysokich energii

Wreszcie, w obszarze skrajnie wysokich energii, zaczynamy obserwować pewne regularności, które zgodnie z przewidywaniami teoretycznymi powinny cechować zderzenia przy energiach zmierzających do nieskończoności. Znaczna część tych regularności ma charakter uniwersalny, a więc niezależny od rodzaju zderzających się cząstek. Jest sprawą w pewnej mierze otwartą, czy już przy obecnie osiągalnych energiach zbliżono się do obszaru „asymptotycznego", a także jak dalece przewidywania teoretyczne, oparte przecież na pewnych założeniach, będą się sprawdzać w praktyce.

Wybór zmiennych kinematycznych

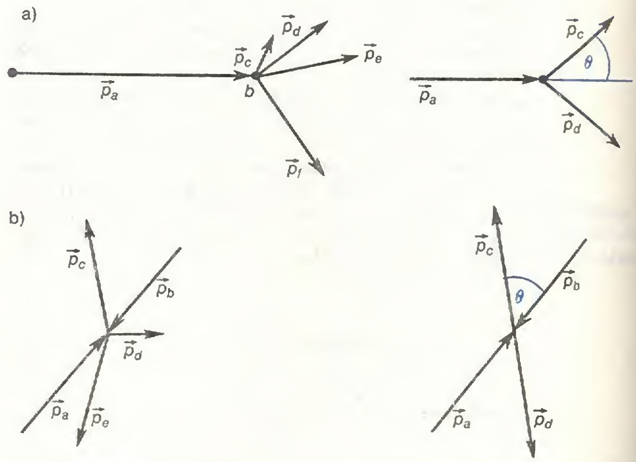
Będziemy rozważać zderzenia zachodzące między dwiema cząstkami w stanie początkowym: cząstką-pociskiem a i cząstką-tarczą b. W stanie końcowym mogą pojawić się dwie cząstki, i wtedy proces nazywa się binarnym, albo więcej cząstek i wtedy proces jest wielocząstkowy. W dalszym ciągu przez cząstkę trwałą należy rozumieć cząstkę trwałą ze względu na oddziaływania silne, a więc o czasie życia rzędu co najmniej 10^{-20} s. Może się zdarzyć, że proces wielocząstkowy zachodzi w ten sposób, że w rzeczywistości są wytwarzane dwie cząstki nietrwałe rozpadające się niemal w miejscu swej produkcji na cząstki trwałe. W takich wypadkach mówi się o procesie kwazibinarnym.

procesy binarne i wielocząstkowe

procesy kwazibinarne

Wśród procesów binarnych można wyróżnić ważną klasę procesów elastycznych, tj. takich, w których cząstki końcowe są fizycznie identyczne z cząstkami początkowymi. Zbliżony charakter mają te procesy, w których jedyna zmiana w porównaniu ze stanem początkowym polega na wymianie ładunku (procesy wymiany ładunku, np. π^{-}-$p \to \pi^{0}$-n) lub też na wymianie dziwności (procesy z wymianą dziwności, np. K^{-}-$p \to \pi^{-}$-Σ^{+}). Niekiedy mówi się też o procesach z wymianą innych liczb kwantowych.

Z czteropędów p_a i p_b cząstek początkowych można utworzyć jeden niezmiennik lorentzowski, a więc wielkość niezależną od wyboru układu odniesienia.



Rys. 3. Określenie układu laboratoryjnego (a) i układu środka masy (b) dla procesów wielocząstkowych i binarnych

zmienna s

Wielkością tą jest zmienna $s = (p_a+p_b)^2$. W układzie laboratoryjnym, zdefiniowanym warunkiem $\vec{p}_b = 0$ (rys. 3a), zmienna s wyraża się wzorem

$$s = m_a^2+m_b^2+2m_b E_a^l(\vec{p}_a), \qquad (1)$$

gdzie m_a i m_b są masami zderzających się cząstek, a E_a^l jest energią cząstki-pocisku w układzie laboratoryjnym. Oczywiście, energia i pęd są związane zależnością relatywistyczną, tj. ($c = 1$)

$$E = \sqrt{\vec{p}^2+m^2}\,. \qquad (2)$$

W układzie środka masy zdefiniowanym warunkiem $\vec{p}_a+\vec{p}_b = 0$ (rys. 3b), zmienna s dana jest wzorem

$$s = W^2, \qquad (3)$$

gdzie W jest energią całkowitą obu cząstek w ukła-

dzie środka masy (nazywa się ją także masą efektywną układu cząstek a i b), daną zatem wzorem

$$W = E_a^* + E_b^* \tag{4}$$

(odtąd gwiazdkami będą zawsze oznaczane wielkości zdefiniowane w układzie środka masy).

zmienna t

Dla każdej pary cząstek, jednej początkowej i jednej końcowej, można zdefiniować wielkość zwaną kwadratem przekazu czteropędu, np. $t = (p_a - p_c)^2$. W wypadku procesów binarnych wprowadza się jeszcze zmienną $u = (p_a - p_d)^2$. Można łatwo wykazać, korzystając z zasady zachowania czteropędu,

zmienna u

$$p_a + p_b = p_c + p_d, \tag{5}$$

że zmienne s, t i u nie są niezależne, gdyż spełniają one równość

$$s + t + u = m_a^2 + m_b^2 + m_c^2 + m_d^2. \tag{6}$$

Zależności — takie jak związek (6) — między zmiennymi kinematycznymi charakteryzującymi dany proces powodują, że proces o n cząstkach w stanie końcowym (i, oczywiście, dwu w stanie początkowym) jest opisywany tylko przez $3n-4$ zmienne niezależne. W procesie binarnym istnieją więc tylko dwie takie zmienne i można je wybrać jako s i t. W każdym ustalonym układzie odniesienia można też wprowadzić dwie inne zmienne, bliżej powiązane z pomiarem. Przykładem takiej pary zmiennych może być (w układzie laboratoryjnym) energia lub pęd cząstki-pocisku oraz kąt między kierunkiem jej lotu i kierunkiem lotu jednej z cząstek końcowych, np. cząstki c (kąt produkcji cząstki c). W układzie środka masy dwiema wybranymi zmiennymi są często wartość bezwzględna pędu cząstek zderzających się, $p_i = |\vec{p}_a| = |\vec{p}_b|$ i kąt między kierunkiem lotu cząstki a i cząstki c, θ^*. Jeśli cząstki a i c są identyczne, to mówimy nie o kącie produkcji, lecz o kącie rozpraszania. Zmienna t wyraża się przez zmienne kinematyczne zdefiniowane w układzie środka masy wzorem

$$t = m_a^2 + m_c^2 - 2E_a^* E_c^* + 2p_a^* p_c^* \cos\theta^*. \tag{7}$$

Ponieważ w procesach elastycznych pęd początkowy p_i jest dokładnie równy pędowi końcowemu $p_f = |\vec{p}_c| = |\vec{p}_d|$, wzór powyższy znacznie się wtedy upraszcza i przybiera postać

$$t = -2p^{2*}(1 - \cos\theta^*). \tag{8}$$

W procesach wielocząstkowych można się także posługiwać zmiennymi zdefiniowanymi niezależnie od wyboru układu odniesienia, lecz często wtedy, gdy dąży się do określenia stanu jednej lub dwu wybranych cząstek, wprowadza się pojęcia nowe, które są tu bardziej przydatne. Wektor pędu cząstki rozkłada się na pęd podłużny $\vec{p}_{\parallel}$, tj. tę składową pędu, która jest skierowana równolegle do pędu cząstki pocisku, oraz pęd poprzeczny $\vec{p}_{\perp}$, znajdujący się w płaszczyźnie prostopadłej do lotu pocisku. Przy transformacjach Lorentza, w których prędkość względna obu układów jest skierowana zgodnie z $\vec{p}_a$, pędy poprzeczne wszystkich cząstek nie zmieniają się, gdyż przekształceniu ulegają tylko zmienne $p_{\parallel}$. Aby dalej uprościć problem przechodzenia od układu do układu wprowadza się zmienną zw. pospiesznością (ang. *rapidity*), zdefiniowaną wzorem

pospieszność

$$y = \frac{1}{2}\ln\left|\frac{E+p}{E-p}\right|, \tag{8a}$$

która ma tę istotną własność, że przy zmianie układu odniesienia pospieszności wszystkich cząstek zmieniają się w szczególnie prosty sposób, a mianowicie do każdej z nich dodaje się ta sama stała odpowiadająca pospieszności względnej obu układów odniesienia.

pseudopospieszność

W doświadczeniu mierzy się niekiedy tylko kąty, pod którymi wylatują poszczególne cząstki. Wygodnie jest wtedy wprowadzić wielkość zw. pseudopospiesznością, η, zdefiniowaną wzorem:

$$\eta = -\ln \operatorname{tg}(\theta/2). \tag{9}$$

Użyteczność tej zmiennej wynika stąd, że jest ona powiązana z pospiesznością wzorem:

$$p_\perp \sinh\eta = (\sqrt{m^2 + p_\perp^2})\sinh y. \tag{10}$$

Widać z tego, że w tych wszystkich wypadkach, w których masa wyprodukowanej cząstki jest mała w porównaniu z jej pędem poprzecznym (zachodzi to często dla mezonów π), pospieszność i pseudopospieszność są w przybliżeniu jednakowe.

amplitudy przejścia

Zgodnie z ogólnymi zasadami fizyki kwantowej, przejścia od stanu początkowego do końcowego opisywane są zespołem funkcji zw. amplitudami przejścia; przy czym każda z nich jest zależna od wymienionych wcześniej $3n-4$ zmiennych kinematycznych. Amplitudy przejścia są funkcjami zespolonymi tych zmiennych, a ich znajomość daje kompletny opis procesu. Liczba amplitud zależy od spinów cząstek zarówno początkowych jak i końcowych. Ponieważ hadron o spinie S może występować w $2S+1$ stanach spinowych, przeto łączna liczba amplitud nie może przekraczać wartości $(2S_a+1)(2S_b+1)(2S_c+1)$... W rzeczywistości jednak jest ona znacznie mniejsza ze względu na to, że w oddziaływaniach silnych zachowywana jest parzystość (a więc amplitudy muszą spełniać warunki wynikające z niezmienniczości teorii względem inwersji). Ponadto dalsze ograniczenia liczby amplitud w procesach elastycznych pochodzą z niezmienniczości teorii względem odwrócenia czasu, a jeśli zderzające się cząstki są jednakowe, mogą powstać dodatkowe zależności między amplitudami wynikające z zasad symetrii lub antysymetrii funkcji falowej zależnie od tego, czy cząstki te są bozonami czy też fermionami. W rezultacie liczba amplitud znacznie się zmniejsza. W dwu szczególnie ważnych wypadkach mamy: dwie niezależne amplitudy dla zderzeń cząstek o spinach 0 i $^1/_2$ oraz 6 amplitud dla zderzeń elastycznych cząstek o spinie $^1/_2$: liczba ta zmniejsza się do 5, gdy zderzające się cząstki są jednakowe (np. dla zderzeń nukleon-nukleon). Oczywiście w procesach nieelastycznych, a szczególnie wielociałowych, liczba niezależnych amplitud, mimo wszystkich możliwych ograniczeń, może być bardzo duża. Pojawienie się efektów zależnych od spinu prowadzi do poważnych komplikacji (co prawda czysto technicznych), których pełne omówienie nie jest tu możliwe. Wobec tego wszędzie tam, gdzie nie będzie to specjalnie zaznaczone, wszystkie hadrony będą traktowane tak, jak gdyby były cząstkami bezspinowymi.

Przegląd danych doświadczalnych

Wielkościami mierzonymi doświadczalnie są m.in.: a) całkowity przekrój czynny $\sigma_{\text{całk}}$ na zderzenie dwu danych hadronów; b) całkowity przekrój czynny na zderzenie elastyczne dwu hadronów, $\sigma_{\text{el całk}}$; c) różniczkowy przekrój czynny na rozmaite procesy binarne, w tym także na proces elastyczny, $d\sigma/d\Omega$; d) rozmaite wielkości charakteryzujące stany spinowe zderzających się cząstek (wektory i tensory polaryzacji); e) rozkłady krotności wyprodukowanych cząstek, w tym przede wszystkim krotność średnia $\bar{n}$; f) rozkłady kątowe jednej ustalonej cząstki wyprodukowanej w zderzeniu danych dwu hadronów (są to rozkłady inkluzywne, które można zapisać symbolicznie w następujący sposób:

rozkłady inkluzywne

$$a + b \to c + X,$$

gdzie X oznacza wszystkie inne cząstki, które zostały

wyprodukowane oprócz cząstki c, ale nie były identyfikowane); rozkłady te dane są wielkością

$$E\frac{d\sigma}{dp_{\parallel}d^2p_{\perp}} \equiv \frac{d\sigma}{dy\,d^2p_{\perp}}; \tag{11}$$

g) podobne rozkłady dla dwu lub więcej cząstek odpowiadające procesom

$$a+b \to c+d+X,$$
$$a+b \to c+d+e+X$$

itd.

Analiza takich danych pozwala na wyznaczenie innych wielkości fizycznych, które mogą mieć głębszy sens fizyczny i które są przedmiotem zainteresowania teorii oddziaływań silnych.

W bardzo schematycznym zarysie istniejące dane doświadczalne przedstawiają się następująco.

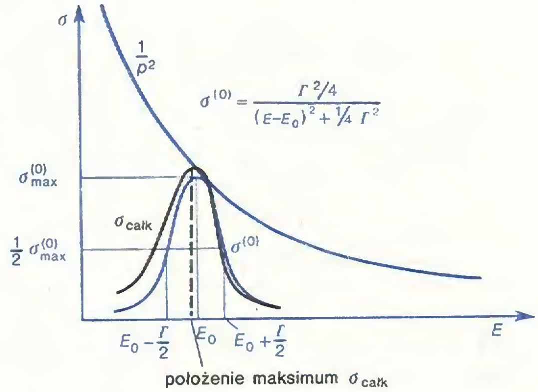
zależność przekroju czynnego od energii

Gdy energia zbliża się do wartości progowej całkowity przekrój czynny reakcji może albo zmierzać do nieskończoności albo do pewnej stałej, niekiedy równej zeru. Pierwszy przypadek odpowiada tym procesom, którym od progu towarzyszy anihilacja (w ogólnym sensie tego słowa); wówczas $\sigma_{całk}$ zmierza do nieskończoności zgodnie z prawem

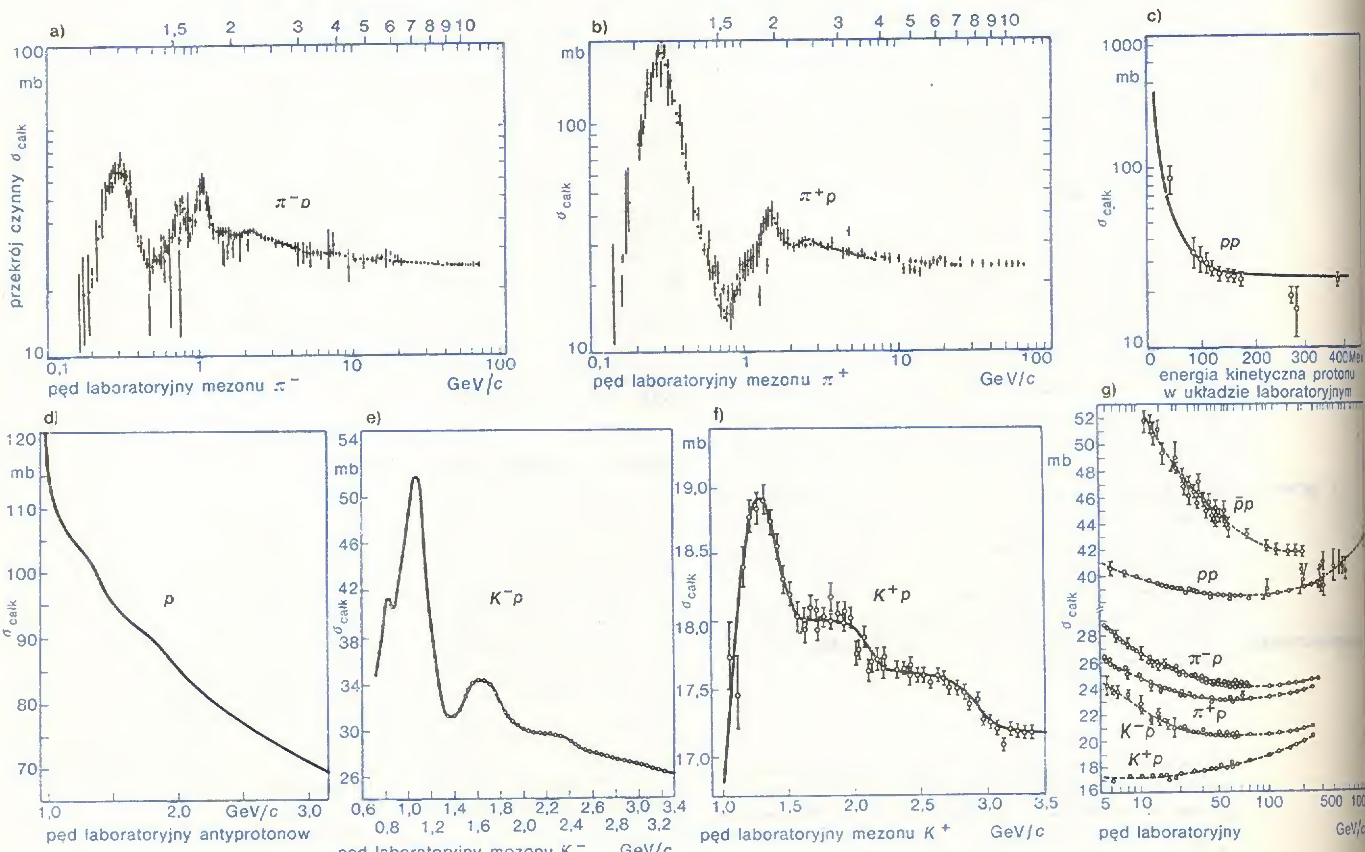
$$\sigma_{całk} \approx \frac{\text{const}}{v}, \tag{12}$$

w którym v — prędkość cząstki padającej w układzie laboratoryjnym. Całkowity przekrój czynny jest bowiem ilorazem dwu czynników, z których licznik przedstawia prawdopodobieństwo zajścia jakiegokolwiek oddziaływania przy danym stanie początkowym, mianownik zaś — gęstość strumienia cząstek padających, proporcjonalną do v. Próg reakcji jest określony przez znikanie prędkości cząstki padającej, tak więc mianownik wyrażenia na $\sigma_{całk}$ zawsze zmierza do zera. Jednakże w procesach, którym już od progu towarzyszą przejścia nieelastyczne, nawet w progu nie znika prawdopodobieństwo przejścia, co prowadzi do nieskończonej wartości ilorazu. Natomiast procesy, które w pobliżu progu są czysto elastyczne, charakteryzują się tym, że w progu znika prawdopodobieństwo przejścia i w rezultacie iloraz (który ma teraz postać 0/0) przybiera pewną wartość skończoną.

Dalszy przebieg zależności $\sigma_{całk}$ od energii zależy od tego, czy w procesie zderzenia pojawiają się rezonanse, czy też nie. W pierwszym wypadku wykres $\sigma_{całk}(E)$ jest krzywą, wykazującą szereg stopniowo coraz to niższych i szerszych maksimów, które po przekroczeniu pewnej wartości energii zanikają. Wartość energii, przy której takie maksimum występuje, odpowiada w przybliżeniu wartości energii



Rys. 4. Całkowity przekrój czynny w pobliżu energii rezonansowej, odpowiadający wzorowi Breita–Wignera



Rys. 5. Wykresy zależności $\sigma_{całk}(E)$ dla różnych zderzeń hadronowych w zakresie energii niskich (a–f) i energii wysokich (g)

rezonansowej. Zachowanie się przekroju czynnego (rys. 4) w otoczeniu takiego punktu można opisać w przybliżeniu fenomenologicznym wzorem Breita–Wignera

wzór Breita–Wignera

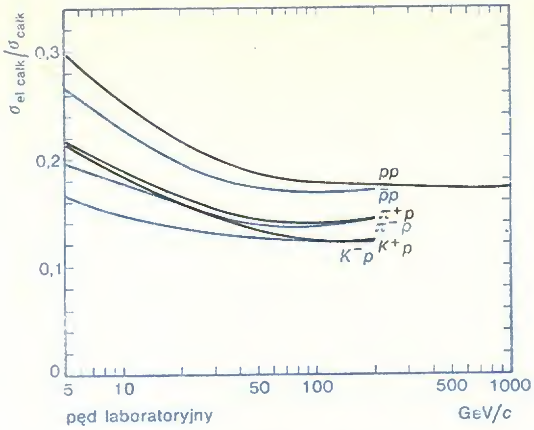
$$\sigma_{\text{całk}} = \frac{(2S+1)}{p^2}\,\frac{\Gamma^2/4}{(E-E_0)^2+{}^1\!/_4\Gamma^2}, \tag{13}$$

gdzie Γ — szerokość połówkowa rezonansu, E_0 — energia rezonansowa, p — pęd w układzie środka masy zderzających się obiektów, a S — spin rezonansu (całkowity moment pędu zderzających się cząstek). Zgodnie z zasadą nieokreśloności, rezonans możemy traktować jako bardzo nietrwałą cząstkę, której średni czas życia wiąże się z Γ wzorem

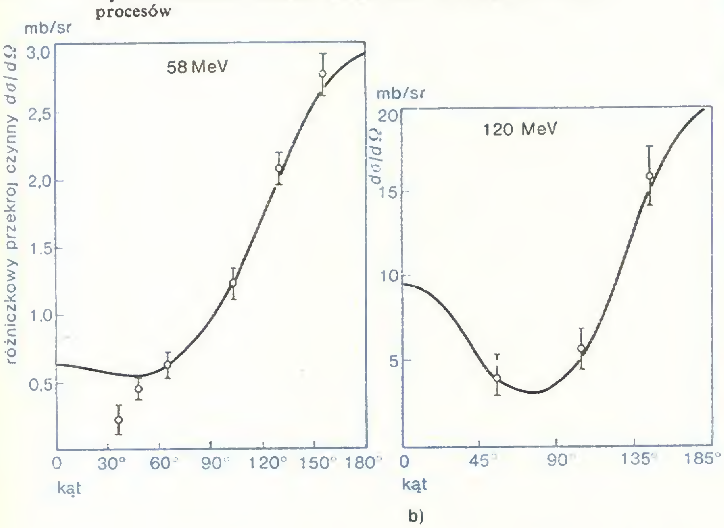
czas życia rezonansu

$$\Gamma \cdot \tau \approx \hbar. \tag{14}$$

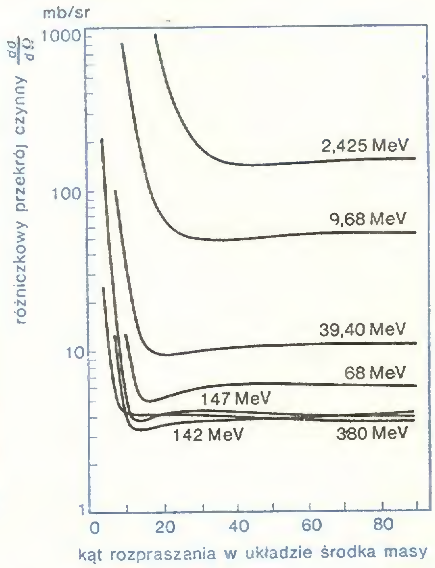
W wypadku drugim, gdy rezonansów nie ma, albo też gdy są one mało widoczne, wykres zależności



Rys. 6. Zależność stosunku $\sigma_{\text{el całk}}/\sigma_{\text{całk}}$ od energii dla różnych procesów

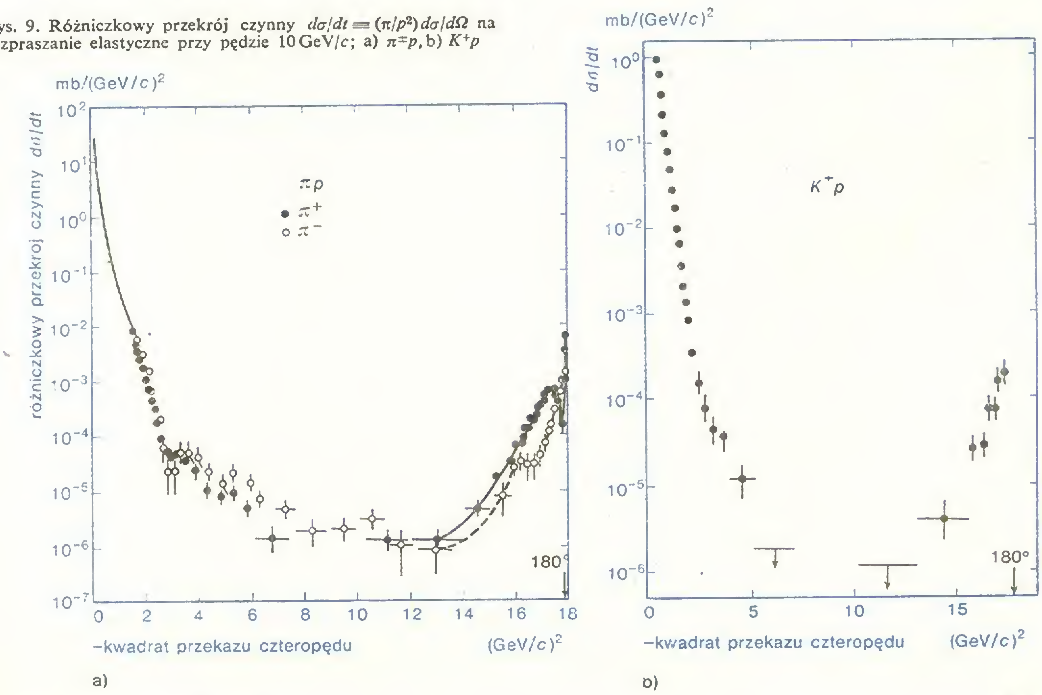


Rys. 7. Różniczkowy przekrój czynny $d\sigma/d\Omega$ na rozpraszanie elastyczne π-p w obszarze niskich energii



Rys. 8. Różniczkowy przekrój czynny $d\sigma/d\Omega$ na rozpraszanie elastyczne p-p w obszarze niskich energii

Rys. 9. Różniczkowy przekrój czynny $d\sigma/dt = (\pi/p^2)\,d\sigma/d\Omega$ na rozpraszanie elastyczne przy pędzie 10 GeV/c; a) $\pi^\pm p$, b) K^+p



$\sigma_{całk}$ od energii jest krzywą gładką, zbliżoną do wykresu funkcji stałej. W obszarze najwyższych energii obserwuje się jednak uniwersalny wzrost $\sigma_{całk}$ proporcjonalny do kwadratu logarytmu energii (rys. 5).

Zależność od energii całkowitego przekroju czynnego na proces elastyczny, a zatem i na wszystkie procesy nieelastyczne, $\sigma_{nieel} = \sigma_{całk} - \sigma_{el\ całk}$, dany jest krzywą zbliżoną do wykresu całkowitego przekroju czynnego, mimo że stosunek tych wielkości nie jest stały i wykazuje pewną zmienność z energią (rys. 6).

procesy elastyczne

Różniczkowy przekrój czynny na zderzenie elastyczne wykazuje silną zależność zarówno od energii, jak i (na ogół) od kąta, jak wreszcie, co najmniej w obszarze niezbyt dużych energii, także od rodzaju zderzających się cząstek. Na rys. 7 i 8 podane są przykładowo różniczkowe przekroje czynne na elastyczne rozpraszanie π^+-p oraz p-p, w obu wypadkach w obszarze niskich energii, natomiast na rys. 9 i 10 — w obszarze wysokich energii. Z rysunków tych widać, że przy przejściu od niskich do wysokich energii obraz całkowicie się zmienia. Różniczkowy przekrój czynny, który w obszarze niskich energii jest niemal izotropowy lub stosunkowo wolno zmienny z kątem rozpraszania, w obszarze energii wysokich jest duży tylko dla bardzo małych kątów, gdzie obserwuje się ostre maksimum. Nachylenie krzywej przedstawiającej to maksimum rośnie na ogół ze wzrostem energii. Przekrój czynny opada wykładniczo, przybierając w obszarze kątów bliskich 90° bardzo małą wartość. Zwykle przed osiągnięciem tej małej wartości krzywa przechodzi przez kilka maksimów i minimów. W obszarze kątów bliskich 180° (w rozpraszaniu proton-proton ze względu na identyczność cząstek jest to obszar fizycznie nieodróżnialny od obszaru kątów bliskich zeru) pojawia się nowe maksimum, znacznie niższe niż dla małych kątów.

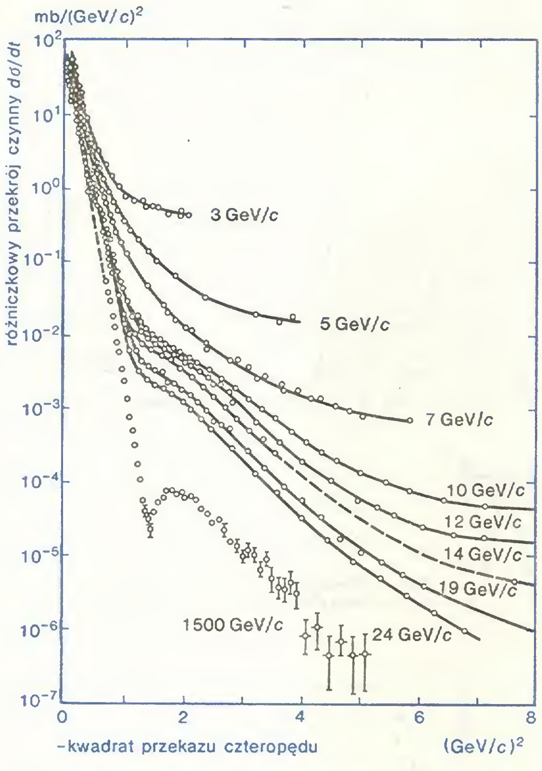
wyznaczanie stosunku Re/Im

Ważną — z teoretycznego punktu widzenia — informacją doświadczalną jest wartość stosunku części rzeczywistej do części urojonej amplitudy rozpraszania. W obszarze niskich energii wielkość tę można zawsze wyznaczyć po dokonaniu analizy fazowej, o czym będzie dalej mowa. Natomiast w obszarze wysokich energii z różnych powodów nie można wykonać analizy fazowej. W tym wypadku wyznacza się stosunek Re/Im badając interferencję amplitudy rozpraszania jądrowego z amplitudą rozpraszania kulombowskiego. W obszarze bardzo dużych energii rozpraszanie kulombowskie jest znaczne tylko dla bardzo małych kątów, a więc tylko tam wyraz interferencyjny może być znaczny i dostępny obserwacji. O rozpraszaniu kulombowskim wiadomo praktycznie wszystko; jego amplituda jest niemal czysto rzeczywista. Tak więc w dobrym przybliżeniu wyraz interferencyjny pochodzi z iloczynu części rzeczywistej obu amplitud. W rezultacie uzyskuje się informacje nie tylko o wartości części rzeczywistej amplitudy uwarunkowanej oddziaływaniem silnym, ale też i o jej znaku. Oczywiście metodę tę można stosować tylko dla cząstek naładowanych elektrycznie, między którymi działa siła kulombowska. Wyniki przedstawione są na rys. 11 i 12. Dla wszystkich zbadanych reakcji Re/Im jest małą wielkością zmniejszającą się wraz z energią. Można powiedzieć zatem, że amplituda rozpraszania elastycznego w obszarze wysokich energii jest w przybliżeniu funkcją urojoną, przynajmniej w obszarze bardzo małych kątów.

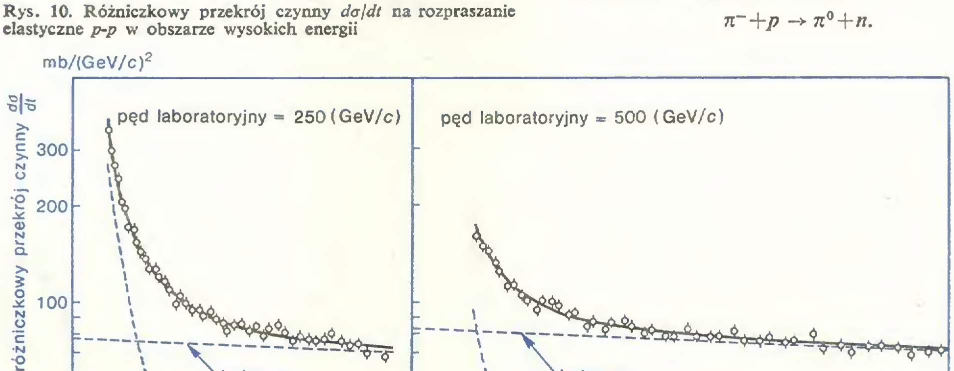
procesy z wymianą ładunku

Oprócz procesów elastycznych bardzo duże znaczenie w teorii oddziaływań silnych mają też inne procesy binarne. Szczególnie wiele uwagi poświęca się analizie procesów z wymianą ładunku. Jednym z najlepiej zbadanych jest proces

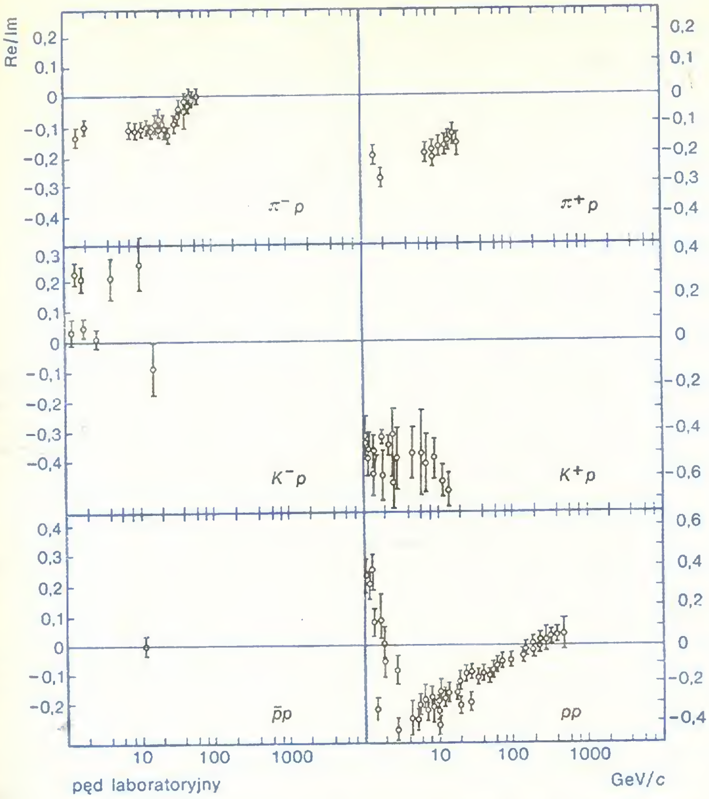
$$\pi^- + p \to \pi^0 + n.$$



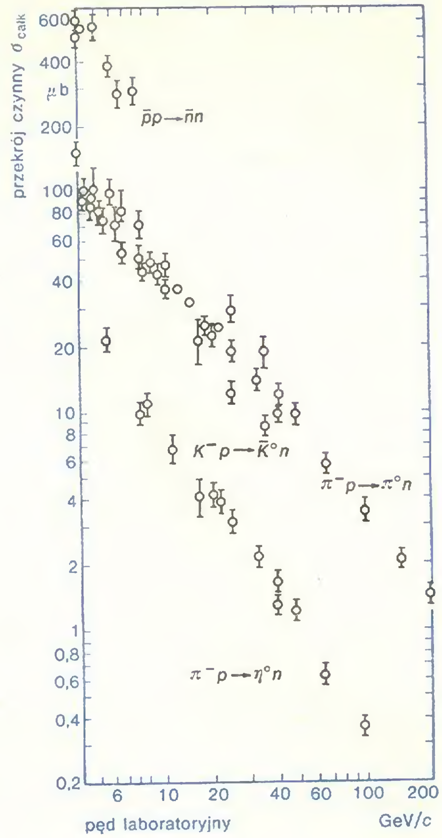
Rys. 10. Różniczkowy przekrój czynny $d\sigma/dt$ na rozpraszanie elastyczne p-p w obszarze wysokich energii



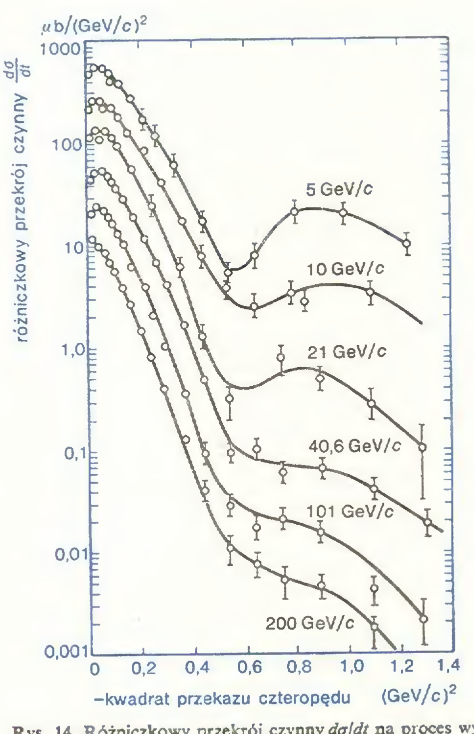
Rys. 11. Różniczkowy przekrój czynny $d\sigma/dt$ na rozpraszanie elastyczne p-p przy wysokiej energii w tym zakresie wartości kwadratu przekazu czteropędu, w którym można zaobserwować interferencję kulombowską



Rys. 12. Zależność od pędu laboratoryjnego stosunku części rzeczywistej do urojonej amplitudy rozpraszania elastycznego rozmaitych hadronów na protonach



Rys. 13. Zależność od energii całkowitego przekroju czynnego na procesy wymiany ładunku



Rys. 14. Różniczkowy przekrój czynny $d\sigma/dt$ na proces wymiany ładunku $\pi^- p \to \pi^0 n$ przy kilku wartościach energii w zakresie $0 < |t| < 1{,}3$ (GeV/c)²

Na rys. 13 przedstawiona jest zależność od energii całkowitego przekroju czynnego na ten proces. Widać z niego, że w obszarze dużych energii wielkość ta szybko maleje z energią. Różniczkowy przekrój czynny na ten proces dla dużych energii pokazany jest na rys. 14. Istnieje cała grupa tego typu procesów binarnych, lecz nieelastycznych i niemal wszystkie wykazują szybki spadek całkowitego przekroju czynnego z energią. Wyjątkowe znaczenie ma więc ta grupa procesów kwazielastycznych, dla których całkowity przekrój czynny maleje z energią bardzo powoli lub nawet w ogóle nie maleje. Zaobserwowano, że dotyczy to przede wszystkim tych procesów, w których nie ma wymiany ładunku, dziwności itd., czyli które spełniają tzw. warunek Morrisona–Gribowa. Właśnie w tych procesach powstają cząstki w stanie końcowym, które różniąc się spinem o ΔS i parzystością o ΔP od cząstki początkowej, mają

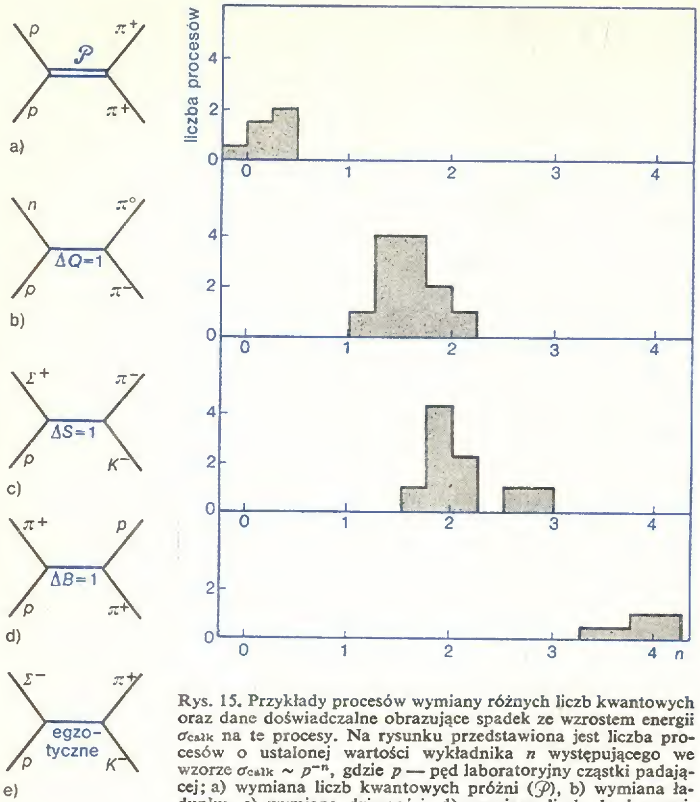
warunek Morrisona–Gribowa

$$\Delta P = (-1)^{\Delta S} \tag{15}$$

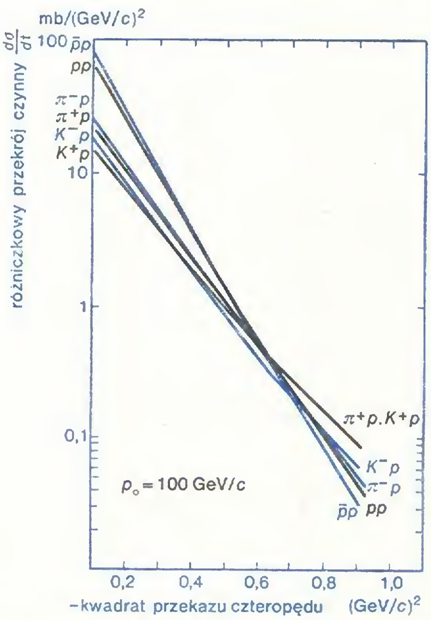
Innymi słowy, jeśli spin cząstki zmienia się o liczbę parzystą, to parzystość cząstki się nie zmienia, natomiast jeśli spin zmienia się o liczbę nieparzystą, to parzystość zmienia się z dodatniej na ujemną lub z ujemnej na dodatnią.

Ta reguła jest przykładem pewnej ogólniejszej zależności, która wiąże szybkość zmniejszania się przekroju czynnego na dany proces z wartościami liczb kwantowych, które są „wymieniane” między oddziałującymi cząstkami. Jest zrozumiałe, że w procesie elastycznym, w którym liczby kwantowe obu cząstek nie zmieniają się, przekazywane są między cząstkami liczby kwantowe próżni (zakładamy, że próżnia jest obojętna elektrycznie, ma dziwność, liczbę barionową, powab, izospin i spin równe zeru, a parzystość zwykłą i ładunkową równą +1). Także w omówionych procesach są wymieniane liczby kwantowe próżni. Okazuje się zaś, że gdy wymieniane liczby kwantowe odpowiadają liczbom kwantowym dowolnego mezonu niedziwnego, spadek przekroju

własności próżni



Rys. 15. Przykłady procesów wymiany różnych liczb kwantowych oraz dane doświadczalne obrazujące spadek ze wzrostem energii $\sigma_{całk}$ na te procesy. Na rysunku przedstawiona jest liczba procesów o ustalonej wartości wykładnika n występującego we wzorze $\sigma_{całk} \sim p^{-n}$, gdzie p — pęd laboratoryjny cząstki padającej; a) wymiana liczb kwantowych próżni ($\mathcal{P}$), b) wymiana ładunku, c) wymiana dziwności, d) wymiana liczby barionowej e) wymiana liczb kwantowych mezonu egzotycznego



Rys. 16. Zjawisko przecinania się wykresów obrazujących różniczkowe przekroje czynne na zderzenia elastyczne cząstki i antycząstki na protonach (p_0 — pęd laboratoryjny cząstki padającej)

czynnego następuje z grubsza biorąc jak s^{-1}, jeśli byłby to mezon dziwny — to jak s^{-2}, jeśli barion — to jak s^{-3}–s^{-4}, a wreszcie, jeśli przenoszonym liczbom kwantowym nie można przyporządkować żadnej znanej cząstki — spadek jest jeszcze szybszy. Przykłady poszczególnych rodzajów procesów wymiany liczb kwantowych oraz dane doświadczalne obrazujące spadek ze wzrostem całkowitego przekroju czynnego na te procesy zostały podane na rys. 15.

Ważne informacje uzyskuje się też, porównując ze sobą wartości przekrojów czynnych na rozpraszanie elastyczne na protonach cząstek (o ładunku ujemnym) i ich antycząstek (o ładunku dodatnim), a więc np. $\bar{p}$-p i p-p, π^--p i π^+-p, K^--p i K^+-p. Jeśli chodzi o wartości całkowitych przekrojów czynnych na rozpraszanie takich dwu cząstek, to stwierdza się, że ich różnica maleje szybko z energią, przy czym $\sigma_{całk}$ dla cząstki jest zawsze większy niż dla antycząstki. Odbija się to także w pewien sposób na postaci różniczkowego przekroju czynnego na rozpraszanie elastyczne, który w obszarze małych kątów jest zawsze większy dla cząstki. Ciekawe jednak, że przy pewnej wartości zmiennej t wykresy obu różniczkowych przekrojów czynnych przecinają się i zaczyna dominować przekrój czynny antycząstki (rys. 16).

wyznaczanie widma mas cząstek nietrwałych

Jednym z najważniejszych zadań fizyki oddziaływań silnych jest wyznaczenie widma mas cząstek nietrwałych (ze względu na oddziaływania silne). Cząstki te mają tak mały średni czas życia (rzędu 10^{-20}–10^{-24} s), że przed rozpadem mogą przebyć tylko bardzo krótki odcinek, praktycznie niedostrzegalny ze względu na ograniczoną zdolność rozdzielczą przyrządów. Innymi słowy, w doświadczeniu zawsze obserwuje się już tylko produkty rozpadu tych nietrwałych obiektów, a o ich istnieniu i własnościach wnioskuje się z analizy tych właśnie produktów rozpadu. W niektórych wypadkach, jak była mowa powyżej, takie nietrwałe obiekty są obserwowane jako rezonanse odpowiadające maksimom na wykresie całkowitego przekroju czynnego na rozpraszanie danych dwu cząstek trwałych. Mówi się wtedy o procesie formacji rezonansów. W pewnych wypadkach, szczególnie w obszarze wyższych energii, rezonansom takim nie odpowiadają żadne widoczne maksima, a o ich istnieniu można się dowiedzieć dopiero po przeprowadzeniu analizy fazowej.

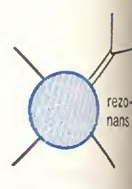


Schemat procesu formacji rezonansu

proces formacji rezonansów

proces produkcji rezonansów

Niekiedy obserwuje się rezonanse powstałe nie w wyniku oddziaływania cząstek początkowych, lecz wskutek oddziaływania w stanie końcowym. Mówi się wówczas o procesie produkcji rezonansów. Rezonanse takie są widoczne na wykresach przedstawiających prawdopodobieństwo pojawiania się pary (rzadziej większej grupy) cząstek, zależnie od ich masy efektywnej, a więc całkowitej energii całego układu tych cząstek w ich układzie środka masy. Przykłady takiej sytuacji są przedstawione na rys. 17. Cząstce nietrwałej odpowiada wyraźnie maksimum na wykresie, które może być najczęściej opisane wzorem Breita–Wignera (13). W pewnych wypadkach (np. rozpraszanie π-π) o istnieniu rezonansów pochodzących z produkcji można się dowiedzieć także po wykonaniu analizy fazowej. Najczęściej jednak takiej analizy nie można wykonać, wobec czego uzyskane dotychczas informacje na temat widma mas cząstek nietrwałych na pewno są niekompletne. Oddzielnym problemem jest wyznaczenie liczb kwantowych cząstek nietrwałych, szczególnie zaś ich spinu i parzystości. I tu znów wielką pomocą jest wynik analizy fazowej, jeśli ją można przeprowadzić. W wielu wypadkach trzeba się jednak uciekać do innych metod, czasem dostosowanych do danego indywidualnego wypadku.



Schemat procesu produkcji rezonansu

pomiary ekskluzywne i inkluzywne

Analizę procesów wielociałowych przeprowadza się w ten sposób, że pomiarem i identyfikacją obejmuje się albo wszystkie wyprodukowane obiekty (pomiar ekskluzywny), albo też tylko niektóre spośród nich (pomiary te nazwaliśmy już inkluzywnymi). Metoda pierwsza, aczkolwiek dostarcza pełnej informacji o stanie końcowym i jest przez to szczególnie cenna, nie może być stosowana wtedy, gdy krotność cząstek jest bardzo duża. Znaczną część wszystkich cząstek stanowią bowiem zawsze cząstki neutralne, które nie są rejestrowane przez większość przyrządów. Gdy rośnie całkowita krotność wyprodukowanych cząstek, wzrasta automatycznie udział tych procesów, w których występuje więcej niż jedna neutralna cząstka, i które nie mogą być dokładnie wyodrębnione z ca-

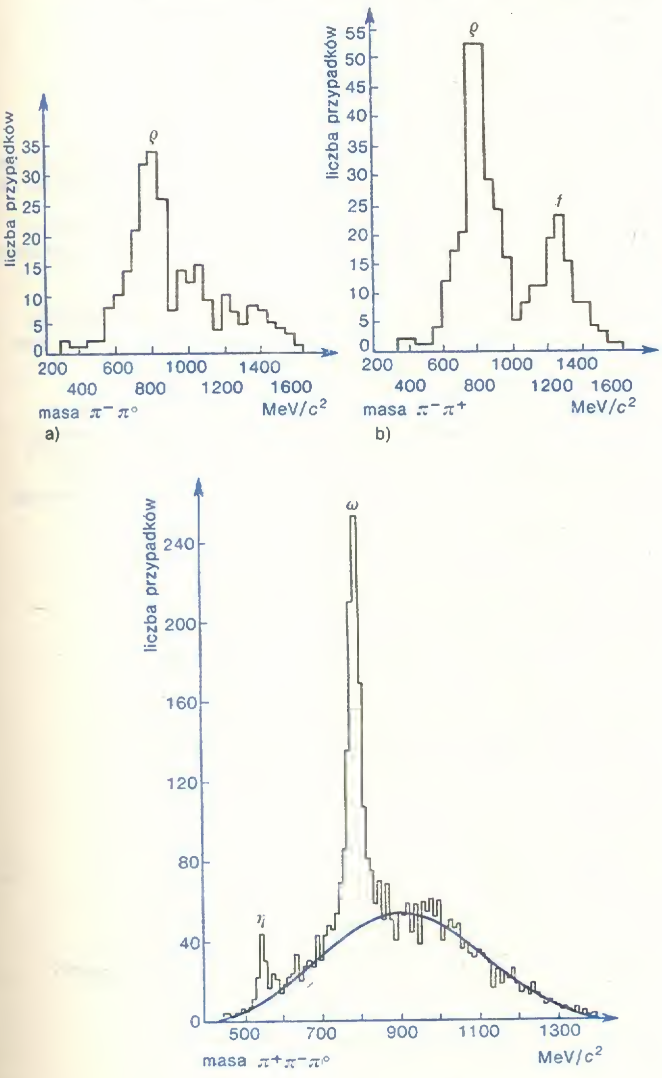
łości danych. Nie znaczy to, że analiza procesów zawierających najwyżej jedną cząstkę neutralną, w których wobec tego można w zasadzie zidentyfikować wszystkie cząstki, traci swoje znaczenie. Jednakże w ostatnich latach uwaga fizyków koncentrowała się przede wszystkim na pomiarach inkluzywnych.

Najprostszym pojęciowo pomiarem inkluzywnym jest pomiar całkowitego przekroju czynnego. Interesującą wielkością jest przecież w tym wypadku prawdopodobieństwo zajścia, przy ustalonym stanie początkowym, procesu, w którym nie identyfikuje się ani jednej cząstki w stanie końcowym. Identyfikując jedną cząstkę końcową, znajduje się pojedynczy rozkład inkluzywny, dwie — rozkład podwójny itd. Rozkłady potrójne, a tym bardziej jeszcze wyższe, są na razie mało zbadane i nie będą tu omawiane.

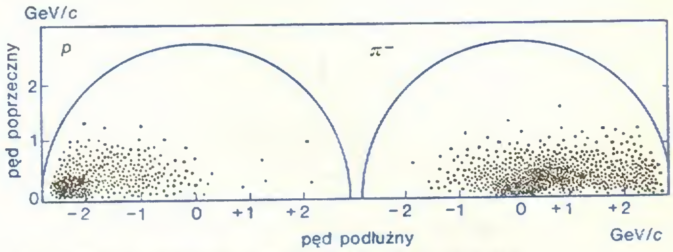
Charakterystyczną cechą rozkładów pojedynczych jest szybki spadek prawdopodobieństwa produkcji cząstek wtórnych ze wzrostem pędu poprzecznego. Zjawisko to jest dobrze jakościowo przedstawione na rys. 18. Jest tam także widoczna inna charakterystyczna cecha produkcji cząstek, a mianowicie istnienie tzw. cząstek wiodących. Przez cząstkę wiodącą rozumie się cząstkę, która jest tego samego typu co cząstka padająca i która kontynuuje z dużym pędem podłużnym kierunek lotu cząstki padającej. I tak np., jeżeli cząstką padającą jest proton, to w dużej liczbie zdarzeń można odnaleźć proton szybki poruszający się zgodnie z kierunkiem lotu protonu początkowego. To samo, choć w mniejszym stopniu, dotyczy też mezonów K, a w jeszcze mniejszym — mezonów π.

cząstki wiodące

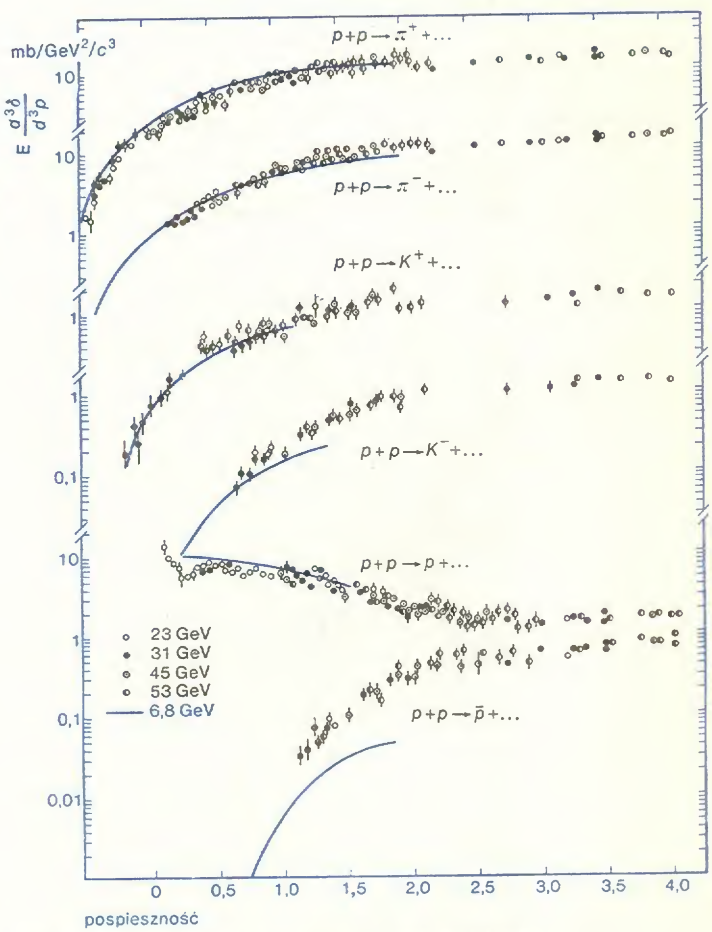
Na rys. 19 i 20 pokazane są dla przykładu rozkłady inkluzywne cząstek produkowanych w zderzeniach proton-proton. Na pierwszym z nich widać (jako funkcję pospieszności) rozkład inkluzywny rozmaitych cząstek powstałych w zderzeniach bardzo wysokiej energii. Na uwagę zasługuje kilka charakterystycznych cech tych rozkładów. Po pierwsze, widać wyraźnie znaczne różnice w wydajności produkcji rozmaitych rodzajów cząstek. Najobficiej produkowane są mezony π, znacznie rzadziej — pary mezonów K, a jeszcze rzadziej — pary barion-antybarion. Ilościowo fakt ten uwidacznia fenomenologiczny



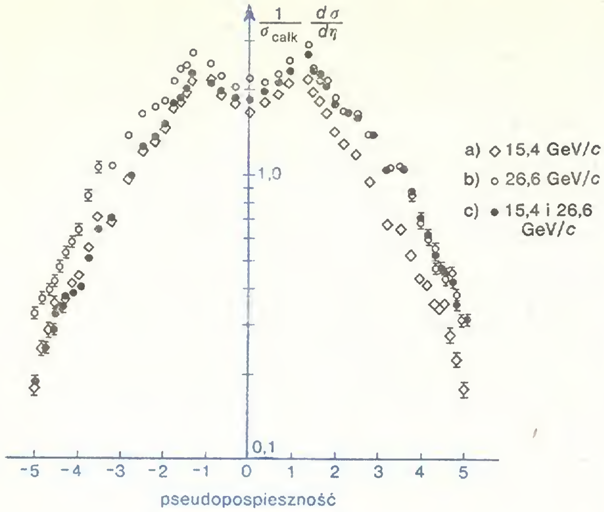
Rys. 17. Maksima w rozkładzie masy efektywnej układów cząstek produkowanych w zderzeniach hadronów, wskazujące na istnienie rezonansów: a) para $\pi^-\pi^0$ (mezon ρ,) b) para $\pi^-\pi^+$ (mezony f i ρ), c) trójka $\pi^+\pi^-\pi^0$ (mezony η i ω)



Rys. 18. Wykres Peyrou dla protonów i mezonów π^- pochodzących z procesu $\pi^-p \to p\pi^+\pi^-\pi^-\pi^0$ przy pędzie laboratoryjnym 16 GeV/c



Rys. 19. Rozkłady inkluzywne rozmaitych cząstek produkowanych w zderzeniach p-p wysokiej energii przy pędzie laboratoryjnym zawartym między 6,8 a 53 GeV/c oraz przy ustalonej wartości pędu poprzecznego $p_{\perp} = 0,4$ GeV/c



Rys. 20. Rozkłady inkluzywne mezonów produkowanych w zderzeniach wiązek przeciwbieżnych przy pędach obu wiązek równych; a) ok. 15,4 GeV/c, b) ok. 26,6 GeV/c, c) przy pędzie jednej wiązki jak w (a), a drugiej jak w (b)

wzór wyrażający podział całkowitej średniej krotności cząstek produkowanych w zderzeniach proton-proton przy pędzie cząstki padającej ok. 1500 GeV/c. Wzór ten ma postać

$$\bar{n}_{\text{całk}} = \bar{n}_{\pm}+\bar{n}_0 = 12\pm 6 = 4{,}7\pi^+ +4{,}5\pi^0+4{,}3\pi^- + +0{,}43K^+ +0{,}31K^- +1{,}6p+0{,}15\bar{p}+\ldots$$

gdzie $\bar{n}_{\pm}$ i $\bar{n}_0$ są to średnie krotności cząstek naładowanych i obojętnych.

zagadnienie krotności

Oczywiście ani średnia krotność ani też (w mniejszym stopniu) podział tej krotności między poszczególne rodzaje cząstek nie są niezależne od energii. Całkowita krotność wzrasta z energią dość szybko, z grubsza zgodnie ze wzorami:

$$\bar{n} = A+B\ln s+C\ln^2 s$$
lub
$$\bar{n} = a+b\ln s+(c\ln s)/s^{1/4}, \tag{16}$$

gdzie A, B, C oraz a, b, c są pewnymi liczbami niezależnymi od energii. Oczywiście, w braku przesłanek teoretycznych, wybór wzoru przedstawiającego wzrost całkowitej krotności produkowanych cząstek naładowanych może być przypadkowy. Jednakże wydaje się, że w każdym razie średnia krotność rośnie szybciej niż $\ln s$. Warto odnotować, że przy każdej wartości energii najwięcej produkowanych cząstek to mezony π, następnie ze wzrostem energii zaznacza się powolny wzrost udziału mezonów K, a w dalszej kolejności także par barion-antybarion.

Informacje na temat całkowitej krotności średniej nie wyczerpują znajomości problemu krotności. Istnieją bowiem dość dokładne dane dotyczące rozkładu krotności cząstek naładowanych przy ustalonej energii w dużym zakresie energii. Są to dane bardzo interesujące, które rzucają wiele światła na mechanizm produkcji cząstek. Gdyby bowiem produkcja cząstek miała charakter czysto probabilistyczny, oczekiwalibyśmy, że rozkład krotności będzie przypominać rozkład Poissona. Jedną z charakterystycznych cech tego rozkładu jest to, że kwadrat dyspersji liczby cząstek, $D^2 \equiv \overline{n^2}-\bar{n}^2$, jest równy średniej liczbie cząstek, $\bar{n}$. Dane doświadczalne przeczą jednak przypuszczeniu, że rozkład krotności cząstek naładowanych dany jest wzorem Poissona. Istnieje fenomenologiczny wzór, podany przez Andrzeja K. Wróblewskiego, który trafnie opisuje dane doświadczalne, i który ma postać

wzór Wróblewskiego

$$D = A(\bar{n}_{\pm}-\alpha), \tag{17}$$

gdzie $A = 0{,}58$ dla wszystkich rodzajów zderzeń, zaś $\alpha \approx 1$ i zależy nieco od zderzających się cząstek. Rozkład krotności cząstek produkowanych jest więc różny od rozkładu Poissona, co wskazuje na istnienie korelacji w procesie produkcji cząstek.

Powracając teraz do analizy danych przedstawionych na rys. 19, należy zauważyć, że pomijając istotny dla rozkładu protonów efekt cząstki wiodącej, rozkłady cząstek różnych rodzajów mają podobny charakter. Ta uniwersalność (przybliżona i nie obejmująca cząstek wiodących) jest widoczna także wtedy, gdy zmieni się cząstka padająca.

Ogólnie rozkłady inkluzywne można scharakteryzować w następujący sposób. Dla wartości pospieszności bliskich wartościom granicznym zaznacza się bardzo interesujący fakt, polegający na tym, że kształt a nawet wartość przekroju czynnego zależy tylko od charakterystyki kinematycznej jednej ze zderzających się cząstek. Aby to lepiej opisać, dogodnie jest wprowadzić trzy obszary kinematyczne: obszar $p_{\parallel}^*$ bliskich pędowi pocisku (obszar fragmentacji pocisku), obszar $p_{\parallel}^*$ bliskich pędowi tarczy (obszar fragmentacji tarczy) i wreszcie obszar $p_{\parallel}^*$ małych (obszar centralny). Okazuje się, że w dwu pierwszych obszarach rozkłady nie zależą od całkowitej energii zderzenia (a więc od s), a tylko odpowiednio od pędu pocisku i pędu tarczy. Jest to wyraźnie widoczne na rys. 20. Natomiast w obszarze centralnym zależność od s, choć słaba, jednak się pojawia. Postulowane w niektórych modelach teoretycznych centralne plateau niezależne od s nie pojawia się (być może dlatego, że dostępne obecnie energie są jeszcze zbyt niskie).

rozkłady ze względu na pęd podłużny i poprzeczny

Rozkłady ze względu na pęd podłużny i poprzeczny są od siebie w przybliżeniu niezależne. Rozkład w pędzie poprzecznym wykazuje wykładniczy spadek z $p_{\perp}$, przy czym wartość średnia pędu poprzecznego jest niewielka, rzędu kilkuset MeV/c, wolno rosnąc z masą wyprodukowanej cząstki.

rozkłady podwójne

funkcje korelacji

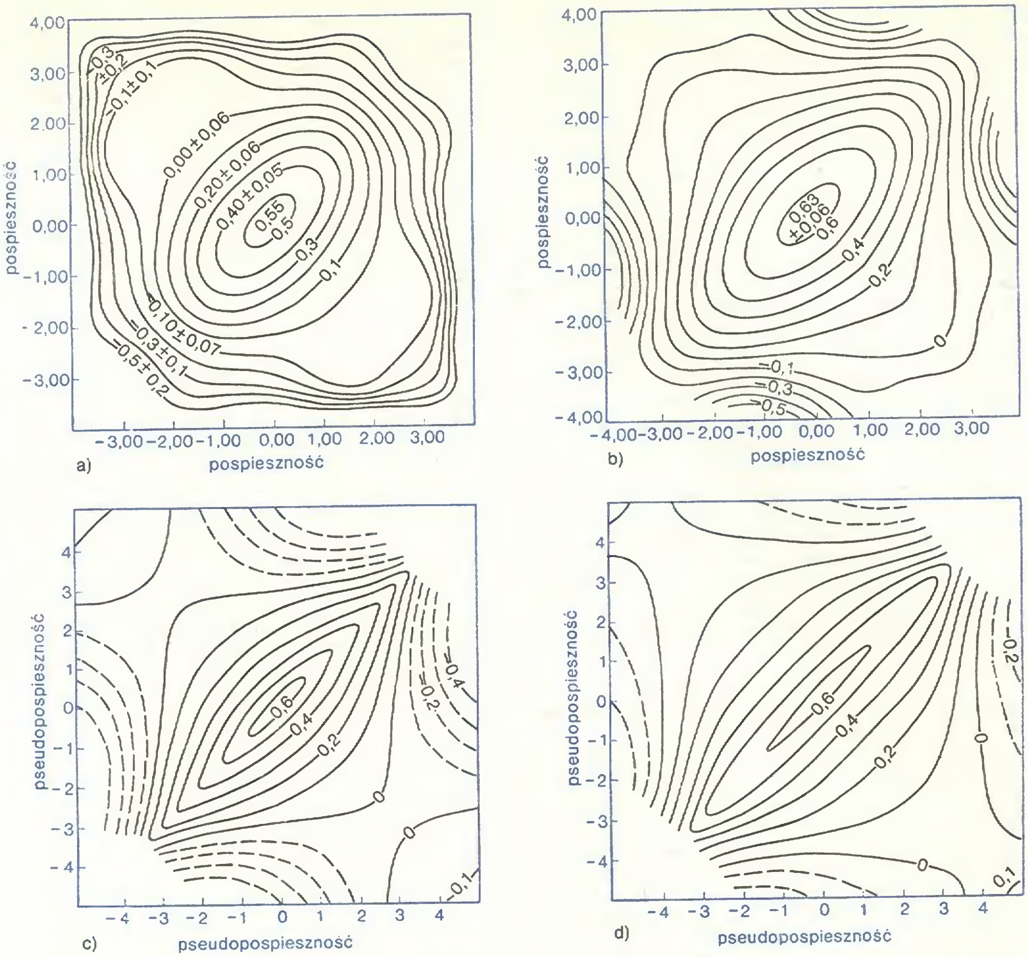
Niezwykle ważnych i cennych informacji dostarczają rozkłady podwójne. Zamiast posługiwać się wprost podwójnym różniczkowym przekrojem czynnym, wprowadza się funkcje korelacji, w celu stwierdzenia, czy produkcja cząstek następuje niezależnie (w tym wypadku podwójny rozkład byłby równy po prostu iloczynowi rozkładów pojedynczych). Interesującą wielkością jest więc różnica między rozkładem podwójnym i iloczynem rozkładów pojedynczych,

$$C(y_1, y_2) = \frac{1}{\sigma}\frac{d^2\sigma}{dy_1 dy_2} - \frac{1}{\sigma^2}\frac{d\sigma}{dy_1}\frac{d\sigma}{dy_2}. \tag{18}$$

Chcąc tę wielkość przedstawić jako funkcję dwu zmiennych, należałoby użyć wykresu trójwymiarowego. Zamiast niego można też posłużyć się metodą poziomic. Poziomice odpowiadające funkcji korelacji dwucząstkowych pokazane są na rys. 21. Podstawową cechą tych korelacji jest silne maksimum funkcji korelacji występujące w pobliżu przekątnej $y_1 = y_2$. Innymi słowy, istnieją wyraźne, dodatnie korelacje między cząstkami o małej różnicy pospieszności (zwane korelacjami krótkozasięgowymi). Na rys. 21 zaznacza się jednak także istnienie korelacji długozasięgowych, w obszarach $y_1 = -y_2$ przy dużych wartościach bezwzględnych obu tych zmiennych. Obraz produkcji cząstek jest więc dość skomplikowany. Najprostsza narzucająca się hipoteza jest taka, że cząstki końcowe są pierwotnie produkowane w pewnych grupach, w jakichś kawałkach materii hadronowej, zwane czasem zgęstkami lub klastrami (ang. *clusters*), które być może choćby częściowo są znanymi już rezonansami. Te zgęstki rozpadałyby się izotropowo w swoim układzie środka masy. Wszystkie mezony π pochodzące z tego samego zgęstka miałyby zbliżone wartości pospieszności i oczywiście byłyby skorelowane dodatnio. Tak więc, znaczna część, może nawet przeważająca, cząstek końcowych byłaby produkowana w procesie dwuetapowym. (Chodzi tu

zgęstki

funkcje korelacji



Rys. 21. Funkcje korelacji dla par mezonów produkowanych w zderzeniach p-p przy różnych pędach: a) 102 GeV/c, b) 400 GeV/c, c) 240 GeV/c, d) 2000 GeV/c

korelacje długozasięgowe

przede wszystkim o te cząstki, które mają małe wartości pospieszności w układzie środka masy). Istnienie korelacji długozasięgowych wskazuje jednak na istnienie jeszcze jakiegoś innego mechanizmu dynamicznego produkcji cząstek, być może w postaci wzbudzenia i tarczy, i pocisku (bez wymiany liczb kwantowych) oraz ich — w następnych chwilach — rozpadu na cząstki końcowe (mechanizm dyfrakcyjny). Zagadnienia te nie są jednak na razie ostatecznie wyjaśnione.

Opis teoretyczny oddziaływań silnych

Ogromne bogactwo danych, które tu zostały przedstawione bardzo fragmentarycznie, wymaga oczywiście jakiejś interpretacji teoretycznej. Powstaje przede wszystkim problem, na jakich podstawowych zasadach teoria ta mogłaby być oparta. W tej kwestii zasadniczo panuje zgodność między fizykami zajmującymi się teorią cząstek. Teoria ta musi być po pierwsze teorią kwantową (bo odnosi się do mikroświata), a po drugie zgodną z teorią względności (relatywistyczną), gdyż dotyczy obiektów, których energia całkowita jest na ogół znacznie większa od ich energii spoczynkowej.

Całkowitego uzgodnienia tych dwu zasadniczych idei teoretycznych nie udało się jeszcze dokonać. Trudności wiążą się między innymi z tym, że w teorii takiej z konieczności musi się rozpatrywać nie pojedyncze cząstki, lecz pola fizyczne, a przy tym nie można pomijać procesów produkcji. Innymi słowy, musi to być teoria zawierająca nieskończenie wiele stopni swobody. Mimo istniejących ogromnych trudności wyłoniło się wiele zasad, o charakterze ogólnym, które służą za przewodnika w gąszczu danych.

Teoria będzie mieć na pewno charakter relatywistyczny, jeśli wszystkie wielkości teoretyczne będą funkcjami tylko niezmienników lorentzowskich. Oczywiście, jak stwierdzono poprzednio, w ustalonym układzie odniesienia od niezmienników tych można zawsze przejść do wielkości zdefiniowanych tylko w tym układzie, takich jak np. kąt rozpraszania.

Kwantowy charakter teorii zmusza do posługiwania się pewnymi funkcjami zespolonymi, które zależą od wymienionych powyżej niezmienników. Funkcje te muszą mieć interpretację probabilistyczną, zgodnie z ogólnymi zasadami fizyki kwantowej.

Macierz S

Przyjmijmy, że stany przed oddziaływaniem (w chwili $t = -\infty$) tworzą zespół stanów opisywanych określonymi wartościami liczb kwantowych i ponumerowanych liczbami i_1, i_2, ... a stany cząstek końcowych (w chwili $t = +\infty$) tworzą zbiór odpowiednich stanów ponumerowanych $f_1, f_2, \ldots$. Amplitudę prawdopodobieństwa tego, że układ cząstek, znajdujących się początkowo w jakimś stanie i przejdzie w wyniku

amplituda prawdopodobieństwa

oddziaływania do jednego ze stanów f, oznacza się symbolem S_{if}. Wielkości S_{if} zależą od pełnej charakterystyki stanu zarówno początkowego jak i końcowego, przy czym od czteropędów, ze względu na postulowaną relatywistyczną niezmienniczość teorii S_{if} zależeć mogą tylko za pośrednictwem opisanych tu niezmienników.

macierz S

Wielkości S_{if} tworzą pewną macierz, nieskończonego rzędu, zw. macierzą S. Jest to kluczowe pojęcie teorii oddziaływań silnych, ponieważ w macierzy tej zawarta jest pełna informacja o wszystkich procesach między cząstkami elementarnymi. Powstają pytania, czy macierz tę można obliczyć w ramach jakiejś teorii, czy jest możliwe choćby powiązanie ze sobą w pewien sposób jej elementów. Rozwiązanie tych problemów stanowi cel teorii oddziaływań silnych między hadronami. W chwili obecnej powstały koncepcje ujmujące częściowo własności macierzy S i pozwalające na pewne przewidywania.

unitarność macierzy S

Probabilistyczna interpretacja macierzy S wraz z postulatem, że wśród stanów i oraz f są wszystkie możliwe stany fizyczne dowolnych układów cząstek prowadzi do wniosku, że macierz ta musi spełniać warunek matematyczny zwany unitarnością, aby było zachowane prawdopodobieństwo w przejściu od stanu początkowego do końcowego. Jeśli symbolem $S^\dagger$ oznaczyć macierz sprzężoną hermitowską z macierzą S, to warunek unitarności można zapisać w postaci

$$SS^\dagger = S^\dagger S = 1, \tag{19}$$

gdzie macierz 1 jest macierzą jednostkową tego samego rzędu co macierz S.

twierdzenie optyczne

Z unitarności macierzy S wypływa prosty wniosek zw. twierdzeniem optycznym. W twierdzeniu tym występuje amplituda rozpraszania elastycznego dwu cząstek a i b (ściślej jej część urojona), przy czym w rozpraszaniu nie może ulec zmianie ani stan spinowy ani nawet pędowy zderzających się cząstek (jest to więc rozpraszanie pod kątem 0°). Twierdzenie optyczne mówi, że wielkość ta jest proporcjonalna do całkowitego przekroju czynnego na jakiekolwiek zderzenie ze stanu wyjściowego $a+b$:

$$\operatorname{Im} f_{a+b\to a+b}(p, \theta = 0°) = (p/4\pi)\sigma_{\text{całk}}(p), \tag{20}$$

gdzie p — pęd zderzających się cząstek.

Właściwy wybór reprezentacji odgrywa istotną rolę w lepszym rozumieniu danych doświadczalnych.

reprezentacja pędowa

W wypadku procesów dwuciałowych reprezentacją możliwą jest reprezentacja pędowa, w której zmiennymi niezależnymi są pędy (w układzie środka masy) cząstki początkowej i końcowej. Ponieważ długość wektora pędu jest wyznaczona przez wartość energii całkowitej W, przeto dogodnymi zmiennymi mogą być W i θ, albo też np. pęd początkowy p_i i θ. Od tej reprezentacji warto niekiedy przejść do reprezentacji momentu pędu, w której zmiennymi niezależnymi są p_i oraz liczba kwantowa orbitalnego momentu pędu l. Element macierzy S odpowiadający procesowi elastycznemu w tej reprezentacji, $S_l(p)$, spełnia nierówność wynikającą z warunku unitarności

$$S_l^* S_l \leqslant 1, \tag{21}$$

przy czym znak równości odpowiada takiej sytuacji kinematycznej, w której z danym procesem nie konkuruje żaden proces nieelastyczny.

Łatwo wykazać, że amplitudę rozpraszania elastycznego można przedstawić w postaci

$$f(p,\theta) = 1/p \sum_{l=0} (2l+1) f_l(p) P_l(\cos\theta), \tag{22}$$

gdzie P_l — wielomian Legendre'a l-tego rzędu, zaś amplituda cząstkowa f_l dana jest wzorem

$$f_l = (S_l - 1)/2i, \tag{23}$$

spełnia zatem warunek

$$|f_l|^2 \leqslant \operatorname{Im} f_l. \tag{24}$$

Warto przy tym zaznaczyć, że jeśli zderzające się cząstki mają spin różny od zera, wzór (22) należy zastąpić wyrażeniem ogólniejszym.

przesunięcie fazowe

Jeżeli proces ma charakter czysto elastyczny, wówczas S_l można sparametryzować wprowadzając przesunięcie fazowe δ_l,

$$S_l = e^{2i\delta_l}. \tag{25}$$

Wzór (25) można stosować nawet wtedy, gdy występują konkurencyjne procesy nieelastyczne, należy tylko przyjąć, że przesunięcie fazowe jest wielkością zespoloną albo też wprowadzić parametr nieelastyczności $\eta_l \leqslant 1$, zdefiniowany wzorem

$$S_l = \eta_l \exp(2i\delta_l). \tag{26}$$

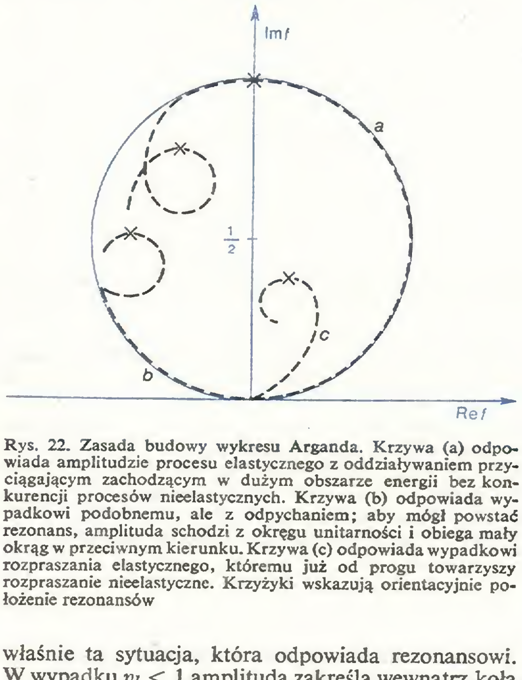
Interpretacja fizyczna przesunięcia fazowego polega na tym, że określa ono różnicę fazy fali padającej i fali końcowej (w obu wypadkach obliczaną w obszarze bardzo dalekim od obszaru oddziaływania).

koło unitarności

Ograniczenie (24) powoduje, że wykres amplitudy cząstkowej rozpraszania na płaszczyźnie zespolonej zawsze zawarty jest w obrębie koła o środku w punkcie $\operatorname{Re} f = 0$ i $\operatorname{Im} f = {}^1/_2$ i o promieniu równym ${}^1/_2$. Koło to nazywamy kołem unitarności. Jest to tzw. wykres Arganda (rys. 22).

wykres Arganda

Widać stąd, że gdy proces jest czysto elastyczny, wartości amplitudy leżą dokładnie na okręgu, a w przeciwnym razie schodzą do wnętrza koła. Można się przekonać, że jeśli $\eta_l = 1$, to amplituda ma wartość maksymalną przy $\delta_l = \pi/2$; jest to



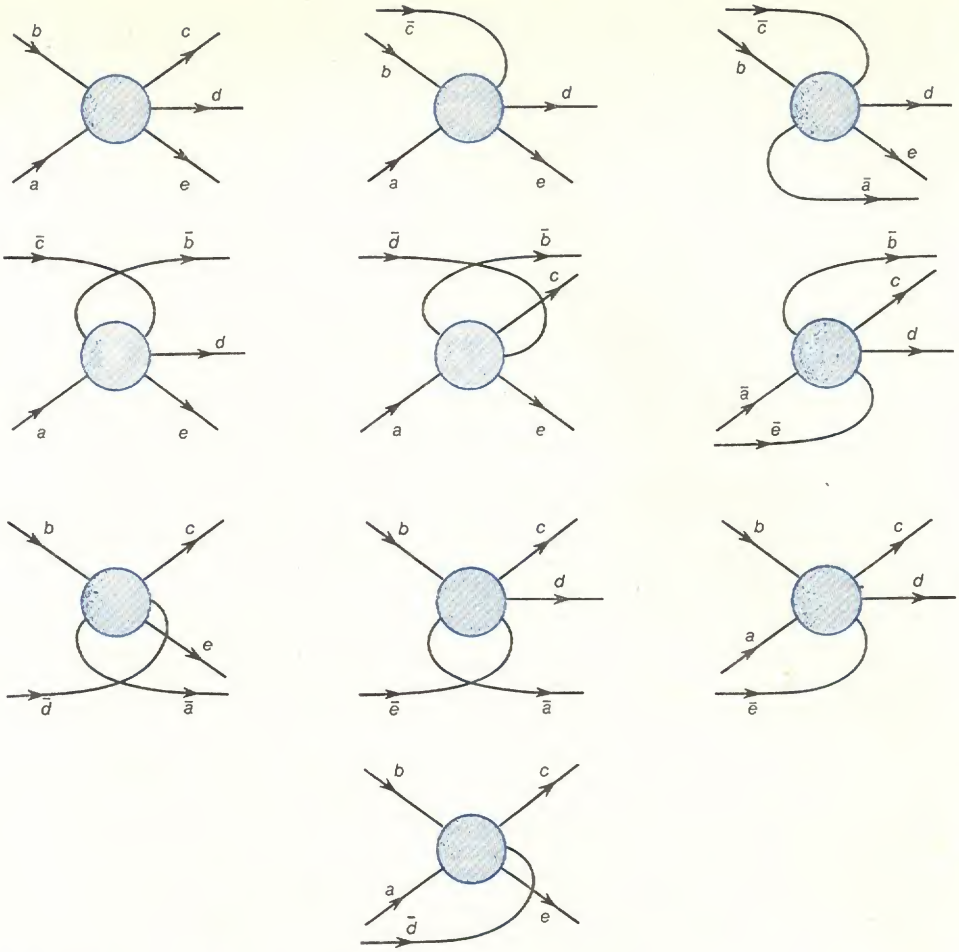
Rys. 22. Zasada budowy wykresu Arganda. Krzywa (a) odpowiada amplitudzie procesu elastycznego z oddziaływaniem przyciągającym zachodzącym w dużym obszarze energii bez konkurencji procesów nieelastycznych. Krzywa (b) odpowiada wypadkowi podobnemu, ale z odpychaniem; aby mógł powstać rezonans, amplituda schodzi z okręgu unitarności i obiega mały okrąg w przeciwnym kierunku. Krzywa (c) odpowiada wypadkowi rozpraszania elastycznego, któremu już od progu towarzyszy rozpraszanie nieelastyczne. Krzyżyki wskazują orientacyjnie położenie rezonansów

właśnie ta sytuacja, która odpowiada rezonansowi. W wypadku $\eta_l < 1$ amplituda zakreśla wewnątrz koła unitarności niekiedy dość skomplikowaną krzywą, lecz zawsze lokalnie w pobliżu rezonansu zakreśla okrąg obiegając go w kierunku przeciwnym do ruchu wskazówek zegara (odpowiada to rozpadowi cząstki nietrwałej; przeciwny kierunek ruchu oznaczałby, że cząstki trwałe skupiają się z powrotem w stanie rezonansowym). Tak więc patrząc na wykres Arganda dla danej fali cząstkowej możemy wykryć rezonanse oraz ustalić ich spin (równy 1).

Jednakże powstaje pytanie, w jaki sposób uzyskać wartości amplitud cząstkowych, które następnie można by odkładać na wykresie Arganda. Zagadnienie to jest przedmiotem analizy fazowej. Punktem wyjścia są wartości całkowitego i różniczkowego przekroju czynnego jako funkcji kąta.

Dla uproszczenia przyjmijmy, że proces jest czysto elastyczny. Korzystając z faktu ortogonalności wie-

procesy skrzyżowane



Rys. 23. Procesy skrzyżowane dla 5 cząstek (łącznie początkowych i końcowych)

analiza fazowa

lomianów Legendre'a można różniczkowy przekrój czynny przedstawić wówczas w postaci

$$\frac{d\sigma}{d\Omega} = |f|^2 = \sum_l P_l(\cos\theta)F_l(p)\frac{1}{p^2}. \qquad (27)$$

Ponieważ oddziaływania silne mają skończony zasięg, przeto parametr zderzenia nie może być zbyt duży, bo w przeciwnym razie nie doszłoby do oddziaływania. Z kolei rząd orbitalnego momentu pędu wynosi $l \approx pb$, gdzie b jest parametrem zderzenia. Tak więc można oczekiwać, że przy każdej skończonej wartości energii wkład do przekroju czynnego wnosić będzie tylko pewna ograniczona (rosnąca z energią) liczba wartości l. W tym przypadku szereg we wzorze (27) można urwać przy pewnej wartości $l_{\max} = L$. Występujące w tym wzorze funkcje F_l zależą od przesunięć fazowych. Gdyby np. przyjąć $L = 0$, to $F_0 = \sin^2\delta_0$. Gdyby L było równe 1, to oprócz F_0 pojawiłaby się jeszcze funkcja $F_1 = \sin\delta_0 \sin\delta_1 (\cos\delta_0 \times \cos\delta_1 + \sin\delta_0 \sin\delta_1)$. Dalsze wyrażenia byłyby oczywiście jeszcze bardziej skomplikowane. Już jednak z tego prostego przykładu wynika, że badając zależność różniczkowego przekroju czynnego od kąta, można wyznaczyć przesunięcia fazowe. Oczywiście jest to problem trudny i pełen niejednoznaczności m.in. z powodu zawsze ograniczonej dokładności pomiaru. Gdy zderzające się cząstki mają spin różny od zera (co zwykle zachodzi), i gdy trzeba jeszcze uwzględnić konkurencyjne procesy nieelastyczne, a liczba fal cząstkowych jest znaczna, zagadnień tych nie można rozwiązywać bez pomocy komputerów. Zasada jednak jest ta sama.

Wartości przesunięć fazowych są ważną informacją m.in. ze względu na możliwość wykrycia stanów rezonansowych. Ponadto znak przesunięć fazowych określa, czy rozpraszanie w danym stanie orbitalnym ma charakter odpychający ($\delta_l < 0$) czy też przyciągający ($\delta_l > 0$). W obszarze najniższych energii przesunięcia fazowe często parametryzuje się wprowadzając zasięg efektywny r_l i długość rozpraszania A_l. Można wykazać, że

$$p^{2l+1}\operatorname{ctg}\delta_l = -\frac{1}{A_l} + \frac{1}{2}r_l p^2 + \ldots \qquad (28)$$

metody nierelatywistyczne

W obszarze energii niezbyt odległym od energii progowej procesu próbuje się często stosować metody nierelatywistyczne oparte na równaniu Schrödingera z odpowiednio dobranym potencjałem. Metody te w wielu wypadkach wystarczają, np. do opisu najprostszych jąder atomowych i rozpraszania przyprogowego dwu nukleonów. W innych sytuacjach można jednak stosować te metody nie po to, aby uzyskać dobrą liczbową zgodność z doświadczeniem, ale po to, by odgadnąć pewne własności macierzy S, które by można było (być może) zastosować nawet tam, gdzie model nierelatywistyczny na pewno przestaje być słuszny.

Przykładem takich własności, częściowo potwierdzonych przez badania teorii relatywistycznych, jest analityczność macierzy S. Analityczność może cechować tylko funkcje zespolone zmiennych zespolonych. Rozważa się więc amplitudę rozpraszania jako funkcję zespolonych argumentów, czy to — niezmienników s, t, u czy też pędu, energii, kąta, a na-

wet liczby l. W tych wypadkach, w których udało się uzyskać ścisłe dowody analityczności macierzy S, własność ta była konsekwencją warunku przyczynowości. W innych wypadkach analityczność macierzy S po prostu się postuluje.

Omówienie choćby drobnej części wiążących się z tym zagadnień przekracza ramy tego artykułu. Wnioski, do których można dojść, są następujące. Przypuśćmy, że istnieje taki stan jednej cząstki trwałej, który ma te same liczby kwantowe zachowywane w oddziaływaniach silnych, co i stan dwu cząstek zderzających się. (Oczywiście przejście dwu tych stanów w siebie nie jest możliwe ze względu na prawo zachowania energii i pędu). W takim wypadku, jeśli M jest masą tego stanu jednocząstkowego, to amplituda rozpraszania cząstek zderzających się mieć będzie biegun pierwszego rzędu w punkcie $s = M^2$ (M — masa stanu jednocząstkowego). W każdym zaś punkcie, w którym istnieje próg procesu dwu- lub więcej-cząstkowego o właściwych liczbach kwantowych, amplituda ma punkt rozgałęzienia.

Postulat maksymalnej analityczności głosi, że jedynymi osobliwościami amplitudy są te właśnie wyżej wymienione bieguny i punkty rozgałęzienia. Aby ten problem przedstawić nieco dokładniej, należy jeszcze wprowadzić ważne pojęcie procesu skrzyżowanego z danym procesem. Jeśli dany jest proces np.

procesy skrzyżowane

$$a+b \to c+d+e,$$

to procesem doń skrzyżowanym nazywa się każdy z poniższych procesów (rys. 23):

$$a+b+\bar{c} \to d+e, \quad b+\bar{c} \to \bar{a}+d+e,$$

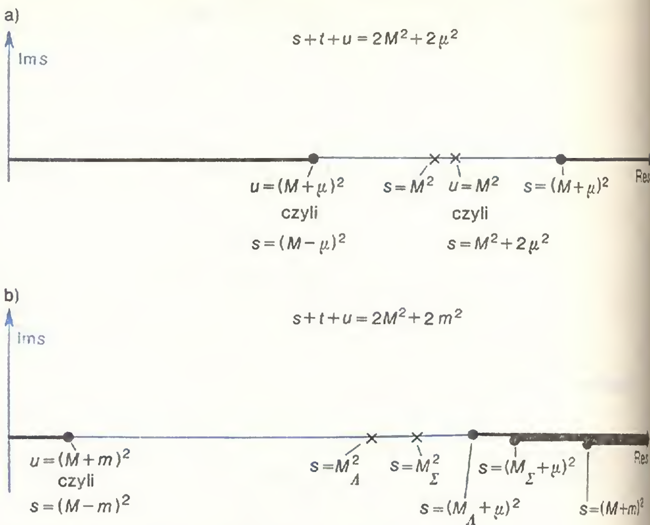
$$a+\bar{d} \to \bar{b}+c+e,$$

i tak dalej, a także każdy z procesów przebiegających w przeciwnym kierunku niż powyższe, albo taki, w którym każda cząstka została zastąpiona przez odpowiednią antycząstkę. Jak widać, proces skrzyżowany różni się od danego procesu tym, że jedna lub więcej cząstek zostaje przeniesiona „na drugą stronę" z jednoczesną zamianą jej na jej antycząstkę.

reguła podstawień

Reguła podstawień mająca podstawowe znaczenie w teorii analityczności macierzy S, głosi, że wszystkie procesy skrzyżowane ze sobą opisywane są funkcjami, które można otrzymać z jednej przez przejście graniczne z wielowymiarowego obszaru zmiennych zespolonych $s, t, u,$... do obszaru, w którym te zmienne przyjmują odpowiadające danemu procesowi wartości rzeczywiste. Konsekwencje tej reguły dla analityczności są bardzo istotne, gdyż automatycznie wszystkie osobliwości jednej z amplitud procesów skrzyżowanych stają się osobliwościami wszystkich innych amplitud.

Warto tu rozpatrzyć pewien konkretny ważny przykład, mianowicie rozpraszanie $\pi^+ + p \to \pi^+ + p$. Istnieją dwa procesy (istotnie różne) skrzyżowane z tym procesem, a mianowicie $\pi^- + p \to \pi^- + p$ oraz $\bar{p} + p \to \pi^+ + \pi$. Jeśli zmienne są zdefiniowane następująco: $s = (p_a + p_b)^2$, $t = (p_b - p_d)^2$ i wreszcie $u = (p_a \quad p_d)^2$, to przejście od rozważanego procesu do pierwszego procesu skrzyżowanego odpowiada zmiana znaku wszystkich składowych czteropędów p_a i p_c. Wówczas zmienna s przybiera postać charakterystyczną dla przekazu czteropędu $(p_b - p_a)^2$, a zmienna u — postać typową dla kwadratu energii całkowitej. W procesie pierwszym istnieje jeden stan jednocząstkowy trwały, a mianowicie nukleon, występuje więc biegun w punkcie $s = M^2$. Podobnie dla procesu drugiego istnieje biegun w punkcie $u = M^2$ odpowiadający nukleonowi. Natomiast w procesie trzecim stan jednej cząstki trwałej się nie pojawia (nie ma takiego trwałego hadronu, który by miał liczby kwantowe pary mezonów π). Tak więc amplitudy rozpraszania wszystkich trzech procesów skrzyżowanych będą mieć dwa bieguny w podanych wyżej punktach, z tym jednak, że interpretacja fizyczna zmiennych s, t, u jest dla wszystkich tych procesów rozmaita. Podobnie można by określić punkty rozgałęzienia amplitud (rys. 24).



Rys. 24. Struktura osobliwości dla: a) procesu $\pi^+p \to \pi^+p$, b) procesu $K^-p \to K^-p$, w obu wypadkach dla $\theta = 0$ (czyli $t = 0$). M oznacza masę nukleonu, μ masę mezonu π, M_Λ masę hiperonu Λ, M_Σ masę hiperonu Σ, m masę mezonu K. Położenia biegunów oznaczone są $\times$, punkty rozgałęzienia •—

związek dyspersyjny

Istnienie osobliwości macierzy S i ich znajomość pozwala na napisanie dla amplitudy rozpraszania $f(s, t)$ reprezentacji całkowej Cauchy'ego, która po pewnych drobnych przekształceniach prowadzi do tzw. związku dyspersyjnego. Typową postać takiego zwiazku przedstawia wzór

$$\mathrm{Re} f(s,t) = \frac{g_{ab}g_{cd}}{s-M^2} + \ldots + \frac{1}{\pi}\int_{s_{\mathrm{prog}}}^{\infty} ds' \frac{\mathrm{Im} f(s',t)}{s'-s} +$$

$$+ \frac{1}{\pi}\int_{u_{\mathrm{prog}}}^{\infty} \frac{\mathrm{Im} f(u',t)}{u'-u} du', \qquad (28a)$$

gdzie całkowanie po każdej zmiennej przebiega od najniższego progu do nieskończoności, a residua w punktach biegunowych są równe iloczynom ładunków silnych charakteryzujących oddziaływanie z odpowiednim stanem jednocząstkowym cząstek początkowych (g_{ab}) oraz cząstek końcowych (g_{cd}). Jeśli stan końcowy i początkowy są identyczne, to residuum staje się równe stałej sprzężenia charakteryzującej oddziaływanie trzech cząstek, tj. cząstek a, b i cząstki pośredniej. Ponieważ część urojona amplitudy rozpraszania ku przodowi może być wyrażona przez całkowity przekrój czynny na podstawie twierdzenia optycznego, a część rzeczywistą amplitudy także się mierzy, przeto z niepewnością wynikającą z nieznajomości danych w obszarze bardzo wysokich energii można wzór dyspersyjny porównywać z doświadczeniem, przy czym w idealnym wypadku jedyną nieznaną w nim wielkością jest właśnie stała sprzężenia. Jest to niemalże jedyna metoda wyznaczania stałych sprzężenia oddziaływań silnych.

Analityczność i inne własności macierzy S pozwalają na ścisłe wyprowadzenie pewnych twierdzeń, zwanych twierdzeniami asymptotycznymi, które odnoszą się do zachowania niektórych wielkości fizycznych w obszarze bardzo wysokich energii (tj. gdy $s \to \infty$). Do najbardziej znanych twierdzeń asymptotycznych należy ograniczenie Froissarta, nie pozwalające całkowitemu przekrojowi czynnemu rosnąć zbyt szybko w nieskończoności,

ograniczenie Froissarta

$$\sigma_{\mathrm{całk}} \leqslant \mathrm{const} \cdot \ln^2 s, \qquad (29)$$

twierdzenie Pomeranczuka

oraz twierdzenie Pomeranczuka, zgodnie z którym całkowite przekroje czynne na zderzenie cząstki a z tarczą b i antycząstki $\bar{a}$ z tą samą tarczą powinny być sobie asymptotycznie równe,

$$\frac{\sigma_{\text{całk}}(ab)}{\sigma_{\text{całk}}(\bar{a}b)} \to 1. \tag{30}$$

Twierdzenia asymptotyczne odgrywają ważną rolę przy testowaniu modeli teoretycznych.

Teoria biegunów Reggego

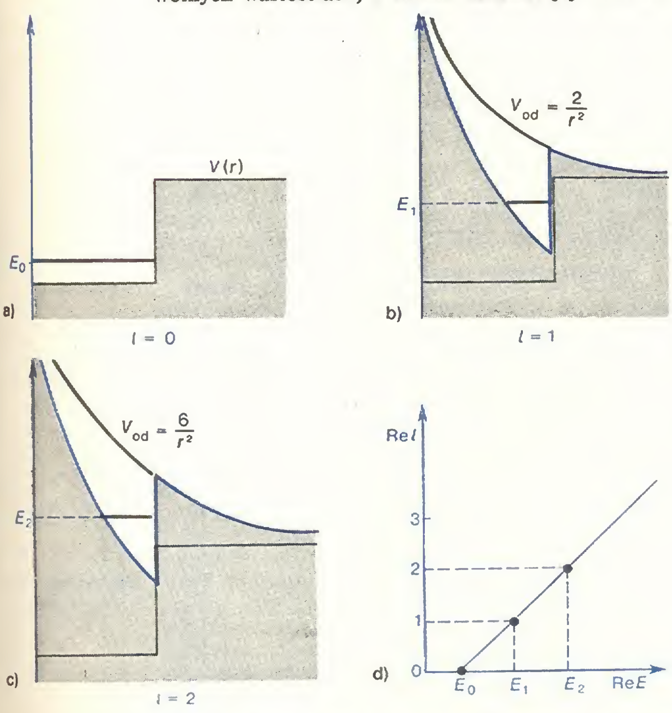
Tullio Regge, badając analityczność nierelatywistycznych amplitud rozpraszania jako funkcji zmiennej zespolonej l, przekonał się, że amplitudy te są meromorficznymi funkcjami l w obszarze $\mathrm{Re}\, l \geqslant -\frac{1}{2}$, dla pewnej klasy potencjałów obejmujących m.in. potencjał Yukawy $V(r) = V_0 e^{-\mu r}/r$. Co więcej, wykazał on, że ewentualne bieguny amplitudy mogą leżeć jedynie w obszarze ograniczonym dodatkowo warunkiem $\mathrm{Im}\, l \geqslant 0$, przy czym położenie takiego bieguna na płaszczyźnie zmiennej zespolonej l zależy na ogół od wartości energii, przy której zachodzi zderzenie. Bieguny te nazywane są biegunami Reggego.

bieguny Reggego

Biegun Reggego ma przejrzystą interpretację fizyczną. Z równania Schrödingera z pewnym ustalonym potencjałem przyciągającym $V(r)$, wynika na ogół istnienie stanów związanych i rezonansowych przy wartościach energii zależnych od liczby kwantowej orbitalnego momentu pędu l. Zależność ta pojawia się dlatego, że do potencjału $V(r)$ dochodzi jeszcze potencjał odpychający związany z siłą odśrodkową, równy

$$+\frac{l(l+1)}{r^2},$$

i efektywny potencjał staje się funkcją liczby l. Oczywiście fizyczne wartości l są całkowite, ale można traktować potencjał ten jako funkcję zmiennej l o dowolnych wartościach, a nawet nadawać jej wartości



Rys. 25. Zasada konstrukcji trajektorii Reggego; a) położenie poziomu związanego, gdy $l = 0$, b) gdy $l = 1$, c) gdy $l = 2$ (w tym wypadku stan przestaje być związany, a staje się stanem rezonansowym); potencjał odśrodkowy $V_{od} = l(l+1)/r^2$; d) wykres Chew–Frautschiego na płaszczyźnie $\mathrm{Re}\, l$, $\mathrm{Re}\, E$

zespolone. Można się przekonać i w końcu do tego sprowadza się odkrycie Reggego, że przy ciągłych zmianach liczby l poziomy energetyczne w potencjale efektywnym, odpowiadające stanom związanym i rezonansowym, przesuwają się także w sposób ciągły. Jeśli biegun przy pewnej wartości s występuje dla wartości (na ogół zespolonej) liczby $l = \alpha(s)$, to funkcję $\alpha(s)$ nazywa się trajektorią bieguna Reggego (rys. 25). Biegun ten jest biegunem amplitudy rozpraszania w tym kanale, w którym s jest kwadratem całkowitej energii zderzających się cząstek w ich układzie środka masy. Residuum tego bieguna (danej amplitudy) jest opisane także pewną funkcją $\beta(s)$, zw. funkcją residualną bieguna Reggego.

trajektoria bieguna Reggego

Regge wykazał, że całkowita amplituda rozpraszania wyraża się w następujący sposób przez wkłady pochodzące od poszczególnych biegunów

$$f(p, \cos\theta) = \sum_j \frac{\beta_j(s) P_{\alpha_j(s)}(-\cos\theta)}{\sin \pi \alpha_j(s)}, \tag{31}$$

gdzie sumowanie obejmuje wszystkie bieguny Reggego. $P_{\alpha_j(s)}(-\cos\theta)$ jest funkcją Legendre'a, która przechodzi w dobrze znany wielomian Legendre'a, gdy $\alpha(s)$ jest liczbą całkowitą nieujemną.

Z wyrażenia na całkowitą amplitudę rozpraszania danego wzorem (31) można przejść do wyrażeń na amplitudy cząstkowe. Okazuje się wtedy, że w otoczeniu punktu o l całkowitym wkład pochodzący od bieguna Reggego reprezentowany jest wyrażeniem Breita–Wignera, a więc biegun ów w takim punkcie przedstawia rezonans, jeżeli tylko $\mathrm{Im}\,\alpha(s) > 0$. Gdy zaś $\mathrm{Im}\,\alpha(s) = 0$, mamy do czynienia ze stanem związanym (szerokość połówkowa Γ jest proporcjonalna do $\mathrm{Im}\,\alpha(s)$, a dla stanów związanych $\Gamma = 0$).

Ponieważ trajektoria bieguna jest funkcją zmiennej s, przeto jeden biegun Reggego zawiera w sobie informacje o wielu rezonansach odpowiadających różnym fizycznym wartościom liczby kwantowej l.

wykres Chew–Frautschiego

Rozpatrując trajektorię Reggego w płaszczyźnie zmiennych ($\mathrm{Re}\, l$, $\mathrm{Re}\, s$) otrzymuje się tzw. wykres Chew–Frautschiego. Okazuje się, że trajektorie obserwowanych rezonansów mają na takich wykresach w przybliżeniu to samo nachylenie,

$$\alpha'(s) \equiv \frac{d\,\mathrm{Re}\,\alpha(s)}{d\,\mathrm{Re}\, s} \approx 1\ \mathrm{GeV}^{-2}. \tag{32}$$

Wyniki doświadczalne zestawione z odpowiednimi trajektoriami Reggego są przedstawione na rys. 26.

Pierwszą zatem korzyścią z wprowadzenia pojęcia bieguna Reggego jest to, że na jednej trajektorii znajduje się wiele stanów związanych i rezonansów, co przyczynia się do uporządkowania widma mas hadronów, a w połączeniu z obserwowaną uniwersalnością nachylenia trajektorii Reggego na wykresie Chew–Frautschiego — pozwala na snucie przewidywań dotyczących mas hadronów dotychczas nie obserwowanych. Należy jednak zaznaczyć, że nie istnieje jak dotąd żadna metoda, pozwalająca na obliczenia czy to trajektorii Reggego czy to funkcji residualnych. Fakt zaś uniwersalności nachylenia i (przybliżonej) prostoliniowości trajektorii jest na razie zupełnie niezrozumiały.

Była już mowa o tym, że zgodnie z prawem podstawień każdy biegun amplitudy jednego z procesów skrzyżowanych jest zarazem biegunem amplitudy pozostałych procesów skrzyżowanych. Dla zilustrowania zagadnienia przytoczymy proste obliczenie kinematyczne. W wypadku, gdy masy wszystkich czterech cząstek biorących udział w procesie są jednakowe, w jednym z kanałów skrzyżowanych („kanale s")

$$s = 4(p_s^2 + m^2), \qquad t = -2p_s^2(1-\cos\theta_s),$$

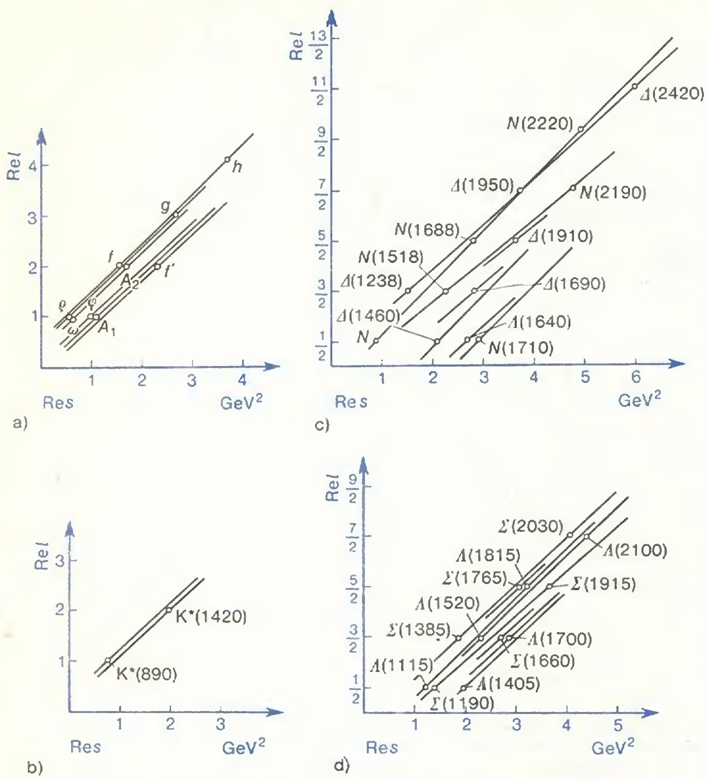
a w drugim („kanale t")

$$s = -2p_t^2(1-\cos\theta_t), \qquad t = 4(p_t^2 + m^2),$$

gdzie wskaźniki „s" i „t" przy pędzie i kącie roz-

wykres Chew-Frautschiego

praszania oznaczają, że dana wielkość jest zdefiniowana w układzie środka masy odpowiedniego procesu, a m jest wartością masy cząstek zderzających się.



Rys. 26. Rozmieszczenie niektórych cząstek trwałych i nietrwałych na trajektoriach Reggego w płaszczyźnie (Rel, Res), tzw. wykres Chew-Frautschiego, a) trajektorie $B = 0$, $S = 0$, b) trajektorie $B = 0$, $S \neq 0$, c) trajektorie $B \neq 0$, $S = 0$, d) trajektorie $B \neq 0$, $S \neq 0$. Spośród trajektorii mezonowych tylko na trajektoriach mezonów ρ i f można umieścić po dwie znane cząstki. Nachylenia pozostałych trajektorii mezonowych dobrano podobnie do wyżej wymienionych

Z powyższych wzorów widać, że małe ujemne wartości niezmiennika t i duże dodatnie wartości niezmiennika s odpowiadają fizycznym wartościom zmiennych p_s i θ_s, ale na pewno niefizycznym wartościom zmiennych p_t i θ_t. Mianowicie, w wymienionym obszarze wartości s i t, $p_t^2 < 0$, zaś $\cos\theta_t \to -\infty$. Jeśli dana jest amplituda reprezentująca wkład od pojedynczego bieguna Reggego w kanale t, zgodnie ze wzorem (31) równa

$$\frac{\beta_j(t)P_{\alpha_j(t)}(-\cos\theta_t)}{\sin\pi\alpha_j(t)}$$

i w obszarze $p_t^2 \geq 0$ i $|\cos\theta_t| \leq 1$ zawierająca wkłady rezonansowe, to ta sama amplituda w innym (niefizycznym) obszarze wartości zmiennym p_t^2 i θ_t zyskuje interpretację amplitudy rozpraszania w kanale skrzyżowanym, tj. w kanale s. W tym właśnie kanale da ona do pełnej amplitudy rozpraszania wkład następujący

$$\bar{\beta}_j(t)(-\cos\theta_t)^{\alpha_j(t)} = \bar{\beta}_j(t)\left[-1-\frac{s}{2p_t^2}\right]^{\alpha_j(t)},$$

gdyż dla bardzo dużych wartości x funkcję $P_l(x)$ można zastąpić przez najwyższą potęgę zmiennej x, czyli przez x^l. Poprawne uwzględnienie wszystkich czynników kinematycznych prowadzi do następującej postaci amplitudy, wyrażającej wkład pojedynczego bieguna Reggego w kanale skrzyżowanym

$$\bar{\bar{\beta}}_j(t)\, s^{\alpha_j(t)-1}$$

(pominięte tu są ważne czynniki odpowiadające tzw. sygnaturze trajektorii). W rezultacie można uzyskać następujące wyrażenia na całkowity przekrój czynny oraz różniczkowy przekrój czynny w kanale s pochodzący od tego samego bieguna Reggego, który w kanale t porządkował stany rezonansowe:

$$\sigma_{\text{całk}} = \text{const}\cdot s^{\alpha(0)-1} \tag{33}$$

oraz

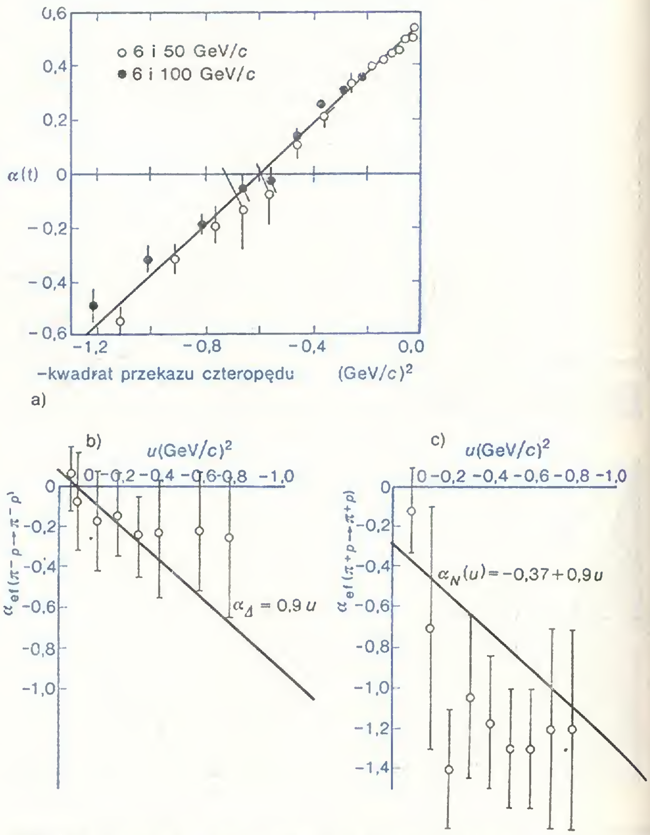
$$\frac{d\sigma}{d|t|} = |\bar{\bar{\beta}}(t)|^2 s^{2(\alpha(t)-1)}. \tag{34}$$

Tak więc trajektorie Reggego nie tylko porządkują widmo mas hadronów, ale także wyznaczają zachowanie się wysokoenergetyczne amplitud w kanale skrzyżowanym.

Hadrony usytuowane na tej samej trajektorii mają te same wartości liczby barionowej B, dziwności S, izospinu I itd., różnią się jedynie masą i spinem. Te wszystkie liczby kwantowe można więc przypisać samej trajektorii. Trajektoria, rozważana z punktu widzenia kanałów skrzyżowanych, opisuje zatem wymianę odpowiednich liczb kwantowych pomiędzy zderzającymi się cząstkami.

Gdyby dla któregokolwiek bieguna Reggego $\alpha(0)$ było większe od 1, wzór (33) musiałby doprowadzić do pogwałcenia ograniczenia Froissarta. Tak więc najwyższą możliwą wartością $\alpha(0)$ jest 1. Odpowiada to zachowaniu się $\sigma_{\text{całk}}$ w obszarze wysokich energii jak const. Takie w przybliżeniu zachowanie się jest istotnie obserwowane, z dokładnością do poprawek logarytmicznych. Biegun o $\alpha(0) = 1$ nazywa się biegunem Pomeranczuka. Trajektoria Pomeranczuka byłaby scharakteryzowana przez liczby kwantowe próżni. Analizując inne trajektorie Reggego z punktu widzenia wymienianych liczb kwantowych, można

trajektoria Pomeranczuka



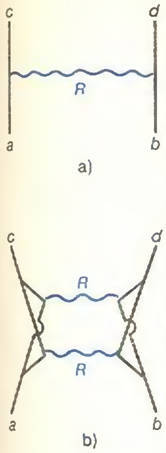
Rys. 27. Porównanie modelu Reggego z wynikami doświadczalnymi: a) trajektoria efektywna dla reakcji wymiany ładunku w zderzeniu π-N, b) i c) opis reggeowski rozpraszania π-N ku tyłowi przy użyciu efektywnych trajektorii nukleonowych (b) — $\pi^- p$, (c) — $\pi^+ p$

uzasadnić zaobserwowane regularności w przebiegu zależności rozmaitych przekrojów czynnych od energii. Warto też zauważyć, że dla prostoliniowej trajektorii

$$\alpha(t) = \alpha_0 + \alpha' t,$$

gdzie α' dane jest wzorem (32), można prawą stronę wzoru (34) przepisać w postaci następującej

$$|\bar{\bar{\beta}}(t)|^2 e^{[2(\alpha_0-1)\ln s]} e^{[2\alpha' t \ln s]}.$$



Rys. 28. Diagram odpowiadający wymianie pojedynczego reggeonu (rys. a); najprostszy diagram dający wkład do cięcia Reggego (wymiana dwu reggeonów, rys. b)

Tak więc wzór ten przewiduje wykładniczy spadek wartości różniczkowego przekroju czynnego ze wzrostem $|t|$ oraz zwiększanie się nachylenia krzywej przedstawiającej ten przekrój czynny ze wzrostem energii (zmiennej s). Oba te wnioski są jakościowo, a często i ilościowo zgodne z danymi doświadczalnymi (rys. 27). Badanie zjawisk wiążących się z nieznikającymi wartościami spinu cząstek, różnicami mas oraz tzw. siłami wymiany wnosi do tego obrazu istotne poprawki, które jednak mają dobre uzasadnienie teoretyczne i znajdują potwierdzenie doświadczalne.

Jak była mowa powyżej, w teorii nierelatywistycznej jedynymi osobliwościami amplitud rozpraszania dla $\mathrm{Re}\, l > -\frac{1}{2}$ są bieguny. Nie jest to jednak prawdą w teorii relatywistycznej. Różnica pochodzi stąd, że w teorii relatywistycznej należy uwzględnić także możliwość produkcji nowych cząstek, a więc stanów wielocząstkowych. Odpowiada to w teorii Reggego istnieniu oprócz biegunów także i cięć w płaszczyźnie zmiennej zespolonej l (tzw. cięć Reggego). Okazało się, że cięcia te w pewnych wypadkach odgrywają istotną rolę w wyjaśnianiu pewnych danych eksperymentalnych, jednakże ich występowanie niezwykle komplikuje teorię Reggego i zmniejsza jej siłę przewidywania (rys. 28).

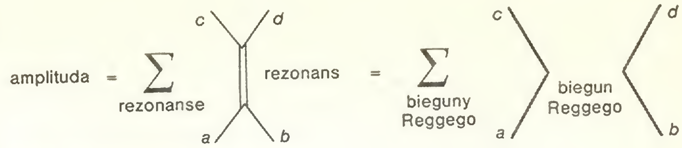
Jako próba konsekwentnego przezwyciężenia tych trudności, powstała w ciągu ostatnich lat tzw. reggeonowa teoria pola, w której operuje się wymianą trajektorii Reggego (tzw. reggeonów), a nie zwykłych cząstek. W teorii tej istnieje możliwość wprowadzenia, bez pogwałcenia ograniczenia Froissarta, także trajektorii o liczbach kwantowych próżni, a więc trajektorii Pomeranczuka o $\alpha(0) > 1$.

reggeonowa teoria pola

Bieguny Reggego znalazły też zastosowanie w teorii procesów inkluzywnych. Dzieje się to w ten sposób, że od procesu $a+b \to c+x$ przechodzi się na podstawie prawa podstawień do procesu $a+b+\bar{c} \to x$, reprezentowanego przez część urojoną amplitudy procesu elastycznego $a+b+\bar{c} \to a+b+\bar{c}$ (zgodnie z uogólnionym twierdzeniem optycznym); tę zaś amplitudę można wyrazić z kolei przez wkłady reggeowskie. To postępowanie doprowadziło do lepszego zrozumienia obszernej klasy zjawisk z udziałem wielu cząstek.

Zaskakującym spostrzeżeniem było to, iż amplitudy opisywane wzorem reggeowskim w obszarze wysokich energii i przedłużone do obszaru energii niższych, gdzie wzór ten nie powinien się stosować, po rozłożeniu na fale cząstkowe wytworzyły na wykresie Arganda okręgi odpowiadające w wielu wypadkach cząstkom obserwowanym w tym samym procesie. Było to dalsze odkrycie, ponieważ do tego czasu bieguny Reggego korelowały zachowanie się wielkości fizycznych przy wysokich energiach w jednym procesie z zachowaniem się rezonansowym — w procesie skrzyżowanym. Sformułowano wówczas hipotezę, że amplituda oddziaływania silnego musi spełniać jeszcze jeden warunek dotychczas nieznany, zwany dualnością. Amplituda spełnia warunek dualności, jeśli w tym samym procesie może być przedstawiona albo jako suma wszystkich wkładów rezonansowych, albo jako suma wszystkich wkładów od biegunów Reggego (rys. 29). Oczywiście, praktycznemu wykorzystaniu tej zasady stoi na przeszkodzie brak informacji o wszystkich możliwych rezonansach i biegunach Reggego, toteż koncepcję tę można sprawdzić tylko z pewnym przybliżeniem. Jednak wnioski uzyskane na jej podstawie niejednokrotnie okazywały się trafne.

warunek dualności



Rys. 29. Zasada warunku dualności

Jednym z kluczowych problemów teorii opartej na pojęciu biegunów Reggego jest status bieguna Pomeranczuka. Jest on niewątpliwie wyróżniony fizycznie przez swoje położenie w płaszczyźnie zmiennej zespolonej l i przez liczby kwantowe próżni, które go charakteryzują. Jednakże również analiza fenomenologiczna danych wskazuje na pewne szczególne cechy tej osobliwości. Po pierwsze, jest to jedyny biegun, którego trajektoria wyłamuje się z ogólnej regularności wartości stałego nachylenia (zob. wzór 32). Nachylenie trajektorii bieguna Pomeranczuka jest znacznie mniejsze; szacuje się je na ok. 0,3 GeV^{-2}. Po drugie, trudno wskazać rezonanse, które by znajdowały się na tej trajektorii. Wprawdzie istnieją takie warianty teorii, w których z trajektorią Pomeranczuka wiąże się mezon f o masie ok. 1270 MeV/c^2, ale takie przyporządkowanie nie jest naturalne i wymaga dość skomplikowanej argumentacji. Po trzecie, wzrost $\sigma_{\text{całk}}$ w obszarze najwyższych obecnie dostępnych energii wskazuje na to, iż amplitudy rozpraszania elastycznego nie mogą być opisywane wyłącznie przez biegun Pomeranczuka, bo w tym wypadku przekrój czynny $\sigma_{\text{całk}}$ przybierałby wartości stałe. Być może zatem to, co nosi obecnie nazwę bieguna Pomeranczuka, jest jakąś inną osobliwością (zespołem cięć). Te trudności w zrozumieniu istoty bieguna Pomeranczuka uwidaczniają się także w próbach włączenia go w schemat teorii dualnej. Niekiedy wypowiada się pogląd, że jest to osobliwość powiązana przez zasadę dualności nie z rezonansami, lecz właśnie z nierezonansowym tłem. Za taką koncepcją mógłby przemawiać fakt, że przebieg wysokoenergetyczny wszystkich amplitud (który jest opisywany wyłącznie przez osobliwość Pomeranczuka) jest taki sam dla tych procesów, w których obserwuje się rezonanse (np. rozpraszania π^+-N), jak i dla tych, w których rezonansów nie widać (np. K^+-N).

biegun Pomeranczuka

Powstaje też problem genezy osobliwości Pomeranczuka. Nie wiadomo, czy osobliwość ta jest tworem samodzielnym, czy też wynikiem jakiegoś sumowania się wkładów od innych, normalnych trajektorii.

Symetrie oddziaływań silnych i modele kwarków

Innym podejściem do teorii oddziaływań silnych, częściowo konkurencyjnym w stosunku do ujęcia dynamicznego i analitycznego, przedstawionego powyżej, a częściowo je uzupełniającym jest ujęcie tej teorii od strony symetrii (→ Cząstki elementarne i ich oddziaływania). Oddziaływania silne są wyróżnione fizycznie m.in. przez bogactwo swych własności symetrii wyrażonych w wielu prawach zachowania rozmaitych wielkości fizycznych. Z symetrii tych wynika nie tylko pewne uporządkowanie cząstek w multiplety (np. izospinu), ale też pewne ważne związki między amplitudami.

Biorąc jako przykład zderzenia mezonów K z nukleonami, można się przekonać, że istnieje pięć fizycznie różnych procesów, a mianowicie: rozpraszanie elastyczne K^+-p, K^+-n, K^0-p, K^0-n oraz proces wymiany ładunku $K^+ + n \to K^0 + p$. Jednakże warunek niezmienniczości izospinowej wyklucza możliwość, aby amplitudy oddziaływań silnych zależały od kierunku wektora izospinu całkowitego układu obu cząstek w przestrzeni izospinu, zezwalając tylko na to, aby amplitudy te zależały od długości tego wektora. Innymi słowy, amplitudy nie mogą za-

niezmienniczość izospinowa

leżeć od liczby kwantowej I_3, a tylko od całkowitego izospinu I. Dla dwu cząstek o izospinie $^1/_2$ (jak N i K) całkowity izospin może przybierać (zgodnie z zasadami składania momentu pędu) tylko dwie wartości, a mianowicie 0 i 1. Powinny więc istnieć najwyżej dwie niezależne amplitudy izospinowe, przez które muszą się wyrażać amplitudy wszystkich pięciu wymienionych wyżej procesów. Związki takie można znaleźć i następnie sprawdzić doświadczalnie. Jak dotąd ani w tym, ani w żadnym innym procesie nie napotkano wypadku łamania symetrii izospinowej przez oddziaływania silne (jest ona łamana przez oddziaływania elektromagnetyczne).

niezmienniczość względem $SU(3)$

Podobnie, choć na wyższym poziomie, przedstawia się zagadnienie niezmienniczości teorii oddziaływań silnych względem przekształceń grupy $SU(3)$. Rozpatrzmy przykład zderzeń binarnych dowolnych dwu cząstek należących jedna do oktetu mezonów pseudoskalarnych, a druga — do oktetu barionów o spinie $^1/_2$ i parzystości dodatniej. Uwzględniając już nawet związki wynikające z niezmienniczości izospinowej, wciąż jeszcze istnieje tu aż 36 *a priori* niezależnych amplitud (rozpraszanie elastyczne π-N, K-N, $\bar{K}$-N, η-N, π-Λ, K-Λ, $\bar{K}$-Λ, η-Λ, π-Σ, K-,Σ $\bar{K}$-Σ, η-Σ, π-Ξ, K-Ξ, $\bar{K}$-Ξ, η-Ξ, oraz amplitudy nieelastycznych procesów binarnych $\pi N \to \eta N$, $\pi\Xi \to \eta\Xi$, $\pi\Lambda \to \eta\Sigma$, $\pi\Sigma \to \eta\Sigma$, $K\Lambda \to K\Sigma$, $\bar{K}\Lambda \to \bar{K}\Sigma$, $\pi\Lambda \to \pi\Sigma$, $\eta\Lambda \to \pi\Sigma$, $\pi N \to K\Lambda$, $\pi N \to K\Sigma$, $\eta N \to K\Lambda$, $\eta N \to K\Sigma$, $\pi\Lambda \to \bar{K}N$, $\pi\Lambda \to K\Xi$, $\eta\Lambda \to \bar{K}N$, $\eta\Lambda \to K\Xi$, $\pi\Sigma \to \bar{K}N$, $\pi\Sigma \to K\Xi$, $\eta\Sigma \to \bar{K}N$, $\eta\Sigma \to K\Xi$). Z drugiej jednak strony z zasad składania reprezentacji grupy $SU(3)$ wynika, że $8 \otimes 8 = 1 \oplus 8 \oplus 8 \oplus 10 \oplus \overline{10} \oplus 27$. Istnieje więc tylko 6 niezależnych amplitud oddziaływań silnych przy zderzeniu, o którym tu mowa, jeśli oddziaływania te są niezmiennicze względem przekształceń grupy $SU(3)$. I te związki mogą być przedmiotem analizy doświadczalnej, choć nie należy oczekiwać, aby były one tak dobrze potwierdzone przez pomiar, jak w wypadku izospinu. Niestety, sprawdzanie jest tu dodatkowo utrudnione przez fakt, że tylko nieznaczna część spośród wymienionych amplitud dostępna jest pomiarom, ze względu na ograniczenia zarówno co do tarcz jak i co do wiązek (tarcze tylko nukleonowe, wiązki tylko π, K i $\bar{K}$).

Sprawdzanie tych relacji może nawet nie mieć, co najmniej w obszarze obecnie dostępnych energii, większego sensu ze względu na znany fakt łamania symetrii $SU(3)$ uwidaczniający się już w widmie mas. Gdyby bowiem symetria $SU(3)$ była ścisła, wszystkie masy cząstek należących do tego samego multipletu byłyby jednakowe, a tak nie jest. Jednakże Susumo Okubo i Murray Gell-Mann podali opis teoretyczny takiego łamania symetrii $SU(3)$, które prowadzi do uzyskania poprawnego widma mas. Założyli oni, że operator masy składa się z kilku części, z których jedna jest skalarem symetrii $SU(3)$ i daje wszystkim cząstkom tego samego multipletu te same masy, podczas gdy pozostałe dwie części są pewnymi składowymi oktetu, tak dobranymi, aby wszystkie cząstki należące do tego samego izomultipletu miały tę samą masę. Uzyskany wzór na widmo mas (wzór Gell-Manna-Okubo) ma następującą postać:

wzór Gell--Manna--Okubo

$$m = a + bY + c[I(I+1) - {}^1/_4 Y^2], \qquad (35)$$

gdzie $Y = B + S$ jest hiperładunkiem cząstki, I — jej izospinem, zaś a, b, c są trzema stałymi, zależnymi od multipletu symetrii $SU(3)$, w którym wzór (35) jest sprawdzany.

W zastosowaniu do barionów wzór (35) sprawdza się bardzo dobrze. Na przykład w oktecie, do którego należy nukleon, przyjmując znane z doświadczenia masy N, Λ i Σ można dostać przewidywaną masę hiperonu Ξ równą ok. 1328 MeV/c^2, podczas gdy masa rzeczywista, znana z doświadczenia, wynosi ok. 1320 Me/V/c^2. Lepszą zgodność z doświadczeniem uzyskuje się dla dekupletu barionów o spinie $^3/_2$.

W zastosowaniu do mezonów wzór (35) nieoczekiwanie nie przynosi tak dobrych rezultatów. W wyniku dalszych badań i poszukiwań wyjaśniono ostatecznie, że nie wszystkie obserwowane w przyrodzie cząstki należą do jakiegoś konkretnego multipletu $SU(3)$, lecz że są też takie które stanowią „mieszaninę" dwu różnych multipletów. Przykładów tej sytuacji jest sporo. Wszystkie one dotyczą mieszania oktetu z singletem w rozmaitych stanach spinu i parzystości. Okazało się np., że mezony φ i ω nie są ani czystymi stanami singletowymi, ani też oktetowymi, lecz pewną mieszaniną obu tych multipletów.

model kwarkowy

Zagadnienie to zostało wyjaśnione na podstawie modelu kwarkowego cząstek elementarnych. Okazało się mianowicie, że mezon φ jest zbudowany przede wszystkim (niemal dokładnie) z pary $s\bar{s}$, podczas gdy mezon ω z par kwarków niedziwnych, co dokładnie odpowiada obserwowanemu mieszaniu multipletów.

Z modeli kwarkowych wynikają także relacje między amplitudami zderzeń hadronów, przy pewnych założeniach dynamicznych dotyczących oddziaływań pomiędzy kwarkami. Taką najbardziej znaną relacją jest związek między całkowitymi przekrojami czynnymi na rozpraszanie π-N i N-N, który ma postać

$$\sigma_{\text{całk}}(\pi N) = {}^2/_3 \sigma_{\text{całk}}(NN). \qquad (36)$$

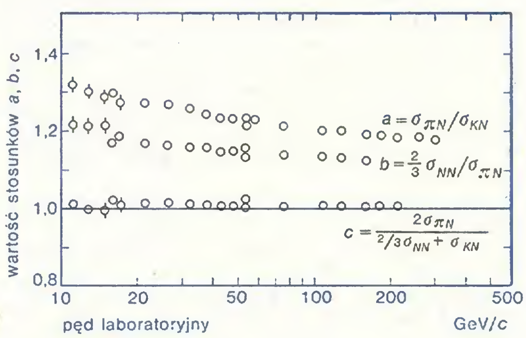
Wynika on bezpośrednio z liczenia kwarków. W mezonach π występuje para kwarków niedziwnych, a w nukleonach — trójka takich samych kwarków. W oddziaływaniu π-N mamy więc 2×3 oddziaływań kwarkowych, a w oddziaływaniu N-N 3×3 takich oddziaływań. Prowadzi to wprost do wzoru (36), przy założeniu, że amplitudy zderzeń indywidualnych kwarków po prostu można dodawać, aby dostać amplitudę zderzenia hadron-hadron. Podobnie z modelu kwarkowego można wysnuć wniosek, iż

$$\sigma_{\text{całk}}(\pi N) = \sigma_{\text{całk}}(KN), \qquad (37)$$

który razem z (36) prowadzi do wzoru:

$$2\sigma_{\text{całk}}(\pi N) = \sigma_{\text{całk}}(KN) + {}^2/_3 \sigma_{\text{całk}}(NN). \qquad (38)$$

Porównanie tych wzorów z danymi doświadczalnymi przedstawia rys. 30.



Rys. 30. Porównanie z doświadczeniem wzorów (36), (37) i (38)

Powstanie jednolitej teorii oddziaływań słabych i elektromagnetycznych, sukcesy modelu partonowego (→ Struktura cząstek elementarnych) w wyjaśnianiu danych dotyczących zderzeń głęboko nieelastycznych hadronów z leptonami i wreszcie wykrycie nowych cząstek, zawierających kwarki powabne, jeszcze bardziej zwiększyło zainteresowanie modelem kwarkowym budowy hadronów. Czyniono próby stworzenia teorii dynamicznej oddziaływań kwarków wewnątrz hadronu. W pracach tych zwykle rezygnuje się z nadziei, że kwarki kiedykolwiek będą mogły być zaobserwowane poza hadronami, a koncentruje się uwagę na wyjaśnieniu faktu uwięzienia kwarków w hadronach.

kwarki uwięzione

chromodynamika kwantowa

Jedną z takich koncepcji teoretycznych, powstałą stosunkowo niedawno, co do której istnieją uzasadnione nadzieje, że może się ona w przyszłości przerodzić w poszukiwaną od tak dawna teorię oddziaływań silnych, jest chromodynamika kwantowa. Jest to teoria opisująca oddziaływania między kwarkami przez wymianę cząstek podobnych do fotonów — gluonów. Zasadnicza różnica między fotonami a gluonami polega na tym, że gluony będąc skalarami zwykłej grupy $SU(3)$, czy też właściwie $SU(4)$ lub może nawet $SU(6)$, byłyby zarazem składowymi oktetu innej grupy $SU(3)$, a mianowicie grupy koloru (stąd nazwa „chromodynamika"). Gluony oddziaływałyby silnie nie tylko z kwarkami, ale też ze sobą nawzajem, w odróżnieniu od fotonów. Być może fakt ten mógłby wyjaśnić, dlaczego gluony nie są emitowane na zewnątrz hadronów (bo wiążą je inne gluony). W rezultacie w przyrodzie nie obserwuje się cząstek kolorowych, a tylko „białe" hadrony.

Chromodynamika kwantowa jak dotąd jednak nie jest w stanie tego wniosku uzasadnić ilościowo. W rezultacie w praktycznych obliczeniach fakt uwięzienia kwarków traktuje się jako dany z góry. Można to zrobić dwojako. W wersji bardziej wyszukanej, zakłada się, że oddziaływania między kwarkami i gluonami powodują powstanie pewnego pola samouzgodnionego, które trzyma wszystkie te cząstki wewnątrz hadronu oraz dają pewne słabe oddziaływania resztkowe, które w pierwszym przybliżeniu można nawet pominąć. Pole wiążące kwarki nazywa się czasem workiem. W modelu worka zakłada się pewną konkretną postać pola. Obliczone wówczas widmo mas hadronów zgadza się na ogół bardzo dobrze z doświadczeniem z wyjątkiem masy mezonu π, której obliczona wartość jest za duża oraz poza tym, że z obliczeń wynika istnienie również takich cząstek, których — jak dotąd — nie wykryto w doświadczeniu. Poza tym model worka nie pozwala na żadne przewidywania co do amplitud procesów między hadronami.

model worka

Mniej ambitna wersja modelu polega na założeniu, że uwięzienie kwarków można opisać wprowadzając po prostu pewien potencjał, który dla dużych odległości między kwarkami zmierza do nieskończoności, nie pozwalając im opuścić hadronu. Typowym przykładem takiego potencjału jest

$$V(r) = A/r + Br, \tag{39}$$

gdzie A i B są pewnymi stałymi. Potencjał taki wstawia się do równania Schrödingera i bada widmo stanów związanych. Wyniki przedstawione dla układu kwarków c-$\bar{c}$ (układ fizyczny zwany czarmonium, od ang. *charm* 'powab'), są w bardzo dobrej zgodności z doświadczeniem. Oczywiście i w tym wypadku nie można badać oddziaływań między hadronami, a tylko widmo mas, oraz inne własności pojedynczych hadronów.

Pomimo wszystkich tych ograniczeń wyniki uzyskane w modelach pochodzących od chromodynamiki kwantowej pozwalają mieć nadzieję na to, że konsekwentna teoria oddziaływań silnych wreszcie powstanie.

Obraz, który się być może wyłania z obecnie dostępnych danych, byłby więc taki, że oddziaływania silne, podobnie jak słabe i elektromagnetyczne byłyby przenoszone przez bozony o spinie 1, a liczba fundamentalnych cząstek obejmowałaby oprócz tych bozonów jeszcze tylko stosunkowo niewielką liczbę fermionów (być może 6 leptonów i 6 kwarków, te ostatnie — każdy w trzech kolorach). Chromodynamika kwantowa (albo inna teoria oparta na modelu kwarkowym) musi przede wszystkim wyjaśnić wszystkie te regularności dynamiczne, które są opisane w modelu biegunów Reggego. Sprawą zaś dalszej przyszłości jest zarysowująca się możliwość unifikacji także oddziaływań silnych ze słabymi i elektromagnetycznymi.

cztery kolory?

Istnieją poglądy, że leptony tworzą czwarty kolor, a więc pełna grupa symetrii koloru byłaby grupą $SU(4)$, a nie $SU(3)$. W tym wypadku kwarki mogłyby samorzutnie przechodzić w leptony, a dokładniej w układy cząstek zawierające leptony. Liczba barionów nie byłaby więc zachowana bezwzględnie, a proton byłby cząstką nietrwałą, z bardzo długim czasem życia, rzędu co najmniej 10^{22} lat. Sprawdzenie doświadczalne tych hipotez może więc być bardzo trudne.

P.D.B. Collins, E.J. Squires *Regge Poles in Particle Physics*, Berlin 1968 (ros. Moskwa 1971); R.J. Eden *High Energy Collisions of Elementary Particles*, Cambridge 1967 (ros. Moskwa 1970); D.H. Perkins *Introduction to High Energy Physics*, London 1972 (ros. Moskwa 1975); H. Pilkuhn *The Interactions of Hadrons*, Amsterdam 1967; Zob. też bibl. do Cząstki elementarne i ich oddziaływania.

Oddziaływania elektromagnetyczne

Michał Święcki

Oddziaływania elektromagnetyczne są to wszelkiego typu reakcje zachodzące między ładunkami elektrycznymi zarówno nieruchomymi, jak i pozostającymi w ruchu. W tym drugim przypadku mamy do czynienia z prądami elektrycznymi, które są źródłem oddziaływań magnetycznych. Wszystkie oddziaływania elektromagnetyczne zachodzą za pośrednictwem pola elektromagnetycznego, które w każdym punkcie wypełnionej nim przestrzeni działa siłą elektryczną na umieszczone w tej przestrzeni ładunki elektryczne, a także siłą magnetyczną — na ładunki, będące w ruchu (prądy, magnesy). Równocześnie źródłem pola elektromagnetycznego są ładunki elektryczne. Nieruchome ładunki oraz stałe prądy elektryczne są źródłami nie zmieniających się z upływem czasu pól statycznych, na przykład pola elektrostatycznego siły Coulomba bądź pola magnetostatycznego nieruchomego magnesu. Ruchome pojedyncze ładunki oraz prądy zmienne wywołują powstawanie zmiennego pola elektromagnetycznego. Zmiana (zaburzenie) pola elektromagnetycznego następuje lokalnie, w miejscu, w którym zaszła zmiana ładunku — źródła pola. Następnie zaburzenie to rozchodzi się po całej przestrzeni w postaci fali elektromagnetycznej. Poza miejscem i chwilą zajścia zaburzenia fala ta rozchodzi się swobodnie, aż do momentu napotkania innych ładunków, na które również lokalnie działa siłami elektrycznymi i magnetycznymi. Przekonamy się dalej, na czym polega owo rozchodzenie się pól elektromagnetycznych — nośników wszystkich oddziaływań elektromagnetycznych.

ładunki elektryczne a pole elektromagnetyczne

rola oddziaływań elektromagnetycznych

Oddziaływania elektromagnetyczne odgrywają wyjątkową rolę w świecie i w procesie jego poznawania. Decydują one o budowie atomów i cząsteczek chemicznych, a także o strukturze wszelkich ciał makroskopowych z wyjątkiem planet i gwiazd, gdzie istotne są również oddziaływania grawitacyjne. Nasze zmysły działają również dzięki oddziaływaniom elektromagnetycznym. Na nich też opiera się funkcjonowanie wszystkich przyrządów pomiarowych. Także tych, za pomocą których przeprowadzamy doświadczenia w świecie cząstek elementarnych. Można powiedzieć, że cała nasza ścisła wiedza o świecie była i jest zbierana za pomocą różnego typu oddziaływań elektromagnetycznych. Nic więc dziwnego, że oddziaływania te zostały najlepiej poznane i opisane przez fizykę. Istnieje bardzo elegancka i doskonale zgodna z doświadczeniem teoria oddziaływań elektromagnetycznych. Dla zjawisk makroskopowych jest to znana teoria Maxwella–Faradaya. W przypadku zjawisk

mikroskopowych, z których wynikają oczywiście wszelkie prawa makroskopowe, teorią tą jest elektrodynamika kwantowa (→ Elektrodynamika, Teoria pola). W tym artykule opiszemy zasady elektrodynamiki kwantowej i jej zastosowanie do opisu oddziaływań elektromagnetycznych cząstek elementarnych.

własności oddziaływań elektromagnetycznych

Nim do tego przejdziemy, zastanówmy się, dlaczego to właśnie oddziaływania elektromagnetyczne pełnią ową wyjątkową rolę w przyrodzie. Jest to związane z dwiema ich własnościami. Po pierwsze mają one nieskończony zasięg działania i łatwo je obserwować na odległościach makroskopowych. Oddziaływania grawitacyjne mają również tę własność, ale są wyjątkowo słabe i zaczynają odgrywać zauważalną rolę dopiero dla obiektów astronomicznych. Oddziaływania zaś silne i słabe mają zasięg działania niesłychanie mały ($\lesssim 10^{-15}$ m), znacznie mniejszy niż zdolność rozdzielcza stosowanej aparatury pomiarowej. Z drugiej strony oddziaływania elektromagnetyczne cząstek elementarnych są na tyle słabe, że przy ich opisie stosować można pewne metody przybliżone (diagramy lub inaczej grafy Feynmana), które opiszemy dalej Oddziaływania silne ważne przy opisie struktury jądra atomowego są natomiast zbyt mocne i dlatego ich opis teoretyczny nie jest już tak prosty.

elektrodynamika kwantowa

Elektrodynamika kwantowa traktuje o oddziaływaniach elektromagnetycznych cząstek elementarnych. W oddziaływaniach tych biorą udział wszystkie cząstki z wyjątkiem neutrin. Wszystkie one bowiem albo są obdarzone ładunkiem, albo zbudowane są ze składników naładowanych, jakimi są np. kwarki w hadronach (→ Cząstki elementarne i ich oddziaływania, Struktura cząstek elementarnych). Kluczową rolę we wszystkich oddziaływaniach elektromagnetycznych odgrywa foton. Nie ma on ładunku, a tylko oddziałuje z ładunkami innych cząstek przenosząc między nimi pęd i energię. Foton więc pełni w świecie cząstek elementarnych taką samą funkcję, jak pole elektromagnetyczne w zjawiskach makroskopowych. Rozpatrzmy dokładniej, na czym polega przenoszenie oddziaływań między cząstkami przez inne cząstki (np. przez foton). W tym celu musimy przypomnieć prawa kwantowe rządzące mikroświatem.

wzory de Broglie'a

Zgodnie z wynikami dotychczas przeprowadzonych doświadczeń cząstki elementarne rozchodzą się jak fale, których częstość i długość wiążą się z energią i pędem cząstek wzorami de Broglie'a:

$$E = h\nu, \qquad p = h/\lambda.$$

Fale te ulegają dyfrakcji i interferencji, zjawiskom charakterystycznym dla procesów rozchodzenia się cząstek. Natomiast bez względu na to, czy rozchodząca się fala elektronowa bądź fotonowa jest kulista czy też płaska (zależy to od własności źródła elektronów, czy fotonów) zawsze przekazuje ona całą swą energię, pęd, ładunek i inne mierzalne swe własności w postaci zlokalizowanych, bardzo małych przestrzennie porcji — w postaci cząstek. Tak więc cząstki elementarne rozchodzą się jak fale, a we wszystkich urządzeniach pomiarowych rejestrowane są jako bardzo małe obiekty. Żeby opisać ten paradoksalny wynik wielu doświadczeń wprowadzamy pojęcie fal prawdopodobieństwa.

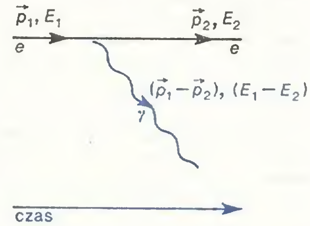
fale prawdopodobieństwa

Fale te opisują wszystkie możliwe drogi i sposoby rozchodzenia się cząstek. Rozchodzenie się to ma właśnie charakter falowy. Natężenie fal prawdopodobieństwa, czyli kwadrat ich amplitudy jest miarą prawdopodobieństwa określonego zachowania się cząstek. Tak więc mówimy jedynie o prawdopodobieństwie uzyskania określonego wyniku pomiaru położenia cząstki i prawdopodobieństwo to rozchodzi się jak fala. A jaki jest związek fali, na przykład fotonowej, z makroskopowym polem elektromagnetycznym? Otóż przy bardzo dużej mocy źródła wysyłającego fotony liczba możliwych rozróżnialnych doświadczalnie dróg, po których mogą się one rozchodzić, staje się znacznie mniejsza niż liczba samych fotonów. Chociaż więc w dalszym ciągu losy pojedynczego fotonu mogą być określone jedynie statystycznie, to jednak prawdopodobieństwo tego, że duża liczba jakichkolwiek fotonów znajdzie się w dowolnym miejscu jest bliskie jedności. I tak właśnie pojawia się makroskopowa fala elektromagnetyczna, która składa się z ogromnej liczby mikroskopowych fal fotonowych. Zerowa masa spoczynkowa fotonu jest odpowiedzialna za fakt, że fale elektromagnetyczne rozchodzą się z maksymalną możliwą prędkością $c \approx 300\,000$ km/s. Dwa możliwe ustawienia spinu fotonu odpowiadają dwom niezależnym stanom polaryzacji tych fal. Wysyłanie i pochłanianie fal elektromagnetycznych to nic innego, jak wysyłanie i pochłanianie fotonów przez ładunki elektryczne.

oddziaływanie elementarne

Z powyższych rozważań wynika, że elementarnym oddziaływaniem elektromagnetycznym jest wysłanie bądź pochłonięcie fotonu przez cząstkę naładowaną elektrycznie. Proces ten, podobnie jak oddziaływania makroskopowe, zachodzi lokalnie, czyli w pewnym punkcie przestrzeni i pewnej chwili czasu. Poza tym punktem i chwilą zarówno naładowana cząstka, jak i foton rozchodzą się zupełnie swobodnie, bez oddziaływań. Wynika z tego, że jedynie oddziaływania elektromagnetyczne cząstek elementarnych nie mających wewnętrznej struktury są łatwe do opisania. Takimi cząstkami prócz samego fotonu są wszystkie leptony. Natomiast hadrony mają bogatą strukturę — złożone są prawdopodobnie z kwarków i tylko oddziaływania elektromagnetyczne tych ostatnich są proste. Najpierw więc opiszemy oddziaływania elektromagnetyczne leptonów, przed omówieniem struktury hadronów.

Podstawowym procesem elektromagnetycznym jest wysłanie fotonu przez inną, naładowaną cząstkę. Jednak proces taki wydaje się niemożliwy z punktu widzenia praw zachowania energii i pędu. Elektron



Rys. 1. Emisja fotonu z elektronu. Masa fotonu nie może być równa zeru

o ściśle określonej (przez pomiar) energii E_1 i pędzie $\vec{p}_1$ (rys. 1) może po wysłaniu fotonu pozostać tym samym elektronem o energii E_2 i pędzie $\vec{p}_2$ tylko wtedy, gdy kwadrat masy spoczynkowej fotonu wynosi

$$q^2 = (1/c^4)[(E_2-E_1)^2-(\vec{p}_2-\vec{p}_1)^2c^2] \leqslant 0,$$

zgodnie z zasadami zachowania oraz relatywistycznym związkiem między energią i pędem. Tymczasem wiemy, że masa fotonu wynosi zero, co w powyższym wzorze może się zdarzyć tylko wtedy, gdy $E_1 = E_2$ oraz $\vec{p}_1 = \vec{p}_2$ (rozpraszanie do przodu).

fotony wirtualne

A jednak mogą istnieć, choć niezbyt długo, fotony o dowolnych wartościach masy. Są to tzw. fotony wirtualne. Ich pojawianie się wiąże się z zasadą nieokreśloności energii i czasu powszechnie obowiązującą w teoriach kwantowych. Zasada ta jest bezpośrednią konsekwencją związku między energią czastki i liczbą drgań na sekundę w fali z nią związanej, $E = h\nu$. Głosi ona, że minimalny czas Δt, po którym jakikolwiek obiekt fizyczny (np. przyrząd pomiarowy) może zarejestrować zmianę stanu badanej cząstki (na przykład fakt wysłania przez nią fotonu) wiąże się z nieokreślonością wartości energii tej cząstki ΔE nierównością.

$$\Delta E \cdot \Delta t \geqslant {}^1/_2\hbar.$$

Zasada ta obowiązuje często ze znakiem równości. Na przykład dla cząstki swobodnej iloczyn nieokreśloności czasu i energii dąży do ${}^1/_2\hbar$, gdy nieokreśloność energii dąży do zera, tzn. gdy energię cząstki wyznaczamy z dużą dokładnością. Nie ma jednak żadnego powodu formalnego (poza, być może, naszą jakże

masy spoczynkowe

często zawodną intuicją), żeby masy spoczynkowe cząstek były ściśle określone. W każdym przypadku omawianej reakcji emisji fotonu energie i pędy cząstek są określone bardzo dobrze — dla elektronów przez pomiary, dla fotonu wirtualnego przez zasady zachowania. Nie ma jednak zasady zachowania masy spoczynkowej. Obowiązuje dla niej tylko związek

$$m_0^2c^4 = E^2 - p^2c^2$$

i stąd właśnie wyznaczamy tę masę. Tak więc w każdym pojedynczym akcie emisji masa jest określona przez ten związek, chociaż ani nie musi być ona np. równa zeru dla fotonu, ani być taka sama we wszystkich aktach emisji. W wielokrotnie powtórzonym doświadczeniu z emisją fotonu energie i pędy elektronów końcowych będą za każdym razem inne dając różne masy spoczynkowe fotonu wirtualnego.

Musimy teraz uwzględnić relatywistyczną niezmienniczość teorii. Wiemy przecież, że wartości zarówno energii jak i czasu (podobnie pędu i położenia) zależą od układu odniesienia i musimy dokonać wyboru odpowiedniego układu, uniwersalnie dla wszystkich przypadków danej reakcji. Najwygodniej jest związać wybrany układ z własnościami fotonu wirtualnego. I tak na przykład, w opisanej reakcji emisji fotonu mamy $q^2 \leqslant 0$ i można wybrać układ odniesienia (tzw. układ Breita), w którym $\vec{p_1} = \vec{p_2}$. Wtedy $E_1 = E_2$, energia fotonu jest równa zeru, a jego pęd wiąże się z wartością masy: $p_f = c\sqrt{-q^2}$. W rozważanej dalej reakcji anihilacji pary elektron-pozyton z wytworzeniem fotonu wirtualnego kwadrat masy fotonu jest dodatni, $q^2 > 0$, i wygodny jest układ odniesienia (tzw. układ środka masy), w którym również $\vec{p_1} = -\vec{p_2}$. Kinematyka jest tu jednak inna i mamy $q^2 = (1/c^4) \cdot [(E_1+E_2)^2 - (\vec{p_1}+\vec{p_2})^2c^2] > 0$, skąd $E_f = c^2\sqrt{q^2}$ przy $p_f = 0$. W tym ostatnim przypadku wybraliśmy więc po prostu układ spoczynkowy fotonu wirtualnego (dla $q^2 \leqslant 0$ układ taki nie istnieje) i w tym układzie nieokreśloność energii fotonu równa się nieokreśloności jego masy ($\Delta E = c^2 \cdot \Delta m$). Ponieważ masa fotonu rzeczywistego wynosi zero, więc dla pojedynczego aktu emisji z określoną wartością q^2 nieoznaczoność masy równa się wartości samej masy. Wtedy z zasady nieokreśloności wynika

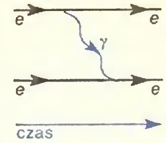
$$\Delta t \cdot \sqrt{q^2} = \hbar/2c^2,$$

gdzie Δt jest czasem trwania emisji, czyli czasem, podczas którego żaden przyrząd pomiarowy nie jest w stanie stwierdzić pojawienia się fotonu o niefizycznej wartości masy. Po upływie tego czasu foton taki może być dostrzeżony i przedtem musi być z powrotem pochłonięty. Zupełnie podobne rozważania można przeprowadzić w przypadku, gdy $q^2 \leqslant 0$. Wtedy jednak należy mówić o nieoznaczoności położenia i pędu w układzie Breita. Mamy podobnie $\Delta p = c \cdot \Delta m$, a w pojedynczym akcie emisji $\Delta m = \sqrt{-q^2}$. Z zasady nieokreśloności pędu i położenia mamy teraz $\Delta x \cdot \sqrt{-q^2} = \hbar/2c$, gdzie Δx oznacza rozmiary obszaru, w którym zaszła emisja. Foton wirtualny nie może się pojawić w odległości większej niż Δx, gdyż wtedy mógłby zostać zaobserwowany. Wprowadzając umownie i w tym wypadku charakterystyczny czas równy $\Delta x/c$ można stale posługiwać się pojęciem czasu oddziaływania, danego uniwersalnie przez związek $\Delta t = \hbar/2\sqrt{q^2}\, c^2$. Wszystko to obowiązuje jedynie dla cząstek swobodnych, dla których nie występuje nieokreśloność energii oddziaływania. Wiemy jednak, że w teorii lokalnej swobodne są wszystkie cząstki, w tym również wirtualne. I tak, w pojedynczym akcie emisji wysłany z elektronu foton o kwadracie masy q^2, wyznaczonym przez pomiary dla elektronu, może zostać zauważony dopiero po czasie

nieoznaczoność czasu emisji fotonu

$$\Delta t = \frac{\hbar}{2\sqrt{-q^2}\, c^2}.$$

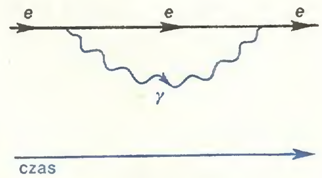
Tyle właśnie wynosi nieoznaczoność czasu emisji fotonu, choć jak wiemy czas ten jest tylko chwilą (punktem czasowym). Po upływie czasu Δt wysłany foton musi zostać pochłonięty przez inną bądź tę samą cząstkę (rys. 2 i 3). W pierwszym wypadku następuje przekazanie pędu i energii drugiej cząstce, czyli jej rozproszenie. I na tym polega przekazywanie oddziaływań elektromagnetycznych przez fotony.



Rys. 2. Rozpraszanie. Foton wirtualny został pochłonięty przez inny elektron

Niedostrzeżony przez czas Δt foton wirtualny może w tym czasie przebyć drogę równą $c \cdot \Delta t$. Ponieważ dla rozpraszania do przodu $q^2 \to 0$, więc $\Delta t \to \infty$ i w tych warunkach droga przebyta przez foton może być dowolnie duża. Dlatego właśnie mówimy, że oddziaływania elektromagnetyczne mają zasięg nieskończony. Odpowiada temu statyczny potencjał oddziaływania $V \sim 1/r$. Gdyby cząstka przenosząca oddziaływanie nie miała zerowej masy, $m_0 \neq 0$ (m_0 jest tu masą cząstki rzeczywistej), to dla $q^2 \to 0$ nieokreśloność

potencjał oddziaływania



Rys. 3. Absorpcja fotonu wirtualnego przez ten sam elektron

masy nie dążyłaby do zera ale do m_0. Wtedy $\Delta t \to \hbar/2m_0c^2$ i zasięg takiego oddziaływania wynosiłby $c \cdot \Delta t = \hbar/2m_0c$. Sytuacja tego rodzaju występuje przy oddziaływaniach silnych oraz słabych i odpowiada jej statyczny potencjał oddziaływania zwany potencjałem Yukawy $V \sim (1/r)\exp[\ (\hbar/m_0c)r]$.

Przejdźmy teraz do obliczenia prawdopodobieństwa zajścia niektórych procesów elektromagnetycznych. Musimy przedtem ustalić, w jaki sposób będziemy opisywali wyniki doświadczeń przeprowadzanych nad cząstkami. Przede wszystkim zauważmy, że z zasady nieokreśloności energii i czasu, a także z podobnej zasady obowiązującej dla pędu i położenia wynika, że reakcje zachodzące między cząstkami elementarnymi mogą być opisywane albo za pomocą położeń i czasu, albo za pomocą energii i pędów cząstek. Opis przy użyciu obu tych zbiorów zmiennych równocześnie nie jest możliwy. W doświadczeniach potrafimy mierzyć energie i pędy cząstek z dokładnością sięgającą jednego procenta, podczas gdy dokładność bezpośredniego pomiaru położenia i czasu jest o wiele rzędów wielkości gorsza niż rozmiary cząstek i czasy trwania procesów zachodzących między nimi. Nic też dziwnego, że przy opisie oddziaływań między cząstkami posługujemy się pojęciem prawdopodobieństwa uzyskania określonych wartości energii i pędów cząstek, nic nie mówiąc o ich położeniach. Podobnie fale opisujące zachowanie się cząstek będą zależały od pędów i energii, podczas gdy ich własności czasoprzestrzenne nie będą nas interesowały. Zamiast pojęcia przestrzennych fal prawdopodobieństwa wprowadzimy ogólniejsze pojęcie amplitudy prawdopodobieństwa. Amplituda ta w wypadku reakcji zachodzących między cząstkami elementarnymi zależy od pędów i energii cząstek. Jej kwadrat daje prawdopodobieństwo uzyskania określonego wyniku doświadczenia. Amplitudami prawdopodobieństwa posługujemy się tak, jak samym prawdopodobieństwem, tzn. mnożymy amplitudy odpowiadające procesom niezależnym bądź warunkowym oraz dodajemy amplitudy odpowiadające procesom nierozróżnialnym doświadczalnie. W tym ostatnim przypadku mamy właśnie do czynienia ze zjawiskami interferencji wszelkiego rodzaju.

amplitudy prawdopodobieństwa

Obliczymy amplitudę prawdopodobieństwa związaną z opisanym uprzednio procesem rozpraszania dwóch cząstek naładowanych, zachodzącym przez wymianę fotonu wirtualnego o kwadracie masy równym q^2. W rachunkach pominiemy spiny cząstek, tak że uzyskane wyniki nie będą ścisłe. Załóżmy też, że rozpraszane cząstki nie są identyczne. Może to być

np. rozpraszanie mionu μ^- na elektronie. W przypadku cząstek identycznych, np. elektronów, pojawią się dwa nierozróżnialne stany końcowe, w których rozproszone cząstki mają zamienione pędy i energie. Amplitudy prawdopodobieństwa zajścia tych dwóch zdarzeń należy dodać, pojawia się interferencja, która nieco komplikuje rachunki.

Amplituda prawdopodobieństwa zajścia procesu emisji fotonu nie może zależeć bezpośrednio od pędów i energii cząstek uczestniczących w tej reakcji. Wtedy bowiem prawdopodobieństwo emisji zależałoby od układu odniesienia, z którego obserwujemy ten proces. Amplituda może więc zależeć jedynie od takich kombinacji pędów i energii cząstek, które nie zmieniają się przy zmianie układu odniesienia. Takimi kombinacjami są masy spoczynkowe cząstek — stałe — oraz zmienna masa $\sqrt{-q^2}$ charakteryzująca wysłany foton wirtualny. Amplituda prawdopodobieństwa wysłania fotonu wirtualnego jest więc funkcją q^2 (pominęliśmy spiny cząstek). Jaka jest ta zależność? Najprościej przyjąć, że amplituda jest wprost proporcjonalna do czasu trwania procesu emisji Δt (oznacza to, że każda chwila wysłania jest równie prawdopodobna), czyli odwrotnie proporcjonalna do $\sqrt{-q^2}$. Stałą proporcjonalności nazywamy ładunkiem elektrycznym cząstki wysyłającej. Taka definicja ładunku jest zgodna z definicją makroskopową (poprzez prawo Coulomba). Tak więc elementarna amplituda prawdopodobieństwa równa się $e_1/\sqrt{-q^2}$, gdzie e_1 jest ładunkiem cząstki wysyłającej. Podobnie amplituda prawdopodobieństwa zajścia procesu absorpcji fotonu wynosi $e_2/\sqrt{-q^2}$, gdzie e_2 jest ładunkiem cząstki pochłaniającej. Wreszcie amplituda zajścia procesu absorpcji pod warunkiem, że uprzednio wystąpił proces emisji jest oczywiście iloczynem obu powyższych amplitud:

ładunek elektryczny

$$A(q^2) = -\frac{e_1 e_2}{q^2}$$

(Często zamiast zmiennej q^2 wprowadza się zmienną $t = q^2$). Prawdopodobieństwo uzyskania w wyniku rozpraszania takiej konfiguracji pędów i energii końcowych, której odpowiada określona wartość q^2, jest dane przez kwadrat tej amplitudy. Przekrój czynny, który jest miarą prawdopodobieństwa rozpraszania, przy odpowiednich jednostkach ładunku wynosi

$$\frac{d\sigma}{dq^2} \sim \frac{(4\pi)^2 e_1^2 e_2^2}{(\hbar c)^2} \frac{1}{q^4}.$$

We wzorze tym nie uwzględniono normalizacji przekroju czynnego (prawdopodobieństwo na jednostkę strumienia) i stąd znak proporcjonalności. Uwzględnienie spinów cząstek komplikuje znacznie rachunki, gdyż amplitudy prawdopodobieństwa mogą być także funkcją takich iloczynów pędów i spinów, które nie zależą od układu odniesienia. Jeżeli jednak badamy rozpraszanie cząstek nie identycznych i nie interesujemy się stanami polaryzacji (ustawienia spinu) cząstek, to powyższy wzór poprawnie opisuje wyniki doświadczalne.

Zupełnie podobnie można obliczyć przekroje czynne innych procesów elektromagnetycznych związanych z wysyłaniem lub pochłanianiem fotonów. Nie będziemy przedstawiać wyników tych obliczeń, musimy jednak zwrócić uwagę na podstawową własność amplitud prawdopodobieństwa, a mianowicie na ich uniwersalność.

uniwersalność amplitud prawdopodobieństwa

Uniwersalność amplitud prawdopodobieństwa polega na tym, że jedna amplituda opisuje wiele procesów. Korzystaliśmy już z tej własności, przyjmując że amplitudy wysłania i pochłonięcia fotonu wirtualnego są dane tym samym wzorem. W tym wypadku było to dosyć oczywiste, gdyż procesy wysyłania i pochłaniania fotonu różnią się formalnie jedynie kierunkiem biegu czasu, który nie występuje w przyjętym przez nas opisie rozpraszania poprzez pędy i energie cząstek. Pochłaniany lub wysyłany foton nie niesie żadnych wielkości (liczb kwantowych), które podlegają prawom zachowania, z wyjątkiem pędu i energii. Foton ma zerowe wartości ładunku elektrycznego, leptonowego oraz barionowego, nie zmienia więc tych ładunków u cząstek, z którymi oddziałuje. Zmienia tylko pęd i energię; a także spin, ale ten nie wnosi tu nic istotnego. Mówimy, że foton jest identyczny ze swoją antycząstką, co z definicji znaczy, że ma zerowe wartości wszystkich zachowywanych liczb kwantowych typu ładunkowego. Dlatego właśnie procesy pochłaniania i wysyłania fotonu są opisywane tą samą amplitudą. Wprawdzie pęd i energia fotonu w pierwszym wypadku dodają się, a w drugim — odejmują od pędu i energii elektronu padającego, ale amplituda zależy jedynie od $q^2 = (1/c^4)[E_f^2 - p_f^2 c^2]$. Znaki E_f oraz p_f nie mają więc znaczenia. Zupełnie inaczej jest w wypadku emisji i absorpcji cząstek, które mają niezerową wartość przynajmniej jednej spośród zmiennych ładunkowych. Wysłanie na przykład elektronu z pewnego układu, to zwiększenie ładunku elektrycznego tego układu o jeden. Pochłonięciu elektronu odpowiada natomiast zmniejszenie ładunku o jeden. Podobnie jest z innym rodzajem ładunków, np. ładunkiem leptonowym elektronu.

Gdyby jednak istniała cząstka, która miała by przeciwny niż elektron znak ładunku elektrycznego oraz leptonowego, a poza tym niczym nie różniła się od elektronu, to emisja jej mogłaby być bardzo podobna do absorpcji elektronu. Taką cząstką jest pozyton, antycząstka elektronu.

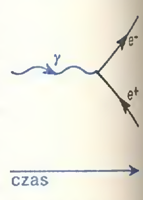
pozyton — antycząstka elektronu

Zwróciliśmy poprzednio uwagę na fakt, że amplituda prawdopodobieństwa wysłania (pochłonięcia) fotonu wirtualnego może zależeć jedynie od mas cząstek (spiny stale pomijamy) — stałych mas spoczynkowych elektronów rzeczywistych i zmiennej masy fotonu wirtualnego. Stałą proporcjonalności we wzorze na amplitudę jest ładunek elektryczny. Jest więc zupełnie oczywiste, że amplituda prawdopodobieństwa wysłania fotonu wirtualnego z pozytonu, cząstki o tej samej masie co elektron, ale przeciwnym ładunku, różni się od odpowiedniej amplitudy dla elektronu jedynie pewnym czynnikiem fazowym (np. znakiem).

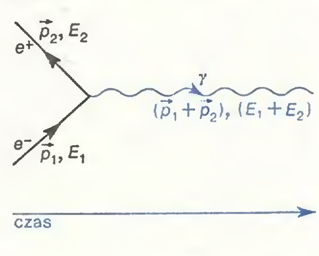
Ta równość amplitud reakcji z udziałem cząstek i antycząstek nosi nazwę niezmienniczości ze względu na operację sprzężenia ładunkowego. Dodajmy tu jeszcze, że oddziaływania elektromagnetyczne są również niezmiennicze ze względu na operacje inwersji przestrzennej i zmiany kierunku czasu. Te dwie ostatnie ich własności znane są już z fizyki klasycznej.

niezmienniczość ze względu na operację sprzężenia ładunkowego

Zauważmy, że proces wysłania fotonu z elektronu można rozumieć w następujący sposób: w pewnym punkcie przestrzeni oraz w pewnej chwili następuje pochłonięcie padającego elektronu i wysłanie elektronu wraz z fotonem. Wtedy podobny proces z udziałem pozytonu powstaje przez zamianę pochłanianego elektronu na wysyłany pozyton, a wysyłanego elektronu na pochłaniany pozyton. Odpowiednie amplitudy różnią się jedynie znakiem. A co będzie, jeżeli tylko jeden z elektronów zamienimy na pozyton (rys. 4, 5)? Powstaną wtedy zupełnie inne reakcje — kreacja pary elektron-pozyton przez foton wirtualny oraz anihilacja pary na foton. Ponieważ założyliśmy raz na za wsze, że amplituda prawdopodobieństwa jest proporcjonalna do nieokreśloności czasu zajścia oddziaływania, a przez stałą proporcjonalności zdefiniowaliśmy ładunek elektryczny, więc nic dziwnego, że te nowe procesy są opisywane znów przez amplitudę



Rys. 5. Kreacja pary elektron-pozyton przez foton wirtualny



Rys. 4. Anihilacja pary elektron-pozyton na foton wirtualny

$e/\sqrt{-q^2}$ z odpowiednią, określoną przez prawa zachowania wartością q^2. Tak więc jedna amplituda opisuje wszystkie omówione poprzednio procesy z udziałem fotonu wirtualnego.

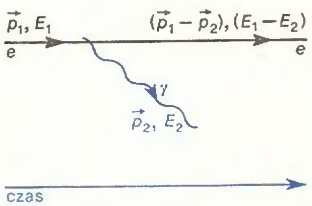
Spróbujmy odwrócić tok powyższego rozumowania. Pokazaliśmy, że amplituda prawdopodobieństwa pochłonięcia fotonu przez elektron przechodzi w amplitudę kreacji pary elektron-pozyton przez foton, jeżeli pęd i energia padającego elektronu zmienią znaki (poprzednio pęd i energia fotonu były różnicami pędów i energii leptonów, teraz są ich sumą) i staną się odpowiednimi wielkościami wysyłanego pozytonu (o przeciwnym znaku ładunku). Korzystaliśmy przy tym z faktu istnienia pozytonu. Nie jest to jednak założenie konieczne, bowiem fakt ten wynika już z samych zasad przedstawianej teorii. Amplituda prawdopodobieństwa absorpcji fotonu, $e/\sqrt{-q^2}$, nie znika przecież, jeżeli pęd i energia elektronu zmienią znak. Tyle tylko, że teraz q^2 dane jest innym niż poprzednio wyrażeniem

$$q^2 = \frac{1}{c^4}[(E_2+E_1)^2-(\vec{p}_2+\vec{p}_1)^2c^2] > 0.$$

Poprzednio $q^2 \leqslant 0$. Funkcja $e/\sqrt{-q^2}$ nie znika ani dla dodatnich, ani ujemnych wartości q^2. Nie znika więc amplituda, a z nią i prawdopodobieństwo zajścia odpowiednich reakcji. Jeżeli prawdopodobieństwo zajścia procesu kreacji pary z fotonu nie znika, to znaczy, że pary takie muszą być czasem produkowane. A zatem musi istnieć antycząstka elektronu o takiej samej co elektron masie, lecz o ładunku elektrycznym przeciwnego znaku. Jest to prawo obowiązujące ogólnie w świecie cząstek elementarnych. Wszystkie one muszą mieć i rzeczywiście mają antycząstki o takich samych masach (i spinach), ale przeciwnego znaku ładunkach wszystkich rodzajów. Niektóre cząstki, jak np. foton, czy też mezon π^0 są tożsame ze swymi antycząstkami.

diagramy Feynmana

Przedstawione poprzednio rysunki obrazujące wysyłanie i pochłanianie fotonów wirtualnych przez elektrony i pozytony (bądź jakiekolwiek inne cząstki naładowane) oraz anihilację i kreację par zwane są diagramami (grafami) Feynmana. Są to elementarne diagramy Feynmana opisywane przez tę samą amplitudę prawdopodobieństwa. Jednak w tych elementarnych procesach nie tylko foton może być wirtualny. Wiemy przecież, że fotony, podobnie jak i fale elektromagnetyczne mogą rozchodzić się swobodnie na ogromne odległości i wtedy ich czas życia jest nieskończony, a masa określona i równa zeru. Jeżeli w elementarnym procesie produkcji foton jest rzeczywisty, to wirtualny musi być jeden z elektronów lub pozytonów. Wymagają tego prawa zachowania



Rys. 6. Emisja fotonu rzeczywistego z elektronu. Elektron końcowy nie może mieć masy równej masie elektronu rzeczywistego

elektron wirtualny

energii i pędu. Tak więc możliwe są procesy, w których np. padający elektron wysyła rzeczywisty foton przechodząc w elektron wirtualny o masie określonej przez prawa zachowania (rys. 6). Amplituda tego procesu jest inna, gdyż nieokreśloność masy wynosi teraz nie $\sqrt{-q^2}$, ale $\sqrt{-q^2+m_0^2}$, gdzie m_0 jest masą spoczynkową elektronu rzeczywistego. W związku z tym $\Delta t \sim 1/\sqrt{-q^2+m_0^2}$, ale, wciąż pomijając spin, jest to jedyna zmiana, gdyż stała proporcjonalności (ładunek) jest znów ta sama, co poprzednio. Moglibyśmy też rozważać procesy, w których nie jedna, ale dwie albo nawet wszystkie trzy cząstki są wirtualne. Rachunki nieco się wtedy komplikują, ale zasada pozostaje ta sama. Tak więc elementarny diagram produkcji wraz ze wszystkimi jego odmianami jest jedyną cegiełką, z której zbudowana jest elektrodynamika. Jest wielkim sukcesem fizyki, że teoria oparta na tak prostych zasadach bezbłędnie opisuje wszystkie bez wyjątku doświadczenia.

ładunki cząstek elementarnych

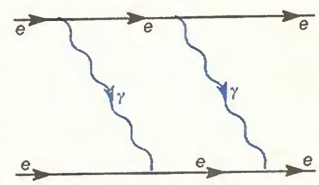
Chcielibyśmy zwrócić jeszcze uwagę na pewną ważną cechę cząstek elementarnych. Mają one mianowicie jedynie całkowite wartości wszystkich swych ładunków. Jeśli chodzi o interesujące nas tu ładunki elektryczne, to okazuje się, że wszystkie znane cząstki pozbawione struktury wewnętrznej mają ładunki ± 1 albo 0 (w jednostkach ładunku elektronu). Leptony (e^-, μ^-) mają ładunki -1, antyleptony $+1$, a neutrina i foton są elektrycznie obojętne. Obdarzone strukturą hadrony mogą mieć ładunki większe, ale zawsze całkowite. Dopiero ich nieobserwowane jako cząstki swobodne składniki — kwarki — mają ładunki wynoszące $^1/_3$ bądź $^2/_3$ ładunku elementarnego. Zarówno owo skwantowanie ładunków elektrycznych u cząstek obserwowanych doświadczalnie, jak i określone odstępstwo od niego dla uwięzionych w hadronach kwarków nie zostały dotychczas wyjaśnione przez teorię.

W związku z tą ciekawą własnością cząstek wygodną miarą prawdopodobieństwa wysłania fotonu z jakiejkolwiek cząstki jest kwadrat ładunku elementarnego. Wprowadza się więc nową bezwymiarową stałą uniwersalną,

stała struktury subtelnej

$$\alpha_{el} = e^2/4\pi\varepsilon_0\hbar c \approx {}^1/_{137},$$

zwaną stałą struktury subtelnej albo stałą sprzężenia oddziaływań elektromagnetycznych. Poza zależnością od q^2 prawdopodobieństwo emisji jest zawsze dane przez α_{el} razy jeden bądź zero ($^1/_9$ lub $^4/_9$ dla kwarków). Zwróćmy uwagę na stosunkowo niewielką wartość elektromagnetycznej stałej sprzężenia, $\alpha_{el} \approx$ $\approx {}^1/_{137} \ll 1$. Jest to bardzo ważna własność. Dzięki temu bowiem mówiąc o rozproszaniu leptonów mogliśmy ograniczyć się do procesów z wymianą tylko jednego fotonu wirtualnego. Prawdopodobieństwo



Rys. 7. Wymiana dwóch fotonów wirtualnych przy rozpraszaniu elektronów jest mało istotna

wysłania (i następnie pochłonięcia) dwóch, trzech i więcej fotonów nie jest równe zeru (rys. 7). Jednak wymiana każdego fotonu wiąże się zawsze z pojawieniem się w amplitudzie prawdopodobieństwa czynnika $\alpha_{el} \approx {}^1/_{137}$. Wszystkie procesy wyższych rzędów są więc bardzo mało prawdopodobne i mogą być pominięte. W dokładnych rachunkach, służących do opisu bardzo precyzyjnych pomiarów, trzeba oczywiście uwzględnić również te nieznaczne wkłady. Jednak zawsze wystarczy ograniczyć się do niewielkiej stosunkowo liczby procesów wielofotonowych. Tak więc jedynie dzięki małej wartości stałej sprzężenia możemy wykonywać efektywne obliczenia dla reakcji pochodzenia elektromagnetycznego. Zupełnie inaczej wygląda sprawa z oddziaływaniami silnymi, których stała sprzężenia jest duża, $\alpha_{sil} \approx 10$. Powoduje to, że kolejne wyższe rzędy oddziaływania są coraz większe i opis teoretyczny bardzo się komplikuje. Dopiero oddziaływania składników hadronów — kwarków — mogą być w pewnych warunkach niezbyt silne. Nie są też oczywiście silne oddziaływania elektromagnetyczne zarówno hadronów, jak i kwarków.

Opiszemy teraz krótko różnego rodzaju reakcje pochodzenia elektromagnetycznego. Każdą z nich będziemy ilustrować odpowiednim diagramem Feynmana, ograniczając się zawsze do diagramów najniższego rzędu. Zaczniemy od przedstawienia oddziaływań elektromagnetycznych leptonów — cząstek bez struktury wewnętrznej.

Rozpraszanie kulombowskie. Stany związane

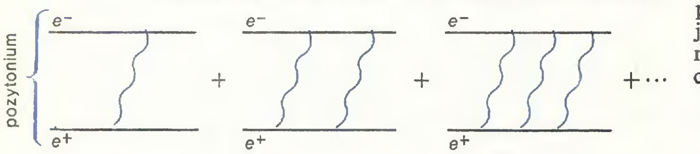
Reakcję tę omawialiśmy szczegółowo poprzednio. Rozpraszać się mogą różne leptony obdarzone ładunkiem. Ponieważ ładunki elektryczne wszystkich leptonów różnią się jedynie znakiem, więc wszystkie przekroje czynne są proporcjonalne do tej samej stałej α^2_{el}. Różnice pojawiają się jedynie w wypadku dwóch cząstek identycznych: interferencja dwóch procesów prowadzących do nierozróżnialnych sytuacji jest wtedy powodem znikania prawdopodobieństwa zajścia rozpraszania pod kątem $\pi/2$ (w układzie środka masy). Pomijając ten wypadek i pomijając spiny cząstek dostajemy zależność:

przekrój czynny na rozpraszanie kulumbowskie

$$\frac{d\sigma}{dq^2} \sim \frac{\alpha^2_{el}}{q^4}.$$

Widzimy, że przy rozpraszaniu nie ma znaczenia, czy rozpraszane cząstki mają ładunki jedno- czy różnoimienne. Nie ma więc znaczenia, czy przyciągają się czy też odpychają wzajemnie. Sytuacja ta niczym istotnym nie różni się od rozpraszania makroskopowych cząstek klasycznych. Klasyczna cząstka rozpraszana na centrum potencjału kulombowskiego porusza się po hiperboli zagiętej w kierunku centrum dla potencjału przyciągającego i odginającej się od centrum dla potencjału odpychającego. Wiązka cząstek przelatujących w różnych odległościach od centrum odchylana jest po różnych hiperbolach, co daje pewne prawdopodobieństwo (przekrój czynny) uzyskania rozpraszania pod pewnym kątem. Rozkład kierunków nie będzie oczywiście zależał od tego, czy kierunki te pochodzą z hiperbol zagiętych w jedną, czy też w drugą stronę. Warto przy okazji zwrócić uwagę na fakt, że w przybliżeniu nierelatywistycznym (potencjalnym) klasyczny wzór na przekrój czynny jest identyczny z wzorem kwantowym. Dlatego tylko Rutherford mógł wyciągnąć ze swych doświadczeń poprawne wnioski o budowie atomu, choć przecież nie znał zupełnie mechaniki kwantowej.

Tak więc znaki ładunków elektrycznych nie mają wpływu na procesy rozpraszania cząstek, decydują jedynie o ewentualnym pojawianiu się stanów związanych cząstek. W ujęciu klasycznym odpowiada to pojawianiu się zamkniętych eliptycznych lub kołowych orbit w przypadku potencjału przyciągającego. Natomiast w ujęciu kwantowym zjawisko to jest związane z pojawianiem się procesów wymiany wielu fotonów



Rys. 8. Wymiana dowolnej liczby fotonów wirtualnych nie może być pominięta dla stanu związanego elektronu z pozytonem

(rys. 8). W reakcjach rozpraszania procesy te są nieistotne ze względu na małą wartość stałej sprzężenia, $\alpha_{el} \approx {}^1/_{137}$. Argument ten jest jednak oparty na założeniu, że czas oddziaływania jest skończony. W przypadku rozpraszania jest to prawda — po krótkim czasie cząstki rozlatują się swobodnie. W zlokalizowanych przestrzennie niewielkich układach związanych cząstek czas oddziaływania jest nieskończony. Cząstki stale pozostają w pobliżu siebie i bez względu na wartość stałej sprzężenia prawdopodobieństwo wymiany dowolnej liczby fotonów jest równe jedności. Wtedy wszystkie przedstawione na rysunku diagramy Feynmana są równie ważne i decyduje ich suma. Z formalnego punktu widzenia szereg perturbacyjny nie jest więc zbieżny dla stanu związanego. Szereg ten w wypadku dwóch cząstek o przeciwnych ładunkach zawiera wyrazy o innej sekwencji znaków niż dla cząstek o ładunkach jednakowych. Tylko w jednej z tych sytuacji szereg jest rozbieżny przy pewnej wartości energii pary cząstek równej, oczywiście, masie stanu związanego. Widać więc wyraźną różnicę, chociaż ścisły dowód na to, że tylko w jednym z tych wypadków otrzymujemy trwały stan związany, nie jest prosty. Warto zwrócić uwagę na fakt, że problem stanów związanych w elektrodynamice kwantowej nie został jeszcze w pełni rozwiązany. Nie umiemy bowiem radzić sobie z problemami, przy których rachunek zaburzeń nie jest zbieżny.

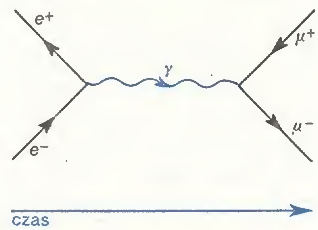
stany związane cząstek

Prócz atomu wodoru i innych atomów znane są również stany związane bardziej egzotyczne: pozytonium — stan związany elektronu i pozytonu oraz atomy egzotyczne (→ Atomy egzotyczne) — stany związane μ^- proton, π^- proton, K^- proton, antyproton proton itp. Te ostatnie nie są stanami absolutnie trwałymi, gdyż nietrwałe są ich składniki, chociaż ich czas życia jest stosunkowo długi w porównaniu z czasem charakterystycznym dla oddziaływań elektromagnetycznych. Jeśli chodzi o pozytonium, to zdarza się, jak się dalej przekonamy, że elektron i pozyton ulegają anihilacji na dwa lub trzy fotony rzeczywiste. Z tego powodu pozytonium też nie jest absolutnie trwałe, choć żyje stosunkowo długo.

atomy egzotyczne

Anihilacja wraz z kreacją pary

Diagram Feynmana takiego właśnie procesu został przedstawiony na rysunku 9. Jeśli para anihilująca i para kreowana są takie same (np. $e^+e^- \to e^+e^-$), to ten sam stan końcowy może być również uzyskany

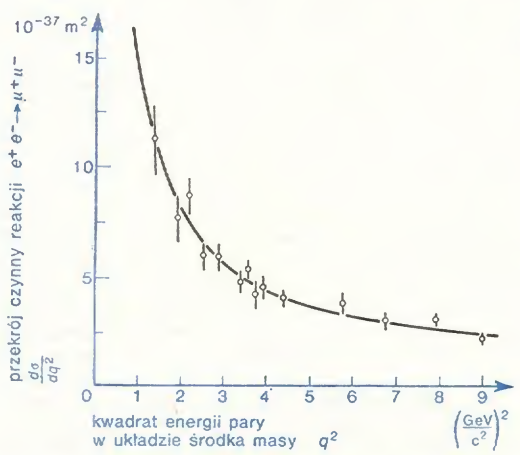


Rys. 9. Anihilacja pary elektron-pozyton wraz z następującą po niej kreacją pary mionów $\mu^-\mu^+$

w wyniku opisanego poprzednio procesu rozpraszania cząstki na antycząstce. Odpowiednie amplitudy trzeba wtedy dodać i dopiero potem obliczyć prawdopodobieństwo przez podniesienie sumy amplitud do kwadratu. Jeżeli obie pary są różne (np. $e^+e^- \to \mu^+\mu^-$), to rozpraszania być nie może i opis jest szczególnie prosty. Jak już wiemy, odpowiedni przekrój czynny jest dany przez to samo wyrażenie, co dla procesu rozpraszania z cząstkami zamienionymi na antycząstki (w tym przypadku $e^-\mu^- \to e^-\mu^-$):

przekrój czynny na anihilację z kreacją pary

$$\frac{d\sigma}{dq^2} \sim \frac{\alpha^2_{el}}{q^4}, \quad q^2 > 0.$$

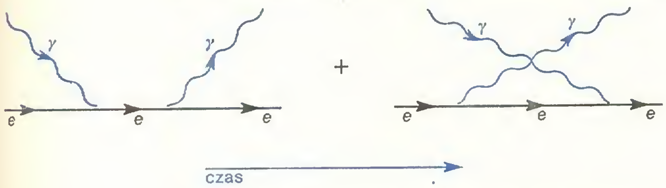


Rys. 10. Wyniki pomiarów przekroju czynnego procesu przejścia pary elektron-pozyton w parę mionów $\mu^-\mu^+$ dla różnych wartości q^2. Krzywa ciągła przedstawia przewidywania teoretyczne

Porównanie tego wyrażenia z danymi doświadczalnymi przedstawiono na rys. 10. Widać, że zgodność jest doskonała.

Rozpraszanie Comptona

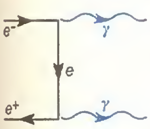
Rozpraszanie Comptona polega na rozpraszaniu fotonów na cząstkach naładowanych. Odpowiednie diagramy podano na rys. 11. Zwr´ćmy uwagę, że teraz



Rys. 11. Diagramy Feynmana opisujące elastyczne rozpraszanie fotonu na elektronie

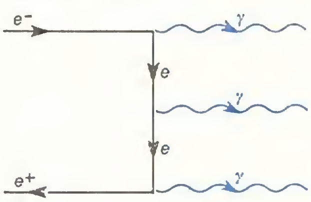
nie foton, a elektron jest wirtualny. Poza tym akty wysyłania i pochłaniania fotonu zachodzą w czasie, którego nieokreśloność jest równa czasowi trwania reakcji. Nie możemy więc stwierdzić, który z tych procesów zachodzi wcześniej i musimy dodać do siebie obie amplitudy odpowiadające diagramom na rysunku. Nie będziemy podawać wzoru na przekrój czynny, zwróćmy tylko uwagę, że jest on znów rzędu α_{el}^2. W wyniku rozpraszania foton zmienia pęd i energię, a stąd i długość fali ($p = h/\lambda$), co nazwano zjawiskiem Comptona. Kinematyka tego zjawiska jest przedstawiona w każdym podręczniku elementarnym. Obliczenie jednak przekroju czynnego, czyli rozkładu kątów fotonów rozproszonych, wymaga posłużenia się aparatem elektrodynamiki kwantowej. Fotony mogą rozpraszać się nie tylko na leptonach, ale również na hadronach, np. jądrach atomowych.

Anihilacja par na fotony rzeczywiste



Rys. 12. Anihilacja pary elektron-pozyton na dwa fotony rzeczywiste

W przeciwieństwie do odpowiedniego procesu z udziałem pojedynczego fotonu, anihilacja pary na dwa i więcej fotonów może zachodzić także dla cząstek rzeczywistych. Proces taki jest obserwowany doświadczalnie. Odpowiedni diagram został przedstawiony na rys. 12. Widzimy, że powstał on z diagramu rozpraszania komptonowskiego przez zamianę rozproszonego elektronu na padający pozyton oraz fotonu padającego na wysłany. Obie reakcje są więc

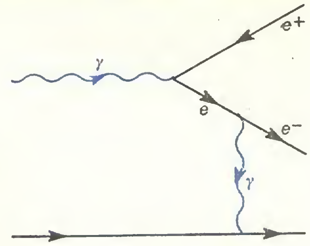


Rys. 13. Anihilacja pary elektron-pozyton na trzy fotony rzeczywiste

opisywane przez tę samą amplitudę, a przekrój czynny jest znów rzędu α_{el}^2. Anihilować mogą nie tylko pary lepton-antylepton ale wszelkie pary cząstka-antycząstka. Na przykład para proton-antyproton może anihilować na fotony, jednak w tym przypadku występuje też bardziej prawdopodobny silny proces anihilacji na mezony π. Prócz anihilacji na parę fotonów obserwuje się też anihilację trójfotonową (rys. 13). Przekrój czynny w tym przypadku jest proporcjonalny do α_{el}^3 i prawdopodobieństwo zajścia reakcji jest ok. sto razy mniejsze.

Kreacja par w polu kulombowskim

W pewnych warunkach może zachodzić również proces kreacji rzeczywistej pary elektron-pozyton przez rzeczywisty, pojedynczy foton. Reakcja taka zdarza się w kulombowskim polu innej cząstki naładowanej, która pełni rolę odbiorcy nieskompensowanego

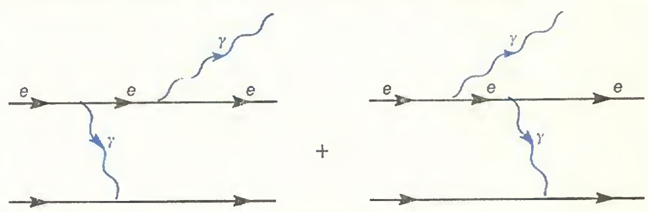


Rys. 14. Kreacja pary elektron-pozyton przez foton rzeczywisty zachodząca poprzez dodatkowe oddziaływanie kulombowskie z inną cząstką, którą może być atomowe jądro

w procesie kreacji pędu. Z rys. 14 widzimy, że zarówno wewnętrzny lepton, jak i foton są teraz cząstkami wirtualnymi. Proces zawiera trzy elementarne akty emisji i jego przekrój czynny jest proporcjonalny do α_{el}^3. Kreacja par jest szczególnie istotna w polu kulombowskim jąder atomowych. Wtedy prawdopodobieństwo zajścia procesu wzrasta Z^2 razy, gdzie Z jest ładunkiem jądra. Fotony przechodzące przez materię dosyć więc często kreują pary. I całe szczęście że często, pojawienie się bowiem pary elektron-pozyton jest jedynym doświadczalnym sygnałem obecności fotonu. Wszystkie stosowane metody detekcji fotonów oparte są na tym zjawisku.

Promieniowanie hamowania

Proces ten zachodzi również jedynie w polu kulombowskim innej cząstki. Promieniowanie hamowania polega na tym, że ruch elektronu zostaje zwolniony w polu np. jądra atomowego w wyniku czego następuje emisja fotonu. Z diagramów przedstawionych na rys. 15 wynika, że proces ten powstaje z procesu kreacji pary (rys. 14) przez odpowiednią zamianę



Rys. 15. Promieniowanie hamowania. Foton rzeczywisty jest wysyłany przez elektron poprzez dodatkowe oddziaływanie kulombowskie z inną cząstką, którą może być jądro atomowe

cząstek. Przekrój czynny jest znów rzędu $Z^2\alpha_{el}^3$. Okazuje się przy tym, że fotony promieniowane są przede wszystkim wewnątrz niewielkich stożków dokoła kierunków lotu elektronu padającego i rozproszonego. Zauważmy, że część amplitudy prawdopodobieństwa związana z wymianą elektronu wirtualnego jest proporcjonalna do $[-q^2+m_0^2]^{-1}$ i staje się bardzo duża, gdy $q^2 \to m_0^2$, czyli gdy elektron wirtualny ma masę niewiele różną od masy elektronu rzeczywistego. Sytuacja taka występuje wtedy, gdy foton wynosi niewiele pędu i energii. W procesie promieniowania hamowania wysyłane są więc przede wszystkim fotony niskoenergetyczne, tzw. miękkie. Często też trudno je wykryć doświadczalnie. Klasycznym odpowiednikiem promieniowania hamowania jest promieniowanie elektromagnetyczne wysyłane przez ładunki elektryczne zmieniające swą prędkość ruchu. Tak promieniują przewodniki z prądem zmiennym, których przykładem jest nadawcza antena radiowa.

Opisane procesy elektromagnetyczne należą do najważniejszych. Prócz nich występują jeszcze znacznie mniej prawdopodobne reakcje z udziałem większej liczby cząstek. Należą tu: promieniowanie hamowania wielu fotonów, produkcja dodatkowych fotonów przy rozpraszaniu Comptona, produkcja pary elektron-

pozyton z wypromieniowanego wirtualnego fotonu hamowania i wiele innych. W reakcjach tych nie występują jednak żadne inne jakościowo zjawiska.

Przejdziemy teraz do opisu oddziaływań elektromagnetycznych hadronów. O niektórych procesach tego typu mówiliśmy już przy dyskusji pewnych reakcji z udziałem leptonów. Teraz skoncentrujemy się na zagadnieniach najważniejszych, na reakcjach będących podstawowym źródłem wiedzy o strukturze hadronów. Zwróćmy przedtem uwagę na dwa ważne fakty. Po pierwsze hadrony oddziałują przede wszystkim silnie i oddziaływania elektromagnetyczne między nimi są przeważnie zupełnie nieistotne. Dlatego przy badaniu oddziaływań elektromagnetycznych hadronów musimy posłużyć się elektronem. Elektron nie bierze udziału w oddziaływaniach silnych i jego rozpraszanie np. na protonie może zajść jedynie poprzez wymianę fotonu wirtualnego (pomijamy oddziaływania słabe elektronu). Foton ten jest pochłaniany przez bogatą strukturę protonu i występowanie tej struktury to drugi istotny dla hadronów fakt. Przedstawiona poprzednio teoria była sformułowana jedynie dla cząstek pozbawionych struktury. Jak zobaczymy dalej teoria ta okaże się płodna również dla hadronów obdarzonych strukturą. Okazuje się bowiem, że hadrony są zbudowane z bardziej elementarnych składników, kwarków, które już nie mają struktury wewnętrznej. Przejdźmy więc do opisu tych doświadczeń z udziałem elektronów i hadronów, których wyniki były decydujące dla podtrzymania hipotezy o istnieniu kwarków.

oddziaływania elektromagnetyczne hadronów

Rozpraszanie elastyczne elektronów na hadronach

Rozpraszanie to jest podobne do rozpraszania kulombowskiego elektronów na mionach $\mu^{\pm}$. Tym jednak razem cząstki rozpraszające mają wyraźną strukturę wewnętrzną. Ładunek elektryczny wewnątrz hadronów jest rozłożony w pewnym obszarze o rozmiarach rzędu 10^{-15} m. W związku z tym wyprowadzony poprzednio wzór na przekrój czynny musi zostać zmodyfikowany o pewną dodatkową zależność od q^2:

$$\frac{d\sigma}{dq^2} \sim \frac{\alpha_{el}^2 F^2(q^2)}{q^4},$$

gdzie funkcja $F(q^2)$ jest tzw. elastycznym czynnikiem postaci hadronu. Jej obecność oznacza, że przy rozpraszaniu efektywny ładunek hadronu wynosi $\pm eF(q^2)$ i przez taki właśnie ładunek jest pochłaniany foton wirtualny o kwadracie masy q^2. Przedyskutujmy własności funkcji $F(q^2)$ z punktu widzenia nieokreśloności Δt czasu emisji i absorpcji fotonu. Czas Δt jest przedziałem czasowym, w którym fakt zajścia emisji czy też absorpcji nie mógł zostać zauważony. Jeżeli czas Δt był bardzo mały, to foton wirtualny wysłany przez elektron przebył niezauważony przez hadron niewielką drogę $c\,\Delta t$, po czym został zaabsorbowany. Sytuacja taka mogła się zdarzyć tylko pod warunkiem, że elektron przebywał w chwili oddziaływania niedaleko środka hadronu, mniej więcej w odległości $c\,\Delta t$. Tak więc foton jest absorbowany tylko przez część ładunku hadronu, co w ujęciu klasycznym odpowiada temu, że ładunek punktowy przelatujący przez obszar naładowany oddziałuje kulombowsko głównie z wewnętrzną częścią ładunku. Oddziaływania warstw zewnętrznych mają tendencję do znoszenia się, a np. dla kuli to znoszenie się jest całkowite. Ponieważ $\Delta t \sim 1/\sqrt{-q^2}$, więc dla $-q^2 \to \infty$ czas $\Delta t \to 0$ i część ładunku hadronu, na której zachodzi rozpraszanie, dąży do zera. Wyciągamy stąd wniosek, że

elastyczny czynnik postaci hadronu

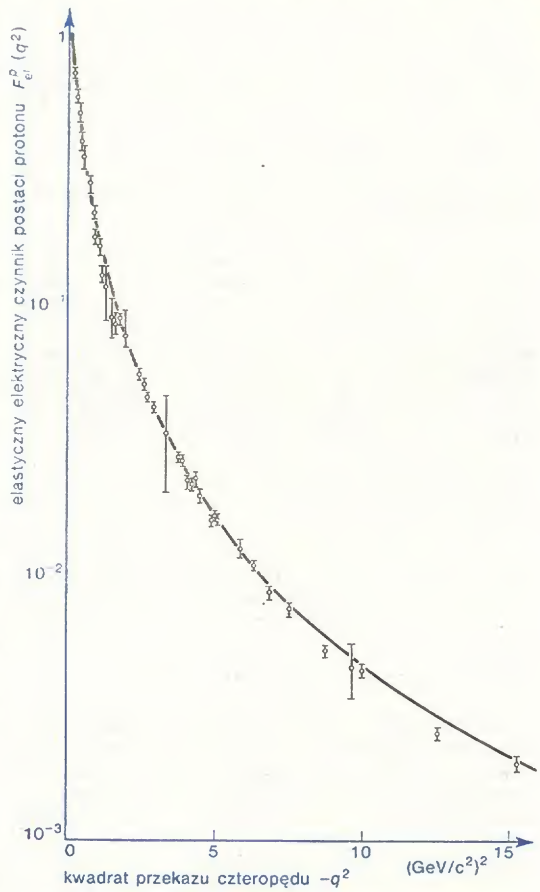
$$\lim_{-q^2 \to \infty} F(q^2) = 0.$$

I na odwrót, dla $\Delta t \to \infty$, czyli $q^2 \to 0$ elektron znajduje się bardzo daleko od środka hadronu (zwróćmy uwagę na fakt, że odległość np. 10^{-7} m może być uważana za niemal nieskończoną w porównaniu ze średnimi rozmiarami protonu, 10^{-15} m) i „widzi" cały jego ładunek. W takich warunkach hadron może być traktowany jako punktowy. Tak więc

$$F(0) = 1 \text{ dla hadronu o ładunku } \pm e,$$
$$F(0) = 0 \text{ dla hadronu o ładunku zero.}$$

Te własności czynnika postaci świetnie zgadzają się z doświadczeniem. Na rys. 16 przedstawiony jest

czynnik postaci protonu



Rys. 16. Wyniki pomiaru elastycznego elektrycznego czynnika postaci protonu dla różnych wartości q^2

zmierzony w doświadczeniu elastyczny czynnik postaci protonu. Jest to tzw. czynnik postaci elektryczny. Proton, jak i wszystkie hadrony obdarzone spinem, ma bowiem również inny czynnik postaci — magnetyczny — o podobnych własnościach. Występowanie dodatkowych czynników postaci dla hadronów wiąże się z tym, że dla cząstek mających wewnętrzny moment pędu (spin) ładunki wewnętrzne składające się na strukturę cząstki wykonują ruchy, które są przyczyną powstawania prądów elektrycznych. Prądy te biorą udział w oddziaływaniach magnetycznych z poruszającym się elektronem i rozkład tych właśnie oddziaływań wewnątrz hadronu odzwierciedlają magnetyczne czynniki postaci. Hadrony obdarzone spinem są więc małymi magnesami o pewnym momencie magnetycznym, którego odpowiednią część (dążącą oczywiście znów do zera, gdy $-q^2 \to \infty$) obserwuje w rozpraszaniu elektron. Zwróćmy przy okazji uwagę na fakt, że moment magnetyczny mają również leptony (o spinie $^1/_2$), ale tylko naładowane. W wypadku leptonów moment magnetyczny nie wykazuje jednak struktury i odpowiedni czynnik postaci, jak i wszystkie czynniki postaci leptonów, wynosi jeden (albo zero dla neutrin) przy wszystkich wartościach q^2.

elektryczny i magnetyczny czynnik postaci

Rozpraszanie nieelastyczne elektronów na nukleonach

Pomiar elastycznych czynników postaci nukleonów stał się najważniejszym argumentem na rzecz istnienia struktury wewnętrznej hadronów. Struktura ta powstaje poprzez oddziaływania silne, w których uczestniczą hadrony i szczegółowa jej analiza, np. obliczenie funkcji $F(q^2)$, jest bardzo złożona. Niewiele możemy powiedzieć o oddziaływaniach, dla opisu których nie możemy stosować rachunku zaburzeń i diagramów Feynmana. Posługujemy się wtedy opisem półfenomenologicznym, wprowadzając nielokalne fenomenologiczne oddziaływanie poprzez wymianę np. wirtualnych mezonów π obdarzonych strukturą, czy też biegunów Reggego (→ Oddziaływania silne). Metody te, chociaż zaczerpnięte z lokalnej kwantowej teorii pola, nie noszą już tych cech ścisłości i jednoznaczności przewidywań, które charakteryzują np. elektrodynamikę kwantową. Stąd elastyczne czynniki postaci hadronów możemy obliczyć posługując się jedynie pewnymi modelami. Lepiej przedstawia się sprawa z tzw. nieelastycznymi czynnikami postaci, których pomiar udowodnił, że struktura hadronów jest zbudowana z cząstek nie mających już struktury, a mianowicie z kwarków.

nieelastyczne czynniki postaci

Nieelastyczne czynniki postaci wyznacza się badając nieelastyczne rozpraszanie elektronów na hadronach (nukleonach). Reakcja ta polega na rozbiciu struktury hadronu przez padający elektron. Wysłany z elektronu wirtualny foton zostaje pochłonięty przez hadron, w wyniku czego powstaje wiele dodatkowych hadronów, głównie mezonów (rys. 17). Wyobraźmy sobie teraz, że np. proton składa się z pewnej liczby składników obdarzonych ładunkami elektrycznymi Q_i i że przez te właśnie składniki jest pochłaniany foton. Reakcja ma więc przebieg, jak na rys. 18. Rysunek ten jest jednak bardzo wyidealizowany. Czas oddziaływania silnego między składnikami protonu jest bowiem zupełnie dowolny i w trakcie pochłaniania fotonu składniki mogą wielokrotnie oddziaływać między sobą zamazując całkowicie obserwowane zjawisko. Jeżeli jednak elektrony mają bardzo dużą energię, to z punktu widzenia elektronu proton porusza się z bardzo dużą prędkością, bliską prędkości światła. Zaczynają wtedy odgrywać rolę efekty relatywistyczne; między innymi wszelkie ruchy i zmiany wewnątrz protonu będą wydawały się naszemu elektronowi bardzo powolne — czas ulega dylatacji. Jeżeli dodatkowo będziemy interesowali się jedynie rozproszeniami, w których q^2 jest duże, to czas trwania procesu pochłaniania fotonu będzie krótki. W takich właśnie kontrolowanych doświadczalnie warunkach, w czasie pochłaniania fotonu przez składnik protonu nie zachodzą prawie żadne oddziaływania silne między składnikami i reakcja przebiega w idealny sposób przedstawiony poprzednio na rysunku. Oczywiście amplituda całego rozpraszania jest sumą amplitud prawdopodobieństwa zajścia rozpraszania na każdym ze składników. Korzystając ze związku de Broglie'a między długością fali i pędem ($\lambda = h/p$) można się przekonać, że dla dużych q^2 (a więc i dużych pędów fotonu) warunek spójności między poszczególnymi amplitudami zostaje silnie naruszony. Nie ma więc interferencji i całkowity przekrój czynny jest po prostu sumą przekrojów czynnych dla poszczególnych rozproszeń:

duże energie, duże q^2

$$\frac{d\sigma}{dq^2} = \sum_i \frac{d\sigma_i}{dq^2}.$$

Znamy wzór na każdy ze składników tej sumy:

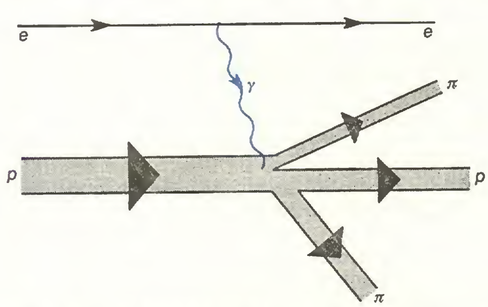
$$\frac{d\sigma_i}{dq^2} \sim \frac{\alpha_{el}^2 Q_i^2}{q^4} F_i^2(q^2),$$

gdzie Q_i — ładunek (w jednostkach ładunku elementarnego), $F_i(q^2)$ — elastyczny teraz czynnik postaci i-tego składnika protonu. Tak więc

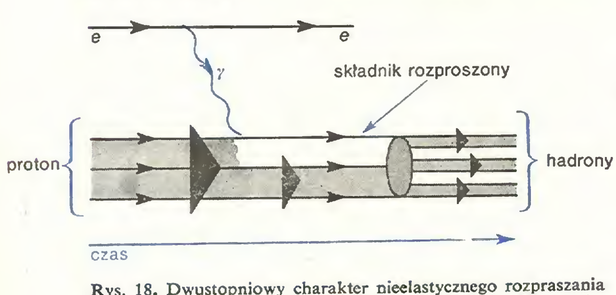
$$\frac{d\sigma}{dq^2} \sim \frac{\alpha_{el}^2}{q^4} \sum_i Q_i^2 F_i^2(q^2).$$

Zatem badając zależność całkowitego przekroju czynnego od q^2 możemy dowiedzieć się czegoś o strukturze samych składników. Wyniki takiego pomiaru zostały przedstawione na rys. 19. Widać wyraźnie, że w opisanych wyżej warunkach (duże energie, duże q^2) wyrażenie $\frac{d\sigma}{dq^2} \Big/ \frac{\alpha_{el}^2}{q^4}$ nie zależy od q^2. Elastyczne czynniki postaci składników protonu (a także neutronu, bo i dla niego przeprowadzono pomiary) nie zależą więc od q^2. Oznacza to, jak wiemy, że $F_i(q^2) \equiv 1$. Składniki nukleonów nie mają więc struktury.

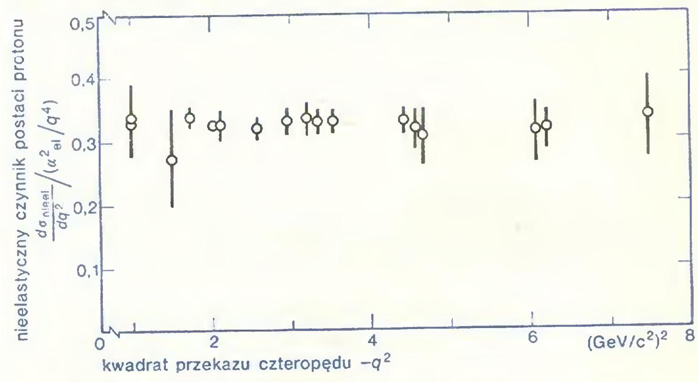
Mogłoby się wydawać, że w tym samym doświadczeniu mierzymy również sumę kwadratów ładunków składników, $\sum_i Q_i^2$, gdyż wszystkie $F_i^2(q^2) \equiv 1$. Tak jednak nie jest, bowiem w powyższym rozumowaniu zrobiliśmy pewne istotne uproszczenie. Nie uwzględniliśmy wcale faktu, że składniki mogą mieć wewnątrz protonu pewne niezerowe wartości pędu, podobnie



Rys. 17. Rozbicie protonu zachodzące w wyniku pochłonięcia fotonu wirtualnego



Rys. 18. Dwustopniowy charakter nieelastycznego rozpraszania elektronu na protonie. Najpierw foton wirtualny jest pochłaniany przez jeden ze składników protonu, a potem zachodzi oddziaływanie wszystkich składników prowadzące do powstania pewnej liczby hadronów



Rys. 19. Wyniki pomiaru przekroju czynnego reakcji nieelastycznego rozpraszania elektronów na protonach dla różnych wartości q^2. Przedstawione dane odpowiadają różnym energiom elektronów większym od wartości 5 GeV

jak elektron wewnątrz atomu wodoru. Od rozkładu tych pędów zależy również całkowity przekrój czynny. Oznaczając przez $f_i(p_z)$ prawdopodobieństwo tego, że wewnątrz protonu i-ty składnik naładowany (ten, z którym oddziałuje elektron) ma pęd p_z (przy bardzo dużych prędkościach składnika względem elektronu istotna jest tylko jedna składowa pędu — w kierunku ruchu elektronu — oznaczyliśmy ją przez p_z), otrzymujemy poprawny już wzór na przekrój czynny:

$$\frac{d\sigma(q^2, p_z)}{dq^2} \sim \frac{\alpha_{el}^2}{q^4} \sum_i Q_i^2 f_i(p_z),$$

gdzie uwzględniliśmy doświadczalnie odkryty fakt braku struktury składników, $F_i(q^2) \equiv 1$. Okazuje się, że nie tylko wartości energii i q^2 mogą być kontrolowane w doświadczeniu. Mierząc całkowitą energię (masę) hadronów powstałych z rozbicia protonu możemy wyznaczyć również wartość p_z. Nie będziemy tego bardzo ważnego faktu bliżej dowodzić. W każdym razie możemy zsumować wyznaczone doświadczalnie przekroje czynne odpowiadające reakcjom ze wszystkimi możliwymi wartościami p_z. Przyjmując, że rozkład kwarków $u(d)$ w protonie jest taki sam, jak kwarków $d(u)$ w neutronie (→ Struktura cząstek elementarnych), można obliczyć średni przekrój czynny rozpraszania elektronów na protonach i neutronach — otrzymuje się wówczas wyrażenie, w którym występuje pewna średnia, niezależna od rodzaju składnika, funkcja rozkładu $f(p_z)$. Oznaczając

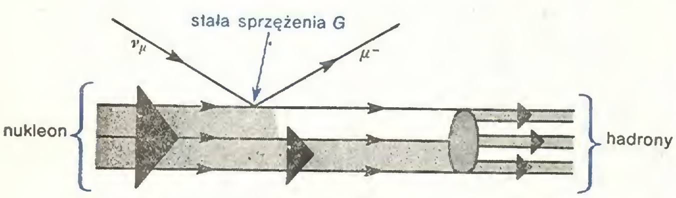
$$f = \sum_{p_z} f(p_z)$$

otrzymamy wtedy

$$\frac{d\sigma_{eN}(q^2)}{dq^2} \sim \frac{\alpha_{el}^2}{q^4} f \frac{1}{2} \sum_i Q_i^2,$$

gdzie stała f oznacza prawdopodobieństwo tego, że całkowity pęd protonu (z punktu widzenia elektronu) jest w całości niesiony przez jego naładowane składniki.

Suma kwadratów ładunków składników nukleonów może zostać wyznaczona, jeżeli zrobimy założenie, że w oddziaływaniach słabych hadronów biorą udział dokładnie te same składniki, co w oddziaływaniach elektromagnetycznych. Możemy wtedy posłużyć się procesem nieelastycznego rozpraszania neutrin na jądrach atomowych, które są na tyle duże, że odpowiednie przekroje czynne mogą być zmierzone. Reakcję $\nu_\mu + N \to \mu^- +$ hadrony badamy znów przy dużych energiach neutrin i dla dużego q^2 (rys. 20). Ko-



Rys. 20. Dwustopniowy charakter nieelastycznego rozpraszania neutrin mionowych na nukleonach. Najpierw zachodzi słaby proces rozpraszania neutrina na jednym z kwarków (zasięg tego oddziaływania jest niemal równy zeru), a potem wszystkie kwarki oddziałują silnie w wyniku czego powstaje pewna liczba hadronów

rzystając z faktu, że stała sprzężenia oddziaływań słabych G (→ Oddziaływania słabe, badany obszar q^2 jest stale taki, że $|q^2| \ll m_W^2$, gdzie m_W jest masą bozonu pośredniczącego w oddziaływaniach słabych) jest uniwersalna dla wszystkich cząstek biorących udział w tych oddziaływaniach, możemy wyprowadzić odpowiedni wzór na przekrój czynny:

$$\frac{d\sigma_{\nu N}}{dq^2} \sim G^2 f.$$

Znając stałą G możemy z obu pomiarów wyeliminować nieznaną poprzednio stałą f. Okazuje się, że $f \neq 1$, co świadczy o tym, że w nukleonach znajdują się nie tylko składniki oddziałujące elektromagnetycznie i słabo. Możemy w ten sposób wyznaczyć jedynie sumę kwadratów ładunków uśrednioną po protonach i neutronach. Otrzymamy wynik

$$\frac{1}{2} \sum_i Q_i^2 = 0{,}27 \pm 0{,}03$$

doskonale zgadza się z hipotezą kwarków. Jeżeli bowiem właśnie kwarki są składnikami hadronów biorącymi udział w oddziaływaniach elektromagnetycznych i słabych, to w protonie znajdują się dwa kwarki o ładunku $^2/_3$ i jeden o ładunku $-^1/_3$ (→ Struktura cząstek elementarnych), a w neutronie na odwrót. Mamy zatem

$$\frac{1}{2}\left[\left(\frac{2}{3}\right)^2 + \left(-\frac{1}{3}\right)^2\right] = \frac{5}{18} \approx 0{,}27!$$

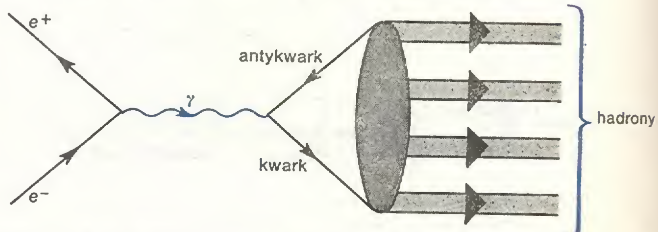
dowód istnienia kwarków

W ten sposób wyniki eksperymentów nieelastycznego rozpraszania elektronów (i neutrin) na nukleonach dowiodły, że składnikami hadronów są kwarki pozbawione struktury wewnętrznej.

Można zadać pytanie, skąd w takim składnikowym modelu bierze się obserwowana silna zależność od q^2 elastycznego czynnika postaci samych hadronów. Pochodzi ona stąd, że przy dużych q^2 kwark, który pochłonął foton, nabiera dużego pędu względem pozostałych kwarków i zostaje silnie wybity z całej struktury. Niechętnie więc zlepia się z resztą kwarków w wyjściowy hadron. Im większe q^2, tym mniejsze prawdopodobieństwo utworzenia na końcu stanu pojedynczego hadronu. Dlatego właśnie prawdopodobieństwo (przekrój czynny) zajścia reakcji elastycznej silnie maleje ze wzrostem q^2. Dlatego też w opisanych wyżej doświadczeniach badano nie tylko reakcje nieelastyczne, ale dodatkowo nie nakładano żadnych ograniczeń doświadczalnych na rodzaj nieelastyczności. Badano więc np. rozpraszanie elektronów na protonach nie interesując się zupełnie, ile i jakie hadrony powstały w procesie. Wtedy i tylko wtedy eliminujemy z opisu wpływ oddziaływania kwarków po rozproszeniu, które to oddziaływanie prowadzi do powstania końcowego stanu wielohadronowego.

Anihilacja pary elektron-pozyton na hadrony

Reakcja $e^+e^- \to$ hadrony zachodzi, podobnie jak opisana poprzednio reakcja $e^+e^- \to \mu^+\mu^-$, przez anihilację pary na foton z $q^2 > 0$ i następnie kreację z tego fotonu różnych hadronów, głównie mezonów. Dla dużych wartości q^2 (w tej reakcji q^2 w układzie środka masy e^+e^- równa się kwadratowi całkowitej energii pary) czas oddziaływania jest znów niewielki i kreacja hadronów zachodzi zapewne przez kreację wszystkich możliwych par składników (kwarków) i dopiero potem oddziaływanie tych par prowadzące do powstania różnych stanów wielohadronowych. W tym wypadku jednak argumenty nie są tak ścisłe jak w wypadku reakcji nieelastycznego rozpraszania elektronów, gdyż nie mamy żadnej metody na kontrolowanie doświadczalne czasu trwania oddziaływań



Rys. 21. Proces anihilacji pary elektron-pozyton ma przy odpowiednio wysokich energiach pary przebieg w pewnym przybliżeniu dwustopniowy. Najpierw foton wirtualny kreuje lokalnie jedną z par kwarków bez struktury, a potem para ta oddziałuje silnie przechodząc w pewną liczbę hadronów

wewnątrzhadronowych i czas ten w zasadzie jest zupełnie dowolny. Jeżeli jednak wyprodukowana para składników ma dużą energię, to rozlatują się one bardzo szybko i prawdopodobnie oddziaływanie przebiega przez dwa kolejne stadia, jak na rys. 21. Jeżeli więc nie interesujemy się, jakie i ile hadronów powstało z wyprodukowanych par, oraz jeżeli $\sqrt{q^2}$ jest dużo większe od wszystkich dostępnych energetycznie progów produkcji tych par, to całkowity przekrój czynny dany jest wzorem analogicznym do wzoru na przekrój czynny reakcji $e^+e^- \to \mu^+\mu^-$:

$$\frac{d\sigma}{dq^2}(e^+e^- \to \text{hadrony}) \sim \frac{\alpha_{\text{el}}^2}{q^4}\sum_i Q_i^2,$$

gdzie sumowanie przebiega po ładunkach wszystkich możliwych par składników hadronów. Przy pisaniu tego wzoru zrobiliśmy założenie, że składniki hadronów nie mają struktury, $F_i(q^2) \equiv 1$, co tutaj znów zostało potwierdzone doświadczalnie w bardzo dużym przedziale wartości q^2. Odpowiedni przekrój czynny został zmierzony przy użyciu akceleratorów ze zderzającymi się przeciwbieżnymi wiązkami elektronów i pozytonów. Dzieląc ten przekrój czynny przez odpowiedni przekrój czynny na produkcję pary $\mu^+\mu^-(\alpha_{\text{el}}^2/q^4)$ otrzymujemy tzw. stosunek R, którego

stosunek R

chowanie się stosunku R, a mianowicie jego wartość $R \approx 2$ poniżej progu produkcji kwarków powabnych i $R \approx 4$ powyżej tego progu jest zgodne z przewidywaniami modelu kwarków. Należy tylko pamiętać, że każdy rodzaj kwarku ma trzy odmiany (tzw. kolory), które nie różnią się między sobą masami ani ładunkami elektrycznymi. W hadronach znajduje się zawsze określona kombinacja tych odmian i np. proton składa się zawsze z trzech kwarków: dwóch kwarków jednego rodzaju (ładunek $^2/_3$) i jednego kwarku innego rodzaju (ładunek $-^1/_3$), z których każdy jest innej odmiany (koloru). Całkowita liczba wszystkich możliwych kwarków zwiększa się w ten sposób trzykrotnie. Poniżej progu produkcji kwarków powabnych mamy więc

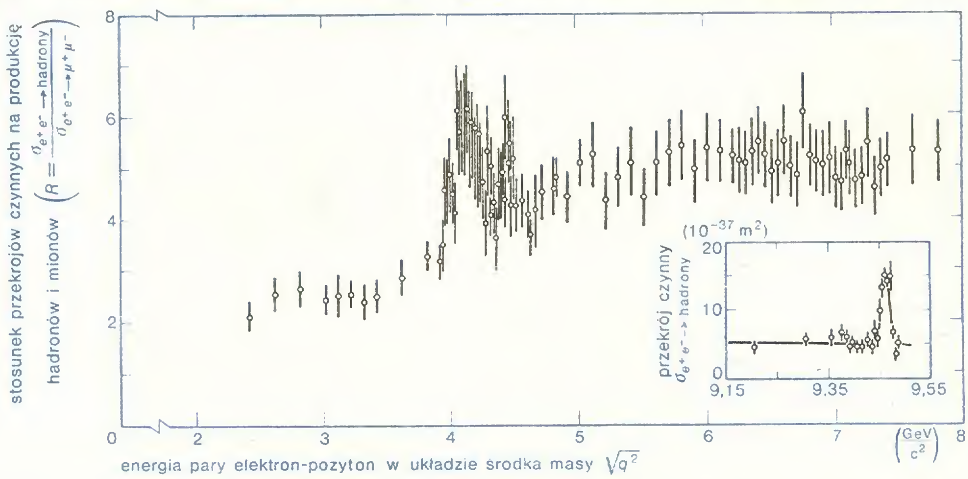
$$R = 3\left[\left(\frac{2}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2\right] = 2,$$

zaś powyżej progu

$$R = 3\left[\left(\frac{2}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 3\frac{1}{3}$$

w dość dobrej zgodności z danymi doświadczalnymi.

Warto jeszcze zwrócić uwagę na ostre maksima pojawiające się w stosunku R w obszarach nieregu-



Rys. 22. Wyniki pomiaru stosunku R całkowitego przekroju czynnego anihilacji par elektron-pozyton na hadrony do przekroju czynnego przejścia tych par na pary mionów $\mu^-\mu^+$ dla różnych wartości q^2. W rogu zamieszczono dane z obszaru powyżej 9 GeV/c^2 nie unormowane do przekroju czynnego na produkcję mionów

wartości doświadczalne przedstawiono na rys. 22. Teoretycznie (choć jedynie w przybliżeniu) stosunek R jest oczywiście dany przez sumę kwadratów ładunków wszystkich możliwych składników wszelkich hadronów:

$$R = \sum_i Q_i^2.$$

Z rysunku widać, że po początkowych nieregularnościach występujących w obszarze wartości q^2 bliskich progom produkcji trzech najlżejszych par kwark-antykwark o ładunku $^2/_3$, $-^1/_3$ i $-^1/_3$ (→ Struktura cząstek elementarnych) stosunek R ustala się — przyjmuje wartość $R \approx 2$ — po czym znowu przy progu produkcji pary kwarków powabnych o ładunku $^2/_3$ występują nieregularności. Dla jeszcze większych wartości q^2 stosunek R ustala się na poziomie $R \approx 5$, po czym jeszcze raz, przy kolejnym progu produkcji jeszcze jednej, piątej pary kwarków, pojawiają się nieregularności. W okolicach progu produkcji kwarków powabnych pojawia się również kreacja pary ciężkich leptonów τ (→ Oddziaływania słabe), które nie zostały wydzielone z danych doświadczalnych. Powyżej tego progu proces produkcji ciężkich leptonów (o ładunku jeden) daje do stosunku R wkład równy jedności i jedynie $R \approx 4$ powinno pochodzić z wkładu od kreacji par kwarkowych. Powyższe za-

larności. Są one wynikiem produkcji krótkożyjących stanów hadronowych, tzw. rezonansów, które szybko rozpadają się na hadrony.

Inne oddziaływania hadronów z fotonami

Przedstawimy teraz krótko niektóre inne, mniej ważne reakcje elektromagnetyczne hadronów. Ich analiza jest oparta na znajomości kwarkowego składu hadronów. Dla dalszych rozważań będzie tylko istotna informacja, że wszystkie bariony składają się z trzech kwarków, natomiast mezony z pary kwark-antykwark. Wysłanie fotonu wirtualnego z kwarku nie prowadzi do jakiejkolwiek jego zmiany, podobnie jak dla wszystkich cząstek pozbawionych struktury. Stąd w oddziaływaniach elektromagnetycznych hadronów jest zachowana trzecia składowa izospinu I_3, hiperładunek Y (lub dziwność S) oraz powab C. Wartości bowiem tych liczb kwantowych służą w modelu kwarków jedynie do ponumerowania różnych rodzajów kwarków. Również odmiana kwarku (kolor) nie zmienia się przy wysyłaniu fotonu. Wszystkie te prawa zachowania zostały sprawdzone w wielu doświadczeniach. Nie ma jednak żadnego powodu, by w oddziaływaniach elektromagnetycznych zachowy-

liczby kwantowe zachowywane

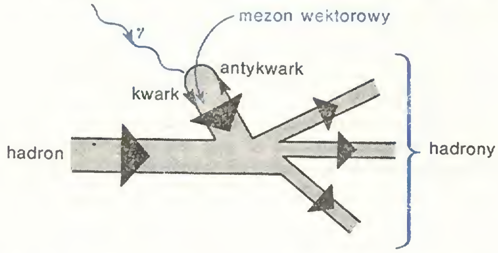
wał się izospin całkowity hadronów. I rzeczywiście nie jest.

rozpraszanie fotonów na hadronach

Oddziaływania elektromagnetyczne hadronów dzielą się na reakcje rozpraszania fotonów oraz rozpady cząstek. Typowymi reakcjami rozpraszania są procesy fotoprodukcji (np. $\gamma+p \to \pi+p$) oraz rozpraszania komptonowskiego (np. $\gamma+p \to \gamma+p$). Końcowe stany hadronowe mogą być i często są wielocząstkowe (np. $\gamma+p \to 2\pi+p$). Procesy rozpraszania fotonów zachodzą przez kreację (lub anihilację) par kwark-antykwark przez foton rzeczywisty (rys. 23). Pary te uczestniczą dalej w oddziaływaniach silnych z hadronem bombardowanym. Powstała z fotonu (lub przechodząca w foton) wirtualna para kwark-antykwark ma liczby kwantowe mezonu, i to mezonu o spinie jeden — równym spinowi fotonu, gdyż obowiązuje zasada zachowania momentu pędu. Stąd np. reakcja fotoprodukcji (rys. 24) zachodzi w ten



Rys. 23. Anihilacja wirtualnej pary kwark-antykwark na foton rzeczywisty



Rys. 24. Proces fotoprodukcji hadronów na hadronach zachodzi za pomocą pośrednictwa pewnego wirtualnego mezonu wektorowego, w który przechodzi padający foton rzeczywisty

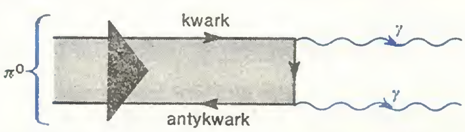
sposób, że foton przechodzi w jeden z mezonów wektorowych o spinie 1 ($\varrho, \omega, \varphi, \ldots$), po czym mezony te rozpraszają się na hadronie bombardowanym. Takie

model dominacji wektorowej

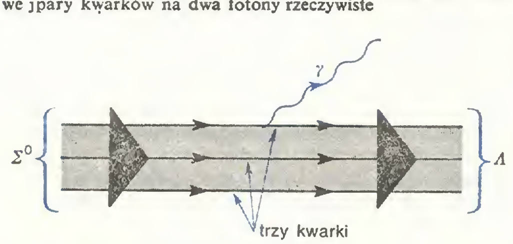
pośredniczące działanie mezonów wektorowych nazwano modelem dominacji wektorowej. Model ten może być również stosowany w przypadku elastycznego pochłaniania fotonu przez składniki hadronu (dostarczając dość dobrej parametryzacji elastycznego czynnika postaci). Oczywiście proces rozpraszania mezonów wektorowych na hadronach bombardowanych należy do zjawisk wywoływanych oddziaływaniami silnymi i jako taki może być opisany jedynie półfenomenologicznie (→ Oddziaływania silne).

rozpady elektromagnetyczne hadronów

Typowymi rozpadami elektromagnetycznymi hadronów są rozpady mezonu $\pi^0 (\pi^0 \to 2\gamma)$ oraz hiperonu $\Sigma^0 (\Sigma^0 \to \Lambda+\gamma)$. Pierwszy z nich to anihilacja mezonowej pary kwark-antykwark na dwa fotony, a drugi to promieniowanie hamowania kwarku hiperonowego zachodzące w polu oddziaływania z innymi kwarkami (hiperony Σ^0 i Λ mają ten sam skład kwarkowy). Oba te procesy zostały przedstawione na



Rys. 25. Rozpad mezonu π^0 zachodzi poprzez anihilację mezonowej pary kwarków na dwa fotony rzeczywiste

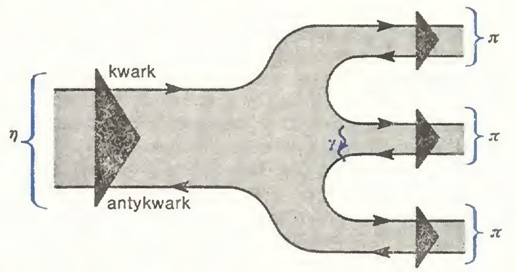


Rys. 26. Rozpad hiperonu Σ^0 zachodzi poprzez promieniowanie hamowania jednego z kwarków znajdującego się w polu oddziaływań silnych kwarków pozostałych. Rozpad ten jest możliwy tylko dlatego, że hiperony Σ^0 i Λ składają się z takich samych trzech kwarków, choć stany kwantowe kwarków są różne w obu przypadkach

rys. 25 i 26. Jak mówiliśmy, prawdopodobieństwo zajścia pierwszej z tych reakcji jest rzędu α_{el}^2, zaś drugiej rzędu α_{el}. W tym drugim jednak przypadku, podobnie jak dla rozpadu β neutronu (→ Oddziaływania słabe), prawdopodobieństwo rozpadu jest silnie stłumione, gdyż różnica mas hiperonów Σ^0 i Λ jest niewielka. W rezultacie czasy życia zarówno mezonu π^0, jak i hiperonu Σ^0 są zbliżone i wynoszą ok. 10^{-16}–10^{-17} s. Zauważmy, że czas życia cząstki jest odwrotnie proporcjonalny do prawdopodobieństwa jej rozpadu. Im więc większa stała sprzężenia oddziaływania wywołującego rozpad, tym krótszy czas życia.

czas życia a stała sprzężenia

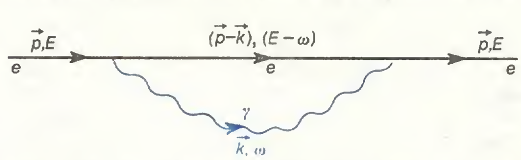
Można by spytać, dlaczego cząstki π^0 i Σ^0 nie rozpadają się przede wszystkim za pomocą oddziaływań silnych, skoro odpowiednie prawdopodobieństwo rozpadu powinno być większe, a czas życia krótszy. Brak rozpadów silnych w tym wypadku bierze się stąd, że nie istnieją odpowiednie cząstki lżejsze od π^0 i Σ^0, na które oba te hadrony mogłyby się rozpadać bez pogwałcenia wszystkich obowiązujących w oddziaływaniach silnych praw zachowania (I, I_3, Y itd.). W przypadku innych hadronów, np. rezonansów, obserwujemy także rozpady elektromagnetyczne (np. $\varrho \to \pi+\gamma$), ale mają one jedynie znaczenie marginesowe i nie decydują o czasie życia. Natomiast wyłącznie elektromagnetycznego pochodzenia są rozpady mezonu η zachodzące czasem z udziałem jedynie fotonów wirtualnych (np. $\eta \to 3\pi$, rys. 27).



Rys. 27. Elektromagnetyczny rozpad mezonu η. Rozpad silny jest wzbroniony przez zasadę zachowania całkowitego izospinu. Wymiana fotonu wirtualnego między pewną parą kwarków zmienia izospin całkowity umożliwiając zachodzenie reakcji

Przedstawiliśmy już wszystkie zasadnicze własności oddziaływań elektromagnetycznych cząstek elementarnych, a także schemat teoretyczny pozwalający w sposób jednoznaczny i ścisły analizować, opisywać i przewidywać te własności. Elektrodynamika kwantowa jest teorią tak elegancką, a jej przewidywania tak świetnie zgadzają się z doświadczeniem (np. pomiar momentu magnetycznego elektronu daje wynik $1{,}0011596577 \pm 0{,}0000000035 e\hbar/2m_e$, a wartość teoretyczna wynosi $1{,}0011596554 \pm 0{,}0000000033 \times e\hbar/2m_e$; podobna zgodność zachodzi dla mionu μ^-), że na wzór tej teorii buduje się ostatnio modele wszystkich rodzajów oddziaływań. W ten sposób powstała teoria kwarkowo-gluonowa oddziaływań silnych (→ Struktura cząstek elementarnych) oraz jednolita teoria oddziaływań elektromagnetycznych i słabych (→ Oddziaływania słabe). Należy jednak zaznaczyć, że współczesna struktura elektrodynamiki kwantowej nie może być uważana za ostateczną. Teoria ta bowiem zawiera pewne luki formalne. Jedną z nich jest zasadnicza niemożność obliczenia mas podstawowych cząstek, np. leptonów. Wiąże się to z faktem, że np. elektron w swym ruchu swobodnym

moment magnetyczny elektronu



Rys. 28. Reabsorpcja fotonu wirtualnego przez elektron. Pęd $\vec{k}$ i energia ω fotonu są dowolne i należy wykonać sumowanie (całkowanie) po wszystkich wartościach tych zmiennych

może wysyłać i reabsorbować fotony wirtualne (rys. 28). Odpowiedni diagram Feynmana wydaje się mały (rzędu α_{el}), ale obliczenie jego wkładu wymaga wykonania sumowania (całkowania) po wszystkich podziałach pędu padającego elektronu na pędy pośredniego elektronu wirtualnego i również wirtualnego fotonu. Suma okazuje się nieskończona (rozbieżna). Logiczną zwartość teorii można wprawdzie uratować za pomocą pewnego dobrze określonego przepisu (zwanego procedurą renormalizacji), ale pojawiają się wtedy dodatkowe dowolne parametry. W ten sposób np. masa elektronu okazuje się dowolna. Choć więc wydawałoby się, że istnieje bardzo dobra teoria, wciąż pewne ciemne jej obszary nie pozwalają fizykom na bezczynność.

G. Białkowski, R. Sosnowski *Cząstki elementarne*, Warszawa 1971; W.B. Bierestecki, E.M. Lifszic, L.P. Pitajewski *Relatywistyczna teoria kwantów*, cz. I, Warszawa 1972; L.N. Cooper *Istota i struktura fizyki*, Warszawa 1975.

Oddziaływania słabe

Andrzej Szymacha

Oddziaływania słabe na tle innych oddziaływań

Oddziaływania słabe są jednym z trzech typów oddziaływań mających znaczenie w fizyce cząstek elementarnych (→ Cząstki elementarne i ich oddziaływania). W zbadanym dotychczas zakresie energii cząstek są one najsłabsze. Silniejsze od nich są oddziaływania elektromagnetyczne oraz jądrowe zwane też oddziaływaniami silnymi.

oddziaływanie w teorii klasycznej i kwantowej

W mechanice klasycznej termin „oddziaływanie" był synonimem siły działającej między rozpatrywanymi ciałami. Ujęcie klasyczne zawodzi jednak w zastosowaniu do mikroobiektów, szczególnie przy opisie zjawisk zachodzących z udziałem niewielkiej liczby cząstek elementarnych. W tym wypadku poprawny opis teoretyczny musi być sformułowany za pomocą pojęć stosowanych w teorii kwantowej i to w dodatku w najbardziej zaawansowanej jej wersji — kwantowej teorii pola. W teorii tej potocznemu pojęciu oddziaływania odpowiada hamiltonian oddziaływania, którego znajomość pozwala (przynajmniej w zasadzie — praktycznie jest to często bardzo trudne) obliczać (przepowiadać) wyniki najrozmaitszych doświadczeń przeprowadzanych z udziałem cząstek elementarnych. Kwantowa teoria pola nie daje się jednak wyłożyć w sposób krótki i prosty. Z drugiej zaś strony oddziaływania słabe — w przeciwieństwie do elektromagnetycznych — są niezwykle krótkozasięgowe i nie mają żadnego odpowiednika w makroświecie. Dlatego opisując je, będziemy przedstawiać konkretne skutki tego oddziaływania, które są znacznie prostsze i pojęciowo uchwytniejsze od abstrakcyjnego hamiltonianu oddziaływania.

skutki oddziaływań

W kręgu zjawisk z udziałem niewielkiej liczby cząstek elementarnych wszelkie oddziaływania mogą się przejawiać w postaci następujących skutków:

a) Wpływu na statyczne parametry cząstek elementarnych i na powstawanie układów związanych. Należą tu takie własności jak: masa cząstek elementarnych, moment magnetyczny, energia wiązania jądra itp.

b) Rozpadów cząstek nietrwałych i układów nietrwałych charakteryzowanych liczbowo przez czas rozpadu (dokładniej średni czas rozpadu bądź czas połowicznego zaniku) i prawdopodobieństwa rozpadów danego obiektu na różne możliwe produkty końcowe.

c) Zachodzenia procesów rozpraszania elastycznego lub nieelastycznego. Pierwszym doświadczeniem tego rodzaju było doświadczenie E. Rutherforda uwieńczone odkryciem jądra atomowego. Wszystkie akceleratory buduje się właśnie po to, by zderzać jedne cząstki z drugimi. Poza tym, zderzenie wysokoenergetycznych cząstek jest jedynym sposobem wytwarzania wszelkich nietrwałych cząstek, których rozpady następnie się bada.

Rola oddziaływań słabych nie jest jednakowo ważna we wszystkich wyżej wymienionych punktach. Są one za słabe, by mogły prowadzić do utworzenia układu związanego. Wszystkie znane układy złożone są utrzymywane bądź przez oddziaływania silne (jądra atomowe), bądź przez elektromagnetyczne (molekuły, atomy, atomy egzotyczne). Wpływ oddziaływań słabych na własności układów złożonych (których istnienie jest uwarunkowane silniejszymi oddziaływaniami) jest znikomo mały. Jedynie w nielicznych przypadkach daje się go zaobserwować przez efekty niezachowania parzystości w stanach wzbudzonych jąder. Interpretacja tych efektów zależy jednak silnie od bardzo złożonych efektów jądrowych i dlatego w ten sposób nie udało się uzyskać niemal żadnych informacji o oddziaływaniach słabych. W ostatnim czasie wielkie zainteresowanie wzbudzają doświadczenia, w których podobnego efektu poszukuje się w zjawisku przejść optycznych w gazach jednoatomowych (wodór atomowy, pary metali ciężkich). Praktycznie efekt taki powinien przejawiać się jako aktywność optyczna niektórych gazów jednoatomowych bądź jako szczątkowa polaryzacja kołowa światła wysyłanego przez rozgrzane gazy. Żaden z tych efektów nie wystąpiłby, gdyby między elektronami i jądrem atomowym działały tylko zwykłe siły elektromagnetyczne. W związku z postępem techniki laserowej i elektroniki osiągnięto już dokładność rzędu spodziewanego efektu. W niedalekiej przyszłości będzie to zapewne bardzo cenna metoda badania oddziaływań słabych.

rozpraszanie neutrin

Procesy rozproszeniowe wywołane oddziaływaniami słabymi są również trudne do obserwacji. Wyjątkiem są procesy rozpraszania neutrin. Są one wprawdzie niesłychanie mało prawdopodobne, wymagające zatem olbrzymich detektorów, dużych natężeń wiązki neutrin i długiego czasu obserwacji, ale ich wielką zaletą jest niewystępowanie „tła" procesów wywołanych innymi oddziaływaniami. Neutrina są jedynymi znanymi cząstkami elementarnymi nie uczestniczącymi ani w oddziaływaniach silnych, ani elektromagnetycznych. Przejawia się to w ich ogromnej przenikliwości przez materię. Neutrino ma realną szansę przejścia nawet na wskroś gwiazdy bez żadnego aktu oddziaływania, a więc bez zmiany swej energii i kierunku lotu! (→ Reakcje jądrowe w gwiazdach). Prawdopodobieństwo tego, że elektron wejdzie w słabe oddziaływanie z materią, przez którą przelatuje, jest tego samego rzędu wielkości co dla neutrina, ale nie sposób wyłowić tych rzadkich przypadków z ogromnej liczby innych procesów wywołanych oddziaływaniami elektromagnetycznymi elektronów z materią (kreacja par, promieniowanie hamowania, rozbijanie jąder i inne).

rozpady słabe

Najwięcej informacji o oddziaływaniach słabych uzyskano obserwując rozpady cząstek elementarnych (i jąder atomowych). Ponieważ badanie reakcji neutrinowych stało się technicznie możliwe stosunkowo późno, oddziaływania słabe przez wiele lat utożsamiano właściwie z rozpadami słabymi, zwanymi też rozpadami powolnymi. Łatwo zrozumieć, dlaczego obserwacja słabych rozpadów nie nastręcza większych trudności. Otóż, jeśli jakaś cząstka może sie rozpaść jedynie wskutek oddziaływań słabych, to niezależnie

od ich słabości proces taki wcześniej czy później musi nastąpić. Prawdopodobieństwo tego zdarzenia wynosi dokładnie 1, podczas gdy prawdopodobieństwo zajścia reakcji rozpraszania przy przechodzeniu neutrina przez określoną warstwę substancji może wynosić np. 10^{-30}! Ewentualna słabość odddziaływania wywołującego rozpad przejawi się jedynie w długości czasu życia rozpadającego się obiektu. Intuicyjnie powinno być zrozumiałe, że im słabsze oddziaływanie, tym dłużej trzeba taką cząstkę obserwować zanim się ona rozpadnie, czyli tym dłuższy będzie jej czas życia.

Typowe czasy życia cząstek elementarnych rozpadających się wskutek oddziaływań słabych wynoszą ok. 10^{-10} s lub więcej. Czy to dużo czy mało? Odpowiedź będzie oczywiście zależeć od tego, z czym ten czas porównujemy.

charakterystyczne wielkości wymiarowe

Jest ogólną regułą w fizyce, że o wielkości mianowanej nie można powiedzieć czy jest duża, czy mała w oderwaniu od innych okoliczności. Małe lub duże mogą być tylko wielkości niemianowane, a więc np. stosunki dwóch wielkości o tym samym wymiarze. W kręgu spokrewnionych zjawisk występują zawsze pewne charakterystyczne wielkości wymiarowe. Na przykład, w teorii względności charakterystyczną prędkością jest prędkość światła c, w fizyce atomowej charakterystyczną wielkością o wymiarze momentu pędu — stała Plancka, a także charakterystyczną jednostką energii — energia jonizacji atomu wodoru itd. Mówiąc, że jakaś wielkość jest bardzo mała (bądź bardzo duża) mamy zawsze na myśli jej stosunek do charakterystycznej dla danego kręgu zjawisk wielkości o tym samym wymiarze. Musimy zatem ustalić charakterystyczne wielkości o różnych wymiarach (w tym także czasu) mające znaczenie dla oddziaływań cząstek elementarnych o wysokich energiach. Niewątpliwie charakterystyczną prędkością jest c, a charakterystyczną długością — średnica cząstki elementarnej. Charakterystycznym momentem pędu jest oczywiście $\hbar$. Z tych trzech wielkości można już utworzyć wszystkie inne mechaniczne wielkości o różnych wymiarach. Kłopot polega jedynie na tym, że w przeciwieństwie do stałych $\hbar$ i c, które mają uniwersalny sens nie związany z tą czy inną cząstką, „średnica cząstki" nie jest dobrze określona, choćby dlatego, że istnieją różne cząstki i nie ma powodu, by któraś z nich była najważniejsza. Jednakże, nie popełnimy wielkiego błędu, jeśli przyjmiemy wartość charakterystycznej długości równą 10^{-13} cm. Teraz można stwierdzić, że czas charakterystyczny dla cząstek elementarnych ma wielkość 10^{-13} cm/$c \approx 3 \cdot 10^{-24}$ s. Wielkością tego rzędu są również niektóre czasy życia cząstek rozpadających się wskutek oddziaływań silnych. Gdyby ten czas przyjąć za jednostkę, to typowy czas życia cząstek rozpadających się wskutek oddziaływań słabych równy 10^{-10} s byłby w tych jednostkach rzędu $3 \cdot 10^{13}$. Jest to liczba ogromna i dlatego rozpady słabe można nazwać rozpadami powolnymi. Przypomnijmy, że typowe czasy życia cząstek rozpadających się silnie zawierają się z grubsza biorąc w zakresie 10^{-24}–10^{-21} s, a rozpadających się elektromagnetycznie — w zakresie 10^{-19}–10^{-16} s.

powolność rozpadów słabych

Liczby te nie tylko pokazują różnicę między poszczególnymi rodzajami oddziaływań, ale mogą być podstawą określenia oddziaływań słabych. Właśnie owa powolność procesu jest jedyną elementarną i zarazem bezpośrednio widoczną cechą charakterystyczną oddziaływań słabych. Są też i inne cechy, bardziej fundamentalne, ale zarazem bardziej abstrakcyjne. Jednej z nich (łamaniu symetrii zwierciadlanej) poświęcimy dalej nieco uwagi.

Rozpady β neutronu i jąder atomowych

Pierwszym rozpadem słabym obserwowanym wiele lat przed pojawieniem się pojęcia oddziaływań słabych był rozpad β jąder atomowych (→ Rozpady jąder atomowych). Istnieją rozpady β^+ i β^-. Najprostszym jądrem ulegającym rozpadowi β^- jest swobodny neutron. Jest zupełnie naturalne wyobrazić sobie, że rozpad β^- jakiegoś bardziej złożonego jądra zachodzi w wyniku rozpadu jednego z neutronów wchodzących w skład jądra. Jak wiadomo, swobodny neutron rozpada się (lepiej by było mówić: samorzutnie się przekształca — słowo rozpad, przyjęte ze względów historycznych, sugeruje, że neutron jest zbudowany z cząstek będących produktami jego rozpadu, a to jest nieprawda) na proton, elektron i antyneutrino elektronowe. Reakcję tę zapisujemy w sposób następujący:

rozpad β^- neutronu

$$n \rightarrow p + e + \bar{\nu}_e + Q_0, \tag{1}$$

gdzie Q_0 — energia kinetyczna cząstek końcowych (produktów rozpadu). Zgodnie ze słynnym wzorem Einsteina, energia ta dana jest wzorem

$$Q_0 = (m_n - m_p - m_e)c^2. \tag{2}$$

Wielkości m_n, m_p i m_e oznaczają masy spoczynkowe odpowiednich cząstek. Pominęliśmy masę neutrina, gdyż zgodnie z doświadczeniem jest ona w granicy dokładności pomiaru równa zeru.

rozpad β^- jądra

Kiedy neutron rozpada się wewnątrz jądra, to wytworzony proton zajmuje jego miejsce w jądrze, co można zapisać tak:

$${}^A_ZX \rightarrow {}^{A}_{Z+1}X + e + \bar{\nu}_e + Q(A, Z),$$

gdzie ${}^A_ZX$ — jądro promieniotwórcze złożone z Z protonów i $A-Z$ neutronów.

Energia wyzwolona w tym procesie wynosi

$$Q(A, Z) = (m_Z - m_{Z+1} - m_e)c^2, \tag{3}$$

gdzie m_Z i m_{Z+1} — masy jąder ${}^A_ZX$ i ${}_{Z+1}^{A}X$. Ale różnica mas jąder nie jest bynajmniej równa różnicy mas spoczynkowych ich składników. W rzeczywistości

$$m_Z = Zm_p + (A-Z)m_n - \frac{B(A, Z)}{c^2}, \tag{4}$$

gdzie B — energia wiązania jądra. Energia wiązania uwarunkowana jest oddziaływaniami silnymi i elektromagnetycznymi nukleonów w jądrze i zależy od A i Z w sposób bardzo złożony. Korzystając z wzorów (2–4) mamy

$$Q(A, Z) = Q_0 + B(A, Z+1) - B(A, Z).$$

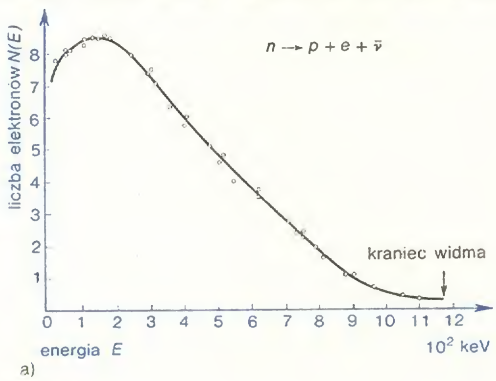
Ze względu na złożony charakter zależności $B(A, Z)$ wielkość $Q(A, Z)$ może być zarówno większa od Q_0, jak też znacznie mniejsza. Najważniejsze jest to, że dla wielu wartości A i Z wielkość $Q(A, Z)$ jest po prostu ujemna i że jądro takie jest trwałe ze względu na rozpad β^-. Gdyby było inaczej, to poza wodorem nie istniałyby we Wszechświecie żadne inne pierwiastki trwałe.

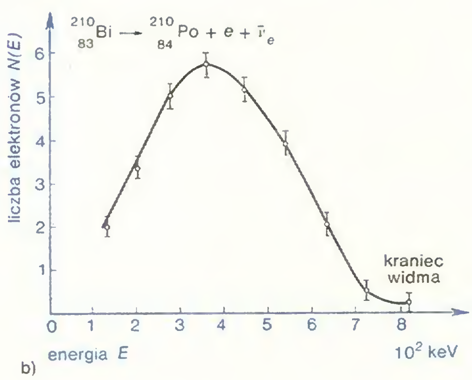
zasady zachowania w rozpadzie β

W procesie rozpadu jest zachowana nie tylko energia ale i pęd. Gdy w stanie końcowym po rozpadzie mamy tylko dwa ciała o określonych masach (spoczynkowych), to łatwo sprawdzić, że dwa prawa zachowania wyznaczają jednoznacznie sposób podziału energii Q między produkty rozpadu. Rozpad β jest jednak rozpadem z trzema cząstkami w stanie końcowym. Uwzględnienie trzeciego ciała nie zwiększa liczby równań, wyrażających prawa zachowania, a jedynie liczbę niewiadomych (energia i pęd trzeciej cząstki, a ponadto rozmaite kąty między kierunkami lotu). Dlatego właśnie prawo zachowania energii i pędu nie wyznacza jednoznacznie sposobu podziału energii Q między cząstki końcowe. Istnieje nieskończenie wiele możliwości zgodnych z prawami zachowania i przyroda w każdym akcie rozpadu wybiera w sposób statystyczny ów sposób rozdziału. Specyfika rozpadu β polega na tym, że jeden z produktów, neutrino, jest cząstką niezwykle przenikliwą, a zatem

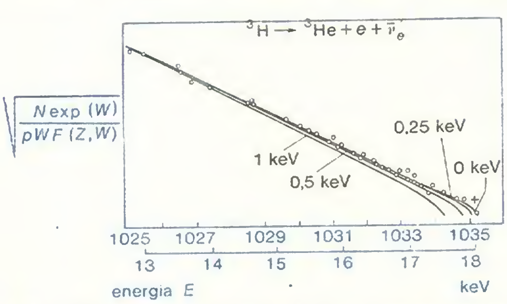
umykającą obserwacji. W praktyce możemy obserwować i mierzyć jedynie energię elektronu i jądra końcowego (bardzo zresztą małą w porównaniu z energią elektronu). Suma tych dwóch energii jest w każdym akcie rozpadu inna i na ogół mniejsza od wartości Q wyznaczonej z wzoru Einsteina (w skrajnym wypadku może być równa). Ponieważ pomijamy w praktyce znikomo małą energię jądra końcowego, więc energia elektronów powstających w wielu aktach rozpadu jest rozrzucona w sposób statystyczny od zera do wartości niemal równej Q, co oznacza, że elektrony z rozpadu β mają widmo ciągłe (rys. 1 i 2). Kształt tego widma był i jest przedmiotem analizy teoretycznej i pomiarów doświadczalnych, gdyż można z niego wiele odczytać. Na przykład jasne jest, że gdyby masa spoczynkowa neutrina była większa od zera, to maksymalna energia elektronów (kraniec widma) powinna być mniejsza od obliczonej z wzoru (2). Przy założeniu, że neutrino istnieje i jest emitowane w procesie rozpadu β, ciągłe widmo energii elektronów nie jest niczym

widmo β





Rys. 1. Typowe widma β. Na osi odciętych odłożona jest energia kinetyczna elektronu E w keV, na osi rzędnych względna liczba elektronów w jednostkowym przedziale energii $N(E) = \Delta N/\Delta E$



Rys. 2. Porównanie widma β rozpadu trytu z przewidywaniami teoretycznymi przy różnych założonych wartościach masy neutrina. Najlepszą zgodność uzyskuje się przy $m_\nu = 0$. E energia kinetyczna elektronu, W całkowita energia elektronu ($E+mc^2$), p pęd elektronu, $F(Z, W)$ znany czynnik uwzględniający m.in. wpływ elektrostatycznego przyciągania jądra końcowego i elektronu β na kształt widma

zagadkowym. Jednakże w czasie, gdy odkryto widmo ciągłe w rozpadzie β nikt nawet nie podejrzewał istnienia neutrina. Przy założeniu zaś, że jądro rozpada się tylko na dwa ciała, widmo ciągłe energii przeczy w oczywisty sposób prawu zachowania energii. Kiedy zmierzono energię i kierunek lotu jądra końcowego, to przekonano się, że również prawo zachowania pędu musiałoby być naruszone. Z kolei analiza spinów cząstek uczestniczących w rozpadzie β (oczywiście przy pominięciu neutrina) wykazała załamanie prawa zachowania momentu pędu. Aczkolwiek niektórzy fizycy traktowali poważnie możliwość naruszenia wszystkich tych praw zachowania, to jednak znaleźli się inni, którzy raczej uwierzyli w hipotezę W. Pauliego postawioną w 1930 r., według której w rozpadzie β musi być emitowana jeszcze jedna cząstka. Wkrótce nawet nadano jej nazwę — neutrino.

hipoteza istnienia neutrina

Jednym z fizyków doceniającym znaczenie hipotezy Pauliego był E. Fermi, który w 1932 r. stworzył pierwszą konsekwentną, acz nie ostateczną teorię oddziaływań słabych, uwzględniającą istnienie neutrina. Przez długie lata była to cząstka hipotetyczna. Dopiero w latach pięćdziesiątych F. Reines i C. Cowan udowodnili bezpośrednio, że neutrino rzeczywiście istnieje, gdyż może (choć niezmiernie rzadko) wywołać nową reakcję w dowolnej odległości od miejsca ego wytworzenia.

rozpad β^+ jądra

Oprócz jąder wykazujących promieniotwórczość β^- znane są także jądra wykazujące promieniotwórczość β^+. Ogólne równanie procesu β^+ ma postać:

$$ {}^A_Z X \to {}^{A}_{Z-1}X + e^+ + \nu_e + Q. \qquad (5)$$

Jasne, że przy $Z = 1$ i $A = 1$ proces (5) byłby rozpadem protonu. Ale dla swobodnego protonu wielkość Q występująca w równaniu (5) jest ujemna (ponieważ proton jest lżejszy od neutronu), zatem swobodny proton jest cząstką trwałą. Mimo to, wygodnie jest wyobrażać sobie, że rozpad β^+ jądra jest rezultatem rozpadu protonu wewnątrz tego jądra. Niezbędną energię „pożycza" proton od całego jądra. Podobny efekt występował także przy rozpadzie β^- wówczas, gdy wyzwolona energia $Q(A, Z)$ była większa od Q_0. Chociaż rozpad swobodnego protonu jest wzbroniony, można rozpad β^- związanego neutronu i rozpad β^+ związanego protonu traktować na zupełnie równych prawach. W tym miejscu ujawnia się nonsensowność rozpatrywania neutronu jako zbudowanego z protonu, elektronu i antyneutrina. Gdyby tak było, to powinniśmy z kolei proton uważać za zbudowany z neutronu, pozytonu i neutrina. Inaczej mówiąc, neutron byłby zbudowany z samego siebie, a ponadto z elektronu, pozytonu, neutrina i antyneutrina, co jest — przynajmniej w dosłownym rozumieniu słowa „zbudowany z" — jawnym bezsensem.

rozpad jako oddziaływanie

Proces rozpadu czyli przekształcania się cząstek jest czymś fundamentalnie nowym, zjawiskiem nieznanym fizyce klasycznej, w której wszelkie procesy sprowadzały się do przegrupowywania składników. Należy wyobrazić sobie, że w momencie rozpadu znika po prostu cząstka pierwotna, na jej miejsce zaś pojawiają się nowe, których przedtem nie było. W takim ujęciu procesy rozpadu są podobne do procesów zderzeń nieelastycznych. Te ostatnie należy rozumieć jako znikanie cząstek początkowych i powstawanie na ich miejsce cząstek końcowych, których nie było przed aktem oddziaływania.

W tym ujęciu termin „oddziaływanie" oznacza możliwość znikania jednych cząstek i pojawiania się na ich miejsce innych.

Zasada krzyżowania

Zajmijmy się teraz dokładniej owym „traktowaniem na równych prawach" procesu rozpadu neutronu (β^-) i rozpadu protonu (β^+). Aby móc sformułować

dokładniej co to znaczy, musimy najpierw rozpatrzyć jaki jest związek między wielkościami charakteryzującymi rozpad neutronu swobodnego, a charakterystykami rozpadu jądra. Nie jest to związek będący równością odpowiednich parametrów. Na przykład, energia maksymalna elektronu $Q(A, Z)$ jest inna niż analogiczna wielkość Q_0, na pewno różne są też czasy życia. Przecież dla wielu wartości A i Z jądro jest trwałe, co odpowiada nieskończonemu czasowi życia tegoż jądra. Porównując odpowiednie dane można się przekonać, że kształty widma elektronów bywają bardzo różne dla różnych jąder (rys. 1). Czyż nie jest to w sprzeczności ze sformułowaniem, że rozpad jądra należy rozumieć jako wynik rozpadu jednego z jego neutronów? Okazuje się, że nie. Rzecz w tym, że wszystkie te procesy są dokładnie opisane identycznym hamiltonianem oddziaływania i jeśli zostanie on poprawnie dobrany do opisu jednego tylko procesu (1), to automatycznie, przez stosowanie uniwersalnych metod teorii kwantowej, ten sam hamiltonian pozwala doskonale opisać wszystko co się da zmierzyć, dla rozpadów rozmaitych jąder. Różnorodność rozpadów β odzwierciedla bogactwo struktury różnych jąder, ale nie wnosi już nic istotnego do poznania własności oddziaływań słabych. Jest to niewątpliwie dużym sukcesem kwantowej teorii pola i dowodem jej wielkiej przydatności do opisu oddziaływań. Pod tym względem teoria ta jest podobna do teorii grawitacji Newtona, dzięki której wystarczy z jednego doświadczenia wyznaczyć stałą grawitacyjną, aby na podstawie tej teorii móc opisywać ruchy wszelkich planet, komet, gwiazd podwójnych, statków kosmicznych itp.

uniwersalność hamiltonianu oddziaływania

Kwantowa teoria pola oprócz tego, że pozwala ze znajomości rozpadu swobodnego neutronu przepowiedzieć dokładnie jak będzie się on zachowywał w jądrze atomowym; przewiduje też, że ten sam hamiltonian opisujący rozpad (1) musi opisać procesy, w których dowolną cząstkę przeniesiemy „na drugą stronę reakcji" z jednoczesną zamianą na antycząstkę (zasada krzyżowania), jak też procesy, w których odwrócimy kierunek przebiegu reakcji.

Jest oczywiste, że reakcję $p \rightarrow n+e^{+}+\nu_e$ można otrzymać z reakcji $n \rightarrow p+e^{-}+\bar{\nu}_e$ przez przeniesienie elektronu i antyneutrina na drugą stronę i zmianę znaku strzałki. Tym samym widzimy, że wszelkie procesy β^- i β^+ są przejawem dokładnie jednego i tego samego oddziaływania. Cała informacja o tym oddziaływaniu jest zawarta w fakcie zachodzenia reakcji (1) oraz w kwantowych regułach obliczeń. Oczywiście przenosząc na drugą stronę dwie cząstki w reakcji rozpadu, dostajemy znów proces rozpadu. Teoria kwantowa wiąże ponadto te rozpady z procesami, w których przenosi się na drugą stronę tylko jedną cząstkę. Są to już procesy rozpraszania, np.

$$\bar{\nu}_e+p \rightarrow n+e^{+}, \tag{6}$$

$$e+p \rightarrow n+\bar{\nu}_e. \tag{7}$$

dowody na istnienie neutrina

Dowód Reinesa i Cowana na istnienie neutrina polegał właśnie na zaobserwowaniu reakcji (6) zainicjowanej przez antyneutrina wydobywające się w ogromnych ilościach z pracującego reaktora jądrowego. Słabość oddziaływania polega na tym, że z miliardów antyneutrin przelatujących przez detektor jedynie kilkadziesiąt wywołało reakcję (6), a reszta przeleciała bez jakichkolwiek oddziaływań. Wspominaliśmy już o tym, że jest zupełnie nierealne próbować obserwować reakcję (7) tymi samymi metodami co reakcję (6). Mimo to proces (7) występuje dość często i jest nieźle zbadany, jednak nie jako proces rozproszeniowy, lecz jako tzw. wychwyt elektronu atomowego, przeważnie z powłoki K. Można się przekonać, że dla pojedynczego protonu i elektronu o prawie zerowej energii zasada zachowania energii zabrania reakcji (7). Wynika to stąd, że

$$m_e+m_p < m_n. \tag{8}$$

Ale dla odpowiednio dobranego jądra (podobnie jak w przypadku rozpadów β^+) można poprawić bilans energetyczny i odwrócić znak nierówności (8). Reakcja

$$e+{}^{A}_{Z}X \rightarrow {}_{Z-1}^{\ \ A}X+\nu_e \tag{9}$$

może zachodzić, jeżeli

$$m_e+m_Z > m_{Z-1}.$$

Nierówność ta jest słabsza od nierówności

$$m_Z > m_{Z-1}+m_e,$$

będącej warunkiem koniecznym rozpadu β^+. A zatem jeśli jądro ulega rozpadowi β^+, to również może następować wychwyt elektronu, ale nie na odwrót. Znane są jądra, dla których nie wystarcza energii na rozpad β^+, a wystarcza na wychwyt elektronu.

Zwróćmy uwagę na to, że z punktu widzenia atomu jako całości, wychwyt elektronu jest procesem rozpadu. Ponieważ w wyniku reakcji wychwytu powstają dwa ciała: neutrino i atom końcowy, końcowa energia atomu (w praktyce jądra) jest ściśle określona. Doświadczalne wykazanie tego faktu pięknie potwierdziło hipotezę o istnieniu neutrina i o spełnieniu prawa zachowania energii w oddziaływaniach słabych. Należy przy tym podkreślić, że czasy życia atomu ze względu na wychwyt elektronu, obliczone na podstawie zasady krzyżowania, są zgodne z doświadczeniem, co przekonuje nas, że wychwyt elektronu jest wywołany tym samym dokładnie oddziaływaniem, co reakcja (1).

Skrętność neutrina i łamanie symetrii zwierciadlanej

różnice między neutrinem a antyneutrinem

Zastanowimy się teraz dlaczego w reakcjach raz występuje neutrino, a raz antyneutrino i czym te cząstki się różnią. W tym celu załóżmy najpierw, że neutrino i antyneutrino są identyczne (podobnie jak foton jest swą antycząstką): $\bar{\nu}_e = \nu_e$. Przepiszemy reakcję (6) z uwzględnieniem powyższego fałszywego założenia

$$\nu_e+p \rightarrow n+e^{+} \tag{10}$$

oraz reakcję jaką dostajemy z (1) po przeniesieniu $\bar{\nu}_e$ na lewą stronę

$$\nu_e+n \rightarrow p+e^{-}. \tag{11}$$

Ze wzorów (10) i (11) wynika, że bombardując „neutrinami" z reaktora, materię złożoną z równej (z grubsza) liczby neutronów i protonów powinniśmy obserwować jednakową liczbę reakcji, w których powstają elektrony i reakcji z utworzeniem pozytonu. Jest to w jawnej sprzeczności z doświadczeniem — obserwujemy w takim wypadku tylko proces (10), a nigdy (11). Innymi słowy, cząstki z reaktora są antyneutrinami, a w reakcji (10) nie wolno opuścić kreski nad symbolem ν_e. Można więc uznać, że antyneutrina to te cząstki, które oddziałując z nukleonami przekształcają się w pozytony (antyelektrony), a neutrina to te, które przekształcają się w elektrony.

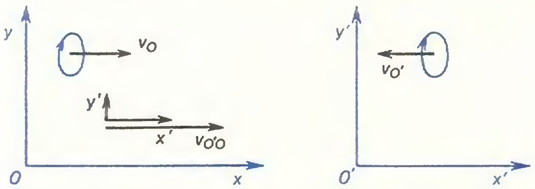
Zgodnie z zasadą krzyżowania, jeśli ani elektronów, ani neutrin nie było na początku (jak w rozpadach $\beta^{\pm}$), to wraz z elektronem zostanie wytworzone antyneutrino, a wraz z antyelektronem (pozytonem) — neutrino. Fakty te pozwalają wprowadzić pojęcie liczby leptonowej i zasadę zachowania tej liczby. Przypisując elektronowi i neutrinu liczbę leptonową 1, a ich antycząstkom liczbę -1, widzimy, że we wszystkich procesach dotychczas omawianych suma tych liczb nie ulega zmianie w wyniku reakcji. Korzystając z prawa zachowania liczby leptonowej, zawsze możemy ustalić czy dana cząstka jest neutrinem czy antyneutrinem (wystarczy znać jej pochodzenie bądź zaobserwować wywołaną przezeń reakcję). Oczywiście można zapytać, czy poza liczbą leptonową neutrino i antyneutrino różnią się jeszcze jakąś inną, mierzalną cechą. Odpowiedź na to jest twierdząca. Neutrina, podobnie jak elektrony i nukleony,

liczba leptonowa

mają własny moment pędu (spin) z tym, że elektron (lub nukleon) może mieć rzut spinu na kierunek lotu równy bądź $+{}^1/_2\hbar$, bądź $-{}^1/_2\hbar$, a neutrino ma zawsze rzut spinu na kierunek pędu równy $-{}^1/_2\hbar$.

skrętność neutrina

Neutrino jest pod tym względem zupełnie wyjątkowe. Jak mówimy obrazowo, neutrino zawsze „wiruje" w lewo. Natomiast antyneutrino zawsze „wiruje" w prawo. Zauważmy, że aby stwierdzenia te miały sens, neutrino musi być cząstką o masie dokładnie równej zeru, a zatem musi zawsze poruszać się z prędkością światła. W przeciwnym wypadku łatwo popaść w sprzeczność. Rozpatrzmy dwóch obserwatorów O i O' (rys. 3). Niech w układzie odniesienia O neu-



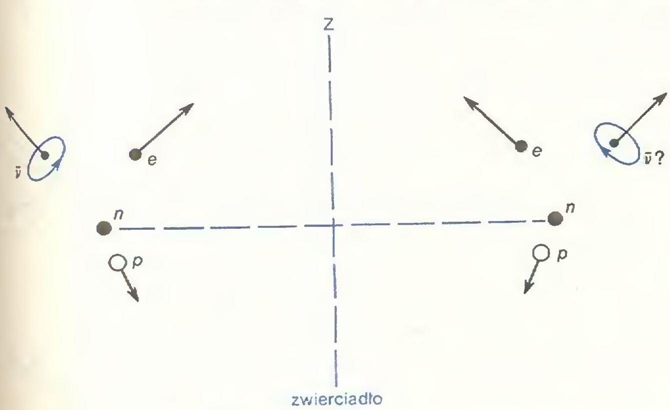
Rys. 3. Neutrino obserwowane przez dwóch obserwatorów: obserwatora O, względem którego porusza się w dodatnim kierunku osi x i obserwatora O' poruszającego się względem O szybciej niż neutrino ($v_{OO'} > v_O$). Obserwator O' widzi, że neutrino porusza się w ujemnym kierunku osi x' i że wiruje w prawo w stosunku do swej prędkości $\vec{v}_{O'}$

trino leci w dodatnim kierunku osi x i wiruje w lewo w stosunku do kierunku swego lotu. Jeśli prędkość neutrina v_O jest mniejsza od c, to przy prędkości $v_{O'O}$ obserwatora O' względem układu O większej od v_O zaobserwuje on w układzie O' ten sam kierunek wirowania względem osi x', ale tym razem kierunek lotu neutrina będzie przeciwny (obserwator O' wyprzedza neutrino). Zatem obserwator O' powie, że neutrino wiruje w prawo w stosunku do kierunku swego lotu.

Fakt określonej, zawsze jednakowej skrętności neutrin ma ważne konsekwencje. Fizyka klasyczna, jak również fizyka kwantowa, ale w zakresie zjawisk wywoływanych oddziaływaniami silnymi i elektromagnetycznymi, wykazywała pewną symetrię zwaną symetrią zwierciadlaną.

symetria zwierciadlana

Przedstawiając poglądowo, symetria zwierciadlana polega na tym, że jeśli jakiś proces jest możliwy (to znaczy zgodny z prawami fizyki), to również zgodny z prawami fizyki, a więc możliwy do realizacji, jest proces podobny do rozpatrywanego, ale w którym wszelkie ruchy i konfiguracje są takie, jakie by widział obserwator obserwujący odbicie tego pierwszego procesu w zwierciadle. Na przykład, śruba prawoskrętna z odpowiednią nakrętką działa zgodnie ze swym przeznaczeniem. Gdybyśmy popatrzyli na śrubę z nakrętką (oraz na proces ich skręcania) w zwierciadle, to by się nam wydawało, że jest to śruba lewoskrętna. Ale taką



Rys. 4. Rozpad β neutronu i jego obraz w zwierciadle Z. Obraz antyneutrina w zwierciadle wiruje w lewo. W rzeczywistości procesu takiego nigdy nie obserwujemy

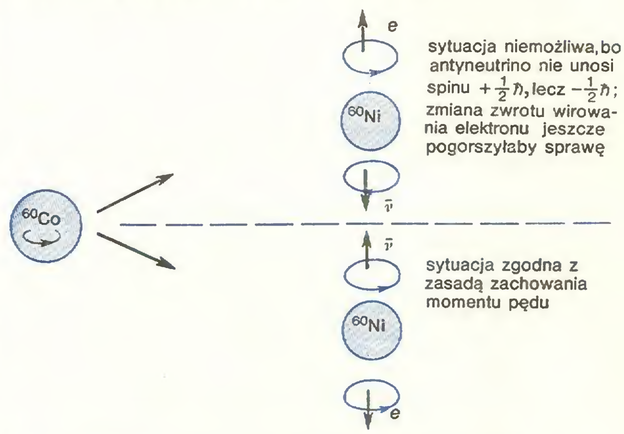
lewoskrętną śrubę można skonstruować naprawdę i będzie ona równie użyteczna w zastosowaniach. Podobnie, przywykliśmy do zasad ruchu prawostronnego na drogach. Jeśli jednak wszystkie pojazdy zmienią ruch na lewostronny, to nadal będzie to konsekwentna metoda unikania kolizji.

łamanie symetrii zwierciadlanej

Przykłady takie można mnożyć w nieskończoność. Tymczasem spójrzmy na obraz w zwierciadle jakiegokolwiek procesu z udziałem neutrin, np. na obraz procesu (1) (rys. 4). Ujrzymy neutron rozpadający się na proton, elektron i pewną cząstkę wirującą w lewo, a więc nie tak jak antyneutrino w rzeczywistym procesie, lecz jak neutrino. Taki proces jest jednak niemożliwy. Przyroda nigdy go naprawdę nie realizuje. (Nie wiemy dlaczego, ale tak po prostu jest!) Można więc stwierdzić, że przy oddziaływaniach słabych nie jest spełniona zasada symetrii zwierciadlanej. Ten sam fakt określa się też jako niezachowanie parzystości w oddziaływaniach słabych. Naruszenie symetrii zwierciadlanej prowadzi do zaskakujących asymetrii, sprzecznych całkowicie z intuicją ukształtowaną przez fizykę klasyczną.

doświadczenie Wu

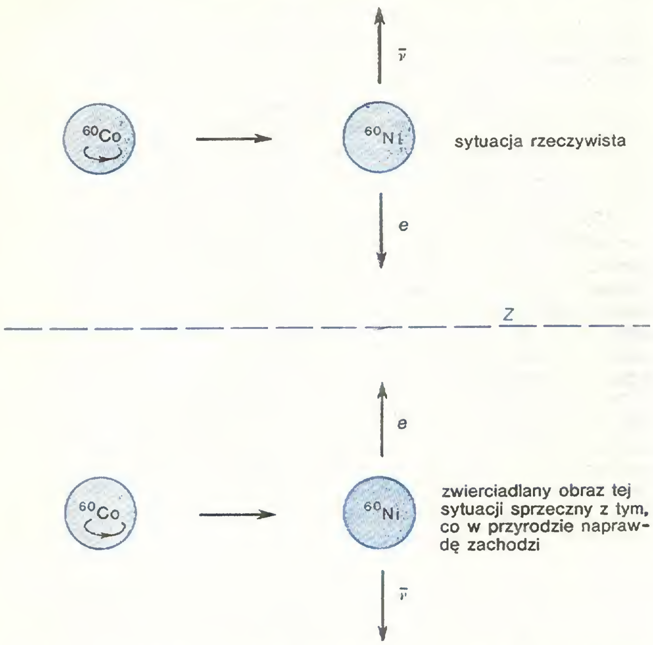
Pierwszym doświadczeniem, w którym taką asymetrię wykryto, było doświadczenie wykonane przez panią C.S. Wu i współpracowników, a zaproponowane przez T.D. Lee i N. C. Yanga. Doświadczenie polegało na badaniu rozpadów β^+ jąder kobaltu ^{60}Co. Decydujące o zaobserwowaniu asymetrii było oziębienie próbki kobaltu do bardzo niskiej temperatury i umieszczenie jej w silnym polu magnetycznym. Jądra kobaltu mają spin i związany z nim moment magnetyczny. W normalnych



Rys. 5. Asymetria „góra-dół" w rozpadzie β^- jądra ^{60}Co. Strzałka na jądrze kobaltu nie symbolizuje całego spinu jądra lecz jedynie różnicę (równą $+1\hbar$) między spinem ^{60}Co i ^{60}Ni. Różnica ta uniesiona musi być przez $\bar{\nu}+e$. Lecące w dół antyneutrino nie może jednak unieść tego momentu pędu, bo musi wirować tak jak na rysunku (w prawo do kierunku swego lotu). Na dolnym rysunku sytuacja jest już zgodna z prawami zachowania i ustaloną skrętnością antyneutrina

warunkach kierunki owych momentów (czyli „płaszczyzny wirowania" jąder) są ustawione chaotycznie. W silnym polu magnetycznym wszystkie momenty magnetyczne ustawiają się równolegle do pola magnetycznego. Można też powiedzieć, że wszystkie jądra wirują w płaszczyznach równoległych i w tę samą stronę. Załóżmy, że płaszczyzna jest pozioma a jądra wirują np. w lewo jeśli patrzymy na próbkę „z góry" (rys. 5) Wiadomo, że jądra ^{60}Co wykazują promieniotwórczość β^-. W wyniku rozpadu powstaje jądro ^{60}Ni, elektron i antyneutrino. Jądro ^{60}Ni ma spin mniejszy o $1\hbar$ od jądra ^{60}Co. Zgodnie z zasadą zachowania momentu pędu ów jednostkowy spin musi być uniesiony przez elektron i antyneutrino. Ponieważ każda z tych cząstek może mieć rzut spinu na oś pionową równy $\pm{}^1/_2\hbar$, to aby z rzutów tych złożyć $+1\hbar$, zarówno spin elektronu jak i spin antyneutrina muszą wynosić $+{}^1/_2\hbar$. Gdyby antyneutrino, które musi kręcić się w prawo w stosunku do swego pędu, leciało

„w dół", to rzut jego spinu na oś pionową byłby $-{}^{1}/{}_{2}\hbar$ i prawo zachowania momentu pędu nie mogłoby być spełnione. W efekcie kierunek „w górę" jest dla antyneutrin uprzywilejowany. Z tego, że suma wszystkich pędów musi być równa zeru, wynika, że dla elektronów uprzywilejowany jest kierunek „w dół" — tam leci faktycznie większość elektronów.



Rys. 6. Dozwolona konfiguracja po rozpadzie ^{60}Co i jej obraz w zwierciadle poziomym. Na obrazie tym antyneutrina lecą w dół, a więc tak jak im faktycznie nie wolno

Spoglądając na odbicie tej sytuacji w zwierciadle płaskim równoległym do płaszczyzny „wirowania" (rys. 6), ujrzymy sytuację sprzeczną z prawem przyrody. W odbiciu tym kierunek „wirowania" jąder nie ulegnie zmianie, a elektron zamiast głównie „w dół" lecieć będzie głównie „do góry", czyli na odwrót niż w rzeczywistym doświadczeniu.

Za postawienie hipotezy, że oddziaływania słabe naruszają symetrię prawo-lewo, fizycy N.C. Yang i T.D. Lee dostali nagrodę Nobla w tym samym roku, w którym doświadczenie pani Wu hipotezę tę potwierdziło (1957 r.).

Oddziaływania słabe mionów

pary leptonów

W przyrodzie oprócz pary leptonów (e, ν_e) występuje inna para leptonów: mion i neutrino mionowe (μ, ν_μ) oraz ich antycząstki. Przekonano się, że oddziaływanie tych cząstek z nukleonami jest identyczne jak pary (e, ν_e). Dokładniej mówiąc, identyczny co do postaci jest hamiltonian oddziaływania, jeżeli zastąpimy w nim symbole elektronu i neutrina elektronowego odpowiednimi symbolami mionu i neutrina mionowego. Nie trzeba zmieniać nawet stałej sprzężenia, jest ona taka sama („uniwersalna"). Oczywiście, obliczając dalej prawdopodobieństwa procesów z udziałem mionów, wszędzie w odpowiednich miejscach wstawiamy masę mionu a nie elektronu i to prowadzi do oczywistych różnic całkowicie jednoznacznie uwzględnianych przez teorię. I tak np. masa mionu jest większa od różnicy mas dowolnych dwóch jąder o tej samej liczbie masowej A — nie obserwujemy więc w ogóle procesów analogicznych do rozpadu β, ale z udziałem mionów. Na odwrót, proces analogiczny do wychwytu elektronu jest energetycznie zawsze możliwy, bo

$$m_\mu + m_p > m_n, \quad m_\mu + m_Z > m_{Z-1},$$

gdyż masa mionu jest równa 106 MeV/c^2, a różnice mas izobarów (jąder o tym samym A) nie przekraczają 10 MeV/c^2. Dodatkowym czynnikiem faworyzującym wychwyt mionu jest fakt, że mion w mezoatomie (atomie, w którym jeden z elektronów zastąpiony jest mionem) jest średnio znacznie bliżej jądra niż elektrony z najgłębszych nawet powłok. Powoduje to, że czas życia mezoatomu ze względu na wychwyt mionu jest o wiele rzędów wielkości krótszy od analogicznego czasu dla wychwytu elektronu. Warto podkreślić jeszcze raz, że i w tym wypadku czas życia można jednoznacznie obliczyć (zgodnie z doświadczeniem) przy założeniu dla mionów tej samej postaci i „intensywności" oddziaływania, którą wyznaczono z procesu (1).

Podobnie reakcja

$$\bar{\nu}_\mu + p \to \mu^+ + n \tag{12}$$

w pełni analogiczna do reakcji (6) była obserwowana. Porównując reakcje (6) i (12) można szczególnie dobitnie wykazać sens uniwersalności oddziaływań słabych mionów i elektronów. Otóż dla tych reakcji przeprowadzonych przy bardzo dużych energiach antyneutrin (dużo większych od energii spoczynkowej mionu) można oczekiwać, że wpływ masy mionu na przekrój czynny powinien być pomijalny. Ale jeśli tak, to w tych warunkach reakcje (6) i (12) powinny zachodzić jednakowo często (przy tej samej energii $\bar{\nu}_e$ i $\bar{\nu}_\mu$) — już bez potrzeby uwzględniania jakichkolwiek poprawek. Po prostu przy tych samych natężeniach antyneutrin w dwóch analogicznych doświadczeniach i tym samym czasie naświetlania powinno być tyle samo reakcji (6) w jednym doświadczeniu, co reakcji (12) w drugim. Wniosek ten jest zgodny z doświadczeniem.

różnice między ν_e a ν_μ

Nasuwa się więc pytanie, co różni ν_e od ν_μ. Otóż jedyna znana różnica polega na tym, że neutrina pierwszego typu wywołują reakcje z elektronami w stanie końcowym, a neutrina drugiego typu — zawsze z mionami. Konkretnie, antyneutrina powstające wraz z mionem w rozpadzie pionu

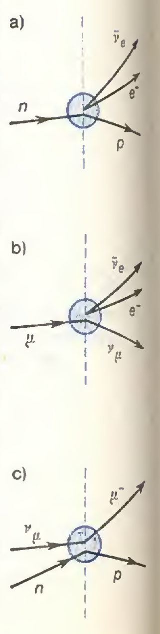
$$\pi^- \to \mu^- + \bar{\nu}_\mu$$

wywołują reakcję (12), a nigdy reakcję (6). Fakt ten pozwala na zdefiniowanie, oprócz wprowadzonej już elektronowej liczby leptonowej L_e, nowej liczby leptonowej L_μ równej $+1$ dla μ^- i ν_μ, -1 dla μ^+ i $\bar{\nu}_\mu$ oraz 0 dla pozostałych cząstek. Obie liczby leptonowe — z dokładnością na jaką pozwala obecna technika doświadczalna — spełniają niezależne prawa zachowania.

Znamy już dwie reakcje (1) i (12), z których przez krzyżowanie można uzyskać wszelkie możliwe procesy z udziałem nukleonów i leptonów określonego typu. Skoro jednak para (n, p) oddziałuje z parą (e, ν_e) i parą (μ, ν_μ) i to w ten sam sposób, to powstaje pytanie, czy pary (e, ν_e) i (μ, ν_μ) mogą oddziaływać ze sobą. Okazuje się, że tak. Prototypem jest tu reakcja rozpadu mionu na neutrino mionowe, elektron i antyneutrino elektronowe:

$$\mu^- \to \nu_\mu + e^- + \bar{\nu}_e. \tag{13}$$

Jest ona bardzo podobna do rozpadu swobodnego neutronu na proton, elektron i antyneutrino elektronowe. (Reakcja (1) przechodzi dokładnie w (13) jeśli zastąpimy $n \to \mu^-$, $p \to \nu_\mu$). Czas życia mionu wynosi 10^{-6} s. Okazuje się, że jeśli do obliczenia tego czasu użyć hamiltonianu oddziaływania o takiej samej postaci jak hamiltonian opisujący rozpad (1) i z tą samą stałą sprzężenia, to dostanie się wynik zaledwie o 5% mniejszy od zmierzonej wartości. Owa rozbieżność jest większa od dokładności pomiaru — powrócimy jeszcze do tego problemu.



Rys. 7. Trzy prototypowe oddziaływania: a) pary (n, p) z parą (e, ν_e); b) pary (μ, ν_μ) z parą (e, ν_e) i pary (n, p) z parą (μ, ν_μ). Przenosząc linię którejkolwiek cząstki na drugą stronę linii przerywanej z jednoczesną zamianą na antycząstkę możemy dostać diagramy wielu innych możliwych procesów

Podsumujmy nasze wiadomości o oddziaływaniach słabych omówionych dotychczas. Sześć cząstek, które rozpatrywaliśmy (n, p, e, ν_e, μ, ν_μ) grupują się w trzy pary (n, p), (e, ν_e) i (μ, ν_μ). Istnieją trzy „elementarne" procesy, w których uczestniczą dwie kompletne pary, a więc cztery cząstki. Można je symbolicznie przedstawić w postaci diagramów (rys. 7).

Powyższe diagramy można interpretować jako ślady torów cząstek uczestniczących w procesie. Po lewej stronie linii przerywanej narysowano cząstki początkowe, po prawej — produkty reakcji. Zgodnie z zasadą krzyżowania, z każdego diagramu można otrzymać wiele innych diagramów przez odchylenie jednej (lub więcej) linii tak, by znalazły się one po przeciwnej stronie linii przerywanej z jednoczesną zmianą cząstki na antycząstkę.

Oddziaływania słabe innych cząstek. Model kwarkowy

Powyżej zostało przedstawione oddziaływanie słabe sześciu cząstek. Zadziwiające jest, że cała różnorodność procesów opisana jest jedną formułą matematyczną (hamiltonian oddziaływania) i jedną stałą sprzężenia (jeśli pominąć 5% rozbieżność). Wiemy jednak, że liczba cząstek jest bardzo duża i wszystkie one uczestniczą w słabych oddziaływaniach. Nie obserwujemy wprawdzie rozpadów słabych wszystkich znanych cząstek nietrwałych, ale to tylko dlatego, że niektóre cząstki są nietrwałe ze względu na inne oddziaływania i żyją zbyt krótko, by powstała realna szansa zauważenia ich słabego rozpadu, który jednak w zasadzie jest możliwy. O tym, że cząstki takie uczestniczą w słabych oddziaływaniach może nas przekonać np. fakt zachodzenia reakcji

$$\nu_\mu + p \rightarrow \mu^- + \Delta^{++}$$

(cząstka Δ^{++} → Cząstki elementarne i ich oddziaływania).

Ale nawet ograniczając się do rozpadów tylko tych cząstek, które nie mogą rozpadać się wskutek oddziaływań innych niż słabe, napotykamy na ogromną różnorodność procesów. Na przykład, mezon K^- (i podobnie K^+) rozpada się na co najmniej 26 znanych różnych sposobów (tabela cząstek w → Cząstki elementarne i ich oddziaływania). Wydaje się zrazu, że idea uniwersalności postaci oddziaływań słabych musi się załamać. Przemawia za tym występowanie procesów powolnych, w których nie uczestniczą bynajmniej cztery cząstki o spinie $^1/_2$, lecz np. trzy lub pięć, z tego część cząstek o spinie 0, np.

$$\pi^- \rightarrow \mu^- + \bar{\nu}_\mu, \quad K^- \rightarrow \pi^- + \pi^+ + e^- + \bar{\nu}_e.$$

W teorii pola, hamiltonian opisujący oddziaływanie czterech cząstek o spinie $^1/_2$ nie może opisywać innych oddziaływań niż czterech cząstek o spinie $^1/_2$. Do opisu powyższych procesów należałoby użyć dwu nowych hamiltonianów. Oczywiście nic nie zabrania Przyrodzie być tak skomplikowaną. Szokujące jest natomiast to, że na przekór pozorom, ideę uniwersalności oddziaływań słabych można utrzymać (uzyskując doskonały opis rzeczywistości), i że w gruncie rzeczy nieznaczne uogólnienie poznanych już oddziaływań pozwala na opisanie różnorodnych rozpadów cząstek elementarnych. Rzecz w tym, że to co nazywamy cząstkami elementarnymi stanowi w większości struktury złożone. Ogromnym postępem w kierunku uzyskania zwartego opisu oddziaływań słabych jest przyjęcie, że wszystkie hadrony są zbudowane z niewielkiej liczby kwarków. Jest to przeniesienie modelu z poziomu jądra, według którego wszystkie jądra są złożone z dwóch rodzajów nukleonów, a rozpady jąder są odzwierciedleniem jednego procesu elementarnego, co bardzo upraszcza sytuację. Gdybyśmy ignorowali fakt złożoności jąder, to do opisu rozpadu β każdego jądra musielibyśmy używać innego hamiltonianu i innej stałej sprzężenia. W tej sytuacji wydawałoby się, że z własności rozpadu jednego jądra nie można wywnioskować nic, co dotyczyłoby rozpadu innego jądra.

model kwarkowy hadronów

Hadrony (znane do 1974 r.) można było zbudować w określony sposób z kwarków trzech rodzajów: d, u, s, w tym nukleony tylko z kwarków d i u o ładunkach $-^1/_3e$ i $+^2/_3e$ (e — ładunek elementarny). W modelu tym proton i neutron to układy złożone z kwarków w sposób następujący:

$$n = (d\ d\ u), \text{ a } p = (u\ u\ d).$$

oddziaływania słabe kwarków

Przyjmijmy, że rozpad nukleonu jest spowodowany rozpadem jednego z jego kwarków składowych. Proces (1) zastępujemy przez

$$d \rightarrow u + e^- + \bar{\nu}_e. \tag{14}$$

Mimo ułamkowych ładunków kwarków widać, że reakcja ta zachowuje ładunek całkowity. Fakt, że kwark rozpada się wewnątrz hadronu, nie stanowi żadnej przeszkody — ten sam problem występował w poprzedniej wersji rozpadu jądra — teoria kwantowa dobrze sobie z tym radzi.

Spójrzmy pod tym kątem na reakcję rozpadu

$$\pi^- \rightarrow \pi^0 + e^- + \bar{\nu}_e. \tag{15}$$

W modelu kwarkowym — π^- to układ związany antykwarku $\bar{u}$ i kwarku d, a π^0 to superpozycja układu ($\bar{u}\ u$) i ($\bar{d}\ d$). Rozpad (15) możemy traktować dokładnie tak, jak rozpad β^- jądra. Zakładamy więc, że d rozpada się na u, e^- i $\bar{\nu}_e$:

$$(\bar{u}\ d) \xrightarrow{\beta^-} (\bar{u}\ u) + e^- + \bar{\nu}_e.$$

Możliwy jest też rozpad antykwarku $\bar{u}$:

$$\bar{u} \rightarrow \bar{d} + e^- + \bar{\nu}_e \text{ (zasada krzyżowania)}$$

prowadzący do procesu

$$(\bar{u}\ d) \xrightarrow{\beta^-} (\bar{d}\ d) + e^- + \bar{\nu}_e.$$

W wyniku równoległego zachodzenia tych procesów dostajemy na końcu dokładnie tę suprepozycję ($\bar{u}\ u$) i ($\bar{d}\ d$), która jest mezonem π^0. Prawdopodobieństwo tego rozpadu można obliczyć dokładnie tak samo, jak rozpadu β^- jądra i wtedy uzyska się — bez żadnych nowych parametrów czy założeń — wynik zgodny z doświadczeniem. Umawiamy się, że zamiast pary (n, p) uważamy odtąd za elementarną parę (d, u).

opis rozpadu cząstek dziwnych

W celu opisania oddziaływań słabych cząstek dziwnych trzeba ustalić, w jaki sposób kwark s (kwark dziwny) uczestniczy w słabych oddziaływaniach. Informacji tych dostarczyć może rozpad

$$K^- \rightarrow \pi^0 + e^- + \bar{\nu}_e. \tag{16}$$

W modelu kwarkowym K^- jest układem ($\bar{u}\ s$), π^0 zaś nie zawiera kwarku s. Zatem kwark ten musi rozpadać się w trakcie procesu (16) na p, e^- i $\bar{\nu}_e$:

$$s \rightarrow u + e^- + \bar{\nu}_e. \tag{17}$$

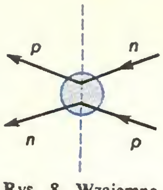
Jest to proces bardzo podobny do (14), którym zastąpiliśmy „złożony" proces (1) — jedyna różnica, to zamiana d na s. Stosując standardową teorię do procesu (16) (z tą samą stałą sprzężenia) dostalibyśmy czas życia zbyt krótki o czynnik ok. 20. Klucz do rozwiązania zagadki może stanowić fakt, że po raz pierwszy napotykamy parę cząstek (s, u), w której składnik u wystąpił już w innej parze (d, u). W rezultacie para (d, u) sprzęga się nieco słabiej niż para (μ, ν_μ) a para (s, u) sprzęga się najsłabiej.

kąt Cabibbo

Wprowadzając jedną parę z kombinacją kwarków d i s

$$(d \cos\theta + \lambda \sin\theta,\ u)$$

sprzężoną równie silnie jak para (e, ν_e) bądź (μ, ν_μ) można uzyskać doskonały opis wszystkich znanych



Rys. 8. Wzajemne oddziaływanie słabe kwarków d i u

procesów, jeżeli tylko kąt θ (kąt abibbo) przyjmiemy równy 0,22;

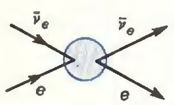
$$\sin\theta = 0{,}22 \quad (\sin^2\theta \approx 1/20 = 5/100).$$

Dlaczego wartość tego kąta wynosi właśnie tyle — nie wiemy. Jest to jeden z dwóch parametrów charakteryzujących słabe oddziaływania (obok stałej Fermiego G), który na obecnym etapie wiedzy brać musimy z doświadczenia.

Nasuwa się pytanie, czy oprócz procesów z udziałem 2 różnych par („dubletów") występują słabe procesy, w których dublet oddziałuje z samym sobą, np.

$$d+u \rightarrow u+d, \tag{18}$$

co symbolizuje rys. 8. Okazuje się, że tak. Nie jest to łatwo potwierdzić. Rzecz w tym, że proces $d+u \rightarrow u+d$ zachodzi też wskutek oddziaływań silnych. Jednakże oddziaływania słabe naruszają parzystość. Chociaż więc są odpowiedzialne za znikomy ułamek potencjału oddziaływania d u (lub n-p — rzeczywisty neutron — rzeczywisty proton), to jedynie ten ułamek może prowadzić do naruszenia symetrii lewo-prawo i w efekcie do tego, że jądra atomowe wysyłają czasami promieniowanie γ o nieznacznej przewadze jednego rodzaju polaryzacji kołowej nad drugą.



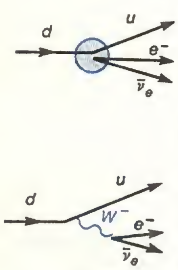
Rys. 9. Oddziaływanie elastyczne $\bar{\nu}_e+e \rightarrow \bar{\nu}_e+e$

Zaobserwowano również (z ogromnym trudem) rozpraszanie elastyczne antyneutrin na elektronach, a więc reakcję

$$\bar{\nu}_e+e \rightarrow \bar{\nu}_e+e,$$

którą przedstawia rys. 9 (diagram).

Najprostszy sposób naturalnego uwzględnienia tych wszystkich procesów polega na przypuszczeniu, że istnieją dwie cząstki o dużej masie, spinie 1 i ładunkach $\pm 1e$, tzw. bozony pośrednie W^+ i W^-, które sprzęgają się (jednakowo silnie!) do każdego z trzech dubletów. Diagram procesu czterofermionowego np. (14) przyjmie więc nową postać. Można go interpretować jako rozpad d na u i W^-, który dalej może zamienić się na parę e i $\bar{\nu}_e$ (rys. 10). Istnienie rozpraszania elastycznego (np. 18) jest w tym obrazie zupełnie naturalne (rys. 11).



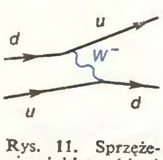
Rys. 10. Przejście od obrazu oddziaływania czterofermionowego do oddziaływania z udziałem bozonu pośredniego

W ten sposób podaliśmy niezbędne informacje potrzebne do opisu oddziaływań słabych znanych do 1973 r. Rok ten przyniósł odkrycie nowego rodzaju sprzężenia, a rok następny odkrycie cząstek, do których opisu stało się niezbędne wprowadzenie czwartego kwarku c (powabnego).

Oba te fakty zostały przewidziane przez słynną teorię Weinberga–Salama oddziaływań słabych (rozwiniętą w pracy Glashowa, Iliopoulosa i Maianiego). Analizując problem rozbieżności w wyższych rzędach rachunku zaburzeń kwantowej teorii pola, doszli oni do wniosku, że przedstawioną tu teorię można i trzeba ulepszyć (zmodyfikować). Niezbędna modyfikacja polega na wprowadzeniu czwartego kwarku c, a raczej czwartego dubletu

$$(-d\sin\theta+s\cos\theta, c) \tag{19}$$

opisującego w pełni słabe rozpady cząstek, w których skład wchodzi ów czwarty kwark c, oraz jeszcze jednego bozonu Z (nienaładowanego) sprzęgającego się w określony sposób z linią, w której nie następuje zmiana tożsamości cząstki (czy będzie to lepton, czy kwark).



Rys. 11. Sprzężenie dubletu (d u) do bozonu pośredniego prowadzi w oczywisty sposób do możliwości oddziaływań elastycznych

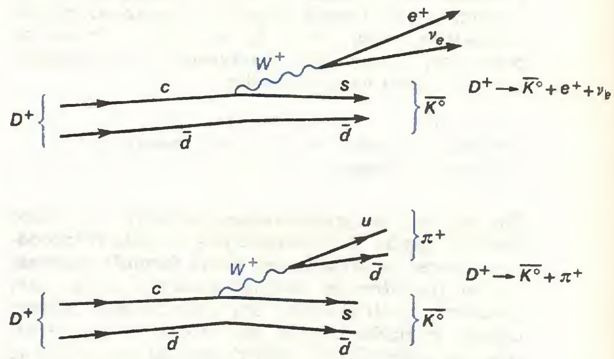
Wiele przewidywań tej teorii zostało już potwierdzonych. Tak więc odkryto w 1977 r. już wszystkie sześć cząstek powabnych o spinie 0 (spokrewnionych z mezonami π i K) i stwierdzono, że rozpadają się one w sposób, który przewiduje teoria. Pomijając $\sin^2\theta$ jako wielkość małą, moglibyśmy w zerowym przybliżeniu dwa dublety kwarków przedstawić w postaci (d, u) i (s, c). W tym przybliżeniu kwark c musi rozpadać się na kwark s i bozon W^+. Ten ostatni może się z kolei przekształcać np. na parę e^+ i ν_e, co prowadzi do procesu

$$D^+ \rightarrow \overline{K^0}+e^++\nu_e$$

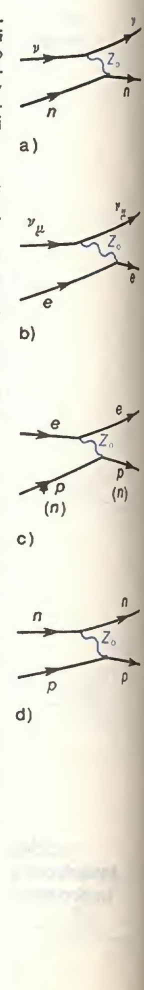
lub

$$D^+ \rightarrow \overline{K^0}+\pi^+,$$

jeżeli W^+ rozpadnie się akurat na parę (u, $\bar{d}$) (rys. 12). Jest tutaj istotne, że w stanie końcowym wystąpi (prawie zawsze) cząstka dziwna. Dokładniej, nastąpi to w 95%, bo tyle wynosi $\cos^2\theta$.



Rys. 12. Dwa spośród wielu możliwych rozpadów słabych cząstki D^+, w skład której wchodzi kwark c rozpadający się głównie na kwark s i bozon W^+. Kwark s na rysunku można zastąpić kwarkiem d, ale według teorii prawdopodobieństwo tych nowych procesów byłoby mniejsze o czynnik $\mathrm{tg}^2\theta_c \approx 1/20$



Rys. 13. Niektóre procesy z wymianą bozonu neutralnego. Procesy a), b) i c) są zupełnie nieoczekiwane w klasycznej wersji teorii słabych oddziaływań. Proces c) powinien prowadzić do łamania symetrii zwierciadlanej w procesach związanych z absorpcją i emisją światła przez atomy. Proces d) wnosi istotny wkład do procesu przedstawionego na rys. 11 modyfikując przewidywanie starej teorii dotyczące łamania symetrii zwierciadlanej w fizyce jądrowej

Natomiast założenie istnienia bozonu Z_0 prowadzi do przewidywania procesów niemożliwych w poprzedniej wersji teorii, a mianowicie: $\nu+d \rightarrow \nu+d$, $\nu_\mu+e \rightarrow \nu_\mu+e$ itp. przedstawionych na rys. 13.

Oba te procesy zostały doświadczalnie potwierdzone. Określa się je jako oddziaływania „poprzez prądy neutralne", co oznacza po prostu, że przy przechodzeniu cząstki przez punkt oddziaływania nie następuje zmiana jej ładunku.

Najciekawszy i najbardziej kontrowersyjny jest proces $e+p$ (lub n) $\rightarrow e+p$ (lub n) przedstawiony na rys. 13c. Jest to proces elastycznego rozpraszania elektronu na nukleonie. Proces ten jest maskowany oczywiście normalnymi oddziaływaniami elektromagnetycznymi, ale podobnie jak przy oddziaływaniach nukleonów prowadzi on do efektywnego potencjału elektron-nukleon (a więc i elektron-jądro) naruszającego parzystość (o efekcie tym wspominaliśmy już we wstępie). Powoduje to m.in. bardzo nieznaczne skręcenie płaszczyzny polaryzacji światła przechodzącego przez jednoatomowe pary metali ciężkich. W pierwszych doświadczeniach (niedokładnie jeszcze sprawdzonych) efektu nie stwierdzono, a w każdym razie nie jest on większy niż ok. $1/5$ tego co przewiduje teoria Weinberga–Salama. Do roku 1979 potwierdzono istnienie tego oddziaływania tak jak przewiduje teoria, czego konsekwencją było przyznanie Weinbergowi, Salamowi i Glashowowi nagrody Nobla z fizyki. Czy jest to nowe wyzwanie, które zmusi do modyfikacji tej pięknej teorii, czy tylko błąd w analizie danych — jeszcze nie wiadomo. Przypuszczalnie problem zostanie rozstrzygnięty, zanim ten artykuł dotrze do Czytelników.

L.A. OKUŃ *Słabe oddziaływania cząstek elementarnych*, Warszawa 1966; R. SOSNOWSKI, G. BIAŁKOWSKI *Cząstki elementarne*, Warszawa 1971.

FIZYKA JĄDRA ATOMOWEGO

Jądra atomowe i ich wzbudzenia · Siły jądrowe · Modele jądrowe · Rozpady jąder atomowych · Reakcje jądrowe · Jądra atomowe w stanach ekstremalnych · Fizyka ciężkich jonów · Spektroskopia jądrowa · Fizyka jądrowa wielkich energii · Hiperjądra · Energia jądrowa · Energia termojądrowa · Radioizotopy ·

Jądra atomowe i ich wzbudzenia

Piotr Decowski

Jądro atomowe składa się z dodatnio naładowanych elektrycznie protonów i obojętnych elektrycznie neutronów, nazywanych wspólnie nukleonami. Liczba protonów Z w jądrze nosi nazwę liczby atomowej lub liczby porządkowej jądra, a liczba wszystkich nukleonów $A = Z+N$, gdzie N jest liczbą neutronów, nosi nazwę liczby masowej jądra. Jądro o liczbach Z, N i A oznacza się krótko ${}^{A}_{Z}X_{N}$, gdzie za X wpisuje się symbol chemiczny odpowiadający liczbie atomowej Z. Na przykład jądro berylu ($Z = 4$) o liczbie masowej $A = 9$ oznacza się ${}^{9}_{4}Be_{5}$. Często w tym oznaczeniu pomija się liczbę neutronów $N=A-Z$.

izotopy, izotony, izobary

Jądra o tym samym Z, a różnych A, nazywamy izotopami. Na przykład ${}^{16}_{8}O_{8}$ i ${}^{18}_{8}O_{10}$ są dwoma izotopami tlenu. Jądra o tym samym N, a różnych A, nazywamy izotonami, np. ${}^{18}_{8}O_{10}$ i ${}^{20}_{10}Ne_{10}$. Jądra o tym samym A, a różnych Z, nazywamy izobarami, np. ${}^{40}_{18}Ar_{22}$ i ${}^{40}_{20}Ca_{20}$. [Ponieważ w normalnych warunkach jądra występują nie same, ale z powłokami elektronowymi, często nazywamy izotopami, izotonami czy izobarami nie grupy jąder o omówionych własnościach, lecz odpowiadające im grupy atomów.]

Liczby Z i N lub Z i A określające skład jądra nie charakteryzują go jeszcze w pełni. Może ono bowiem znajdować się w różnych stanach. Każdy stan określony jest przez pewien zespół cech, do którego należą: energia E, całkowity moment pędu (spin, kręt) I i parzystość P. Parzystość jest cechą czysto kwantową (jest własnością funkcji falowej opisującej stan jądra); może być dodatnia (+) lub ujemna (−). Wśród wszystkich stanów jądra wyróżniony jest stan o najmniejszej energii E_0. Nazywamy go stanem podstawowym. Wszystkie stany pozostałe są stanami wzbudzonymi. Energia wzbudzenia jest różnicą energii stanu wzbudzonego i stanu podstawowego jądra.

energia wzbudzenia

Własności jąder w stanie podstawowym

siły jądrowe

Jądro istnieje dzięki działaniu jądrowych sił wzajemnego przyciągania nukleonów (→ Siły jądrowe). W porównaniu z siłami, z jakimi spotykamy się w świecie makroskopowym, są to siły olbrzymie. Dwa nukleony, mikroskopijne obiekty o rozmiarach rzędu 10^{-13} cm, mogą się przyciągać z siłą równą ciężarowi masy około 10 ton! Siły te jednak działają tylko wtedy, gdy odległość między nukleonami jest bardzo mała — rzędu 10^{-13} cm, a już w odległości paru femtometrów zanikają. Zatem „efektywna" siła działająca między nukleonami w jądrze jest kilkadziesiąt razy mniejsza, ale w porównaniu z naszymi wyobrażeniami jeszcze olbrzymia. Krótki zasięg sił jądrowych powoduje, że nukleon w jądrze „czuje" obecność tylko otaczających go najbliższych paru sąsiadów, niezależnie od liczby pozostałych nukleonów. Dlatego też gęstość materii jądrowej jest stała, nie zależy od liczby nukleonów w jądrze i wynosi około $1{,}7 \cdot 10^{38}$ nukleonów/cm^3. Nasuwa się analogia do kropli nieściśliwej cieczy, w której siły van der Waalsa działają tylko między najbliższymi cząsteczkami. Z tych względów objętość jądra jest proporcjonalna do liczby nukleonów (a więc do liczby masowej A danego izotopu).

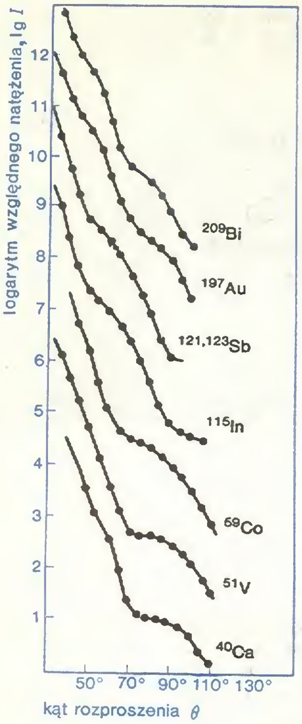
rozkład ładunku

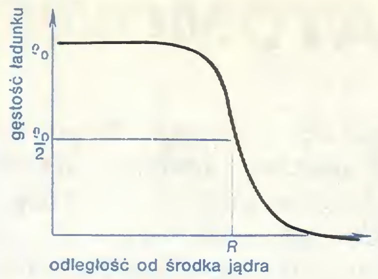
Wyznaczenie rozmiarów tak małego obiektu jak jądro atomowe nie jest łatwe. Większość metod opiera się na badaniu przestrzennego rozkładu ładunku jądra. Informacji o rozkładzie ładunku dostarcza na przykład obserwacja strumienia elektronów mających bardzo dużą energię, rozpraszanych w różnych kierunkach przez badaną próbkę. Długość fali elektronów o energii rzędu kilkuset MeV jest bliska rozmiarom jądra i obraz rozpraszania ma charakter dyfrakcyjny, podobnie jak obraz dawany przez wiązkę światła napotykającą na swej drodze przeszkodę o rozmiarach porównywalnych z długością fali świetlnej (rys. 1). Analizując elektronowy obraz dyfrakcyjny można odtworzyć kształt rozkładu ładunku, a więc i rozkładu protonów w jądrze. Z innych ważniejszych sposobów pomiaru rozkładu ładunku można wymienić pomiar mas jąder o tej samej liczbie masowej A, lecz różniących się liczbą neutronów i protonów o 1: ${}^{A}_{Z}X_{N}$ i ${}^{A}_{Z+1}Y_{N-1}$ (są to tzw. jądra zwierciadlane). Główną przyczyną występowania różnicy mas jest energia kulombowska dodatkowego protonu w polu elektrostatycznym pozostałych protonów zależna od ich rozkładu przestrzennego. O rozkładzie ładunku można również wnioskować z badania widma promieniowania towarzyszącego wychwytowi na orbitę atomową „ciężkiego elektronu" — mionu. Orbita mionu, dzięki jego dużej masie, może znajdować się w bezpośrednim sąsiedztwie jądra, w związku z czym rozkład ładunku może mieć istotny wpływ na jej parametry. Zgodnie z przewidywaniami gęstość ładunku w dużym obszarze jądra jest prawie stała, zaś w pobliżu powierzchni jądra szybko spada do zera. Zależność gęstości od odległości od środka masy opisuje tzw. rozkład Fermiego (rys. 2):

promień ładunkowy

$$\varrho(r) = \frac{\varrho_0}{1+e^{(r-R)/a}},$$

w którym promień jądra, zwany promieniem ładunkowym, wynosi $R = 1{,}07\ A^{1/3}$ fm, natomiast a (tzw. rozmycie powierzchni jądra) równa się 0,55 fm.





Rys. 2. Zależność gęstości ładunku od odległości od środka jądra

Rys. 1. Rozkłady natężenia strumienia elektronów o energii 185 MeV rozpraszanych w różnych kierunkach przez badane tarcze. Występują w nich periodyczne zmiany natężenia przypominające zmiany natężenia fali świetlnej ugiętej przy przejściu w pobliżu przegrody o małych rozmiarach

Użycie elektronów do sondowania jądra ma pewną istotną zaletę. Oddziałują one ze składnikami jądra tylko poprzez siły elektromagnetyczne — nie „czują" sił jądrowych. Dlatego też opisują rzeczywisty rozkład ładunku. Inaczej jest, gdy do badania rozmiarów jąder stosuje się rozpraszanie wiązek nukleonów, cząstek α czy też innych cząstek oddziałujących z nukleonami również siłami jądrowymi. Uzyskany wówczas rozkład gęstości materii jądrowej jest podobny do rozkładu gęstości ładunku, z tym, że wyznaczony promień jądra, noszący nazwę promienia potencjałowego, jest większy: $R=1{,}21\ A^{1/3}$fm. Jest to zrozumiałe, gdyż pojawienie się sił jądrowych powoduje, że cząstka bombardująca zaczyna „odczuwać" jądro już w pewnej odległości od jego powierzchni.

promień potencjałowy

Masa jądra (wyraża się ją ze względu na małą wartość, w jednostkach masy atomowej u, 1u jest równa $^1/_{12}$ masy jądra izotopu węgla ^{12}C, czyli $1{,}660531\cdot 10^{-27}$ kg) jest mniejsza od sumy mas tworzących je nukleonów. Spowodowane to jest oddziaływaniem nukleonów w jądrze. Gdyby zgrupować swobodne nukleony, to na skutek ich wzajemnego przyciągania się wydzieliłaby się, kosztem ich łącznej masy, pewna energia, zwana energią wiązania jądra; taka energia potrzebna jest do rozbicia jądra na pojedyncze składniki. Energię wiązania można obliczyć ze wzoru:

energia wiązania

$$B(A, Z) = Zm_{\mathrm{H}}+(A-Z)m_n-M(A, Z),$$

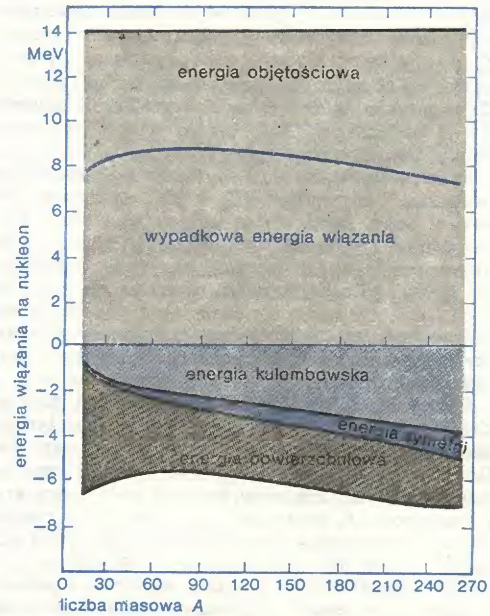
w którym A i Z oznaczają odpowiednio liczbę masową i porządkową jądra, $M(A, Z)$ — masę atomu pierwiastka (tzn. masę jądra wraz z Z elektronami), m_{H} — masę atomu wodoru, m_n — masę neutronu.

Zgodnie z zastosowaniem do jądra analogii kropli cieczy, średnia energia wiązania przypadająca na jeden nukleon, $B(A, Z)/A$, winna być w przybliżeniu stała dla wszystkich jąder. Warunkować ją powinna liczba nukleonów otaczających dany nukleon, zależna jedynie od gęstości materii jądrowej. W jądrze nie można jednak pominąć wpływu dodatkowych czynników, z których najważniejszy jest wpływ nukleonów znajdujących się na powierzchni jądra. Są one słabiej związane, podobnie jak cząsteczki znajdujące się na powierzchni cieczy (co prowadzi do powstania efektu napięcia powierzchniowego), z tym że w jądrze, w związku z małą liczbą nukleonów, efekty powierzchniowe są znacznie bardziej istotne.

Wpływ różnych czynników na energię wiązania można ująć ogólnie (→ Modele jądrowe) za pomocą tzw. wzoru Weizsäckera:

wzór Weizsäckera

$B(A, Z)=a_1A-a_2A^{2/3}-a_3(A-2Z)^2/A-a_4Z^2/A^{1/3}+\delta$.
Człon drugi we wzorze, proporcjonalny do powierzchni jądra ($R^2 \sim A^{2/3}$), uwzględnia zmniejszanie energii wiązania przez efekty powierzchniowe. Człon trzeci, proporcjonalny do kwadratu różnicy liczby neutronów i protonów, mówi o tendencji do utrzymania równej ilości neutronów i protonów w jądrze (jest to tzw. człon symetrii wynikający z braku wyróżnienia jakiegokolwiek z dwu rodzajów nukleonów — nadmiar protonów lub neutronów zakłóca tę symetrię i jest energetycznie niekorzystny). Człon czwarty uwzględnia fakt istnienia odpychania elektrostatycznego protonów (energia potencjalna naładowanej kuli jest proporcjonalna do Z^2/R). Wreszcie człon piąty uwzględnia wynikające z natury sił jądrowych zjawisko dążenia nukleonów do łączenia się w pary — jądra o parzystej liczbie neutronów i protonów są bardziej trwałe (jest to efekt podobny do efektu nadprzewodnictwa gazu elektronowego niektórych związków w niskich temperaturach). Współczynniki a_1, a_2, a_3, a_4 oraz człon δ wyznaczone doświadczalnie na podstawie zmierzonych mas jąder (głównie metodami spektrografii masowej, przez obserwację torów jonów poruszających się ze znaną prędkością w polu magnetycznym o znanym natężeniu) opisują z dużą dokładnością energię wiązania i masy jąder o dowolnym A i Z. Wpływ poszczególnych członów na energię wiązania ilustruje rys. 3.



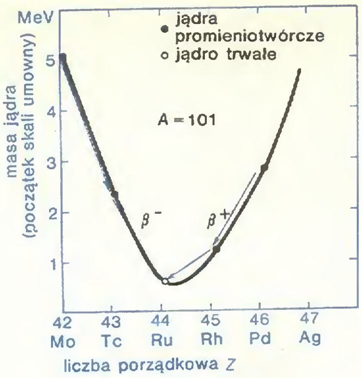
Rys. 3. Wkład do średniej energii wiązania nukleonu różnych członów wzoru Weizsäckera w zależności od liczby masowej jądra. Od energii objętościowej odejmuje się energia powierzchniowa, kulombowska i energia symetrii (człon związany z łączeniem się nukleonów w pary pominięto) dając w rezultacie wartość wynoszącą dla większości jąder około 8 MeV

Wzór Weizsäckera mówi, że energia wiązania, a więc i masa jąder o tym samym A (izobarów), zależy parabolicznie od liczby porządkowej Z (rys. 4). Najbardziej związanymi, a tym samym trwałymi izobarami, są jądra o liczbie porządkowej w pobliżu minimum paraboli mas. We współrzędnych $N=A-Z$ (liczba neutronów) i Z układają się one wzdłuż tzw. ścieżki trwałości (rys. 5). Jądra poza nią są nietrwałe i ulegają przemianom promieniotwórczym prowadzącym do tworzenia jąder bardziej trwałych.

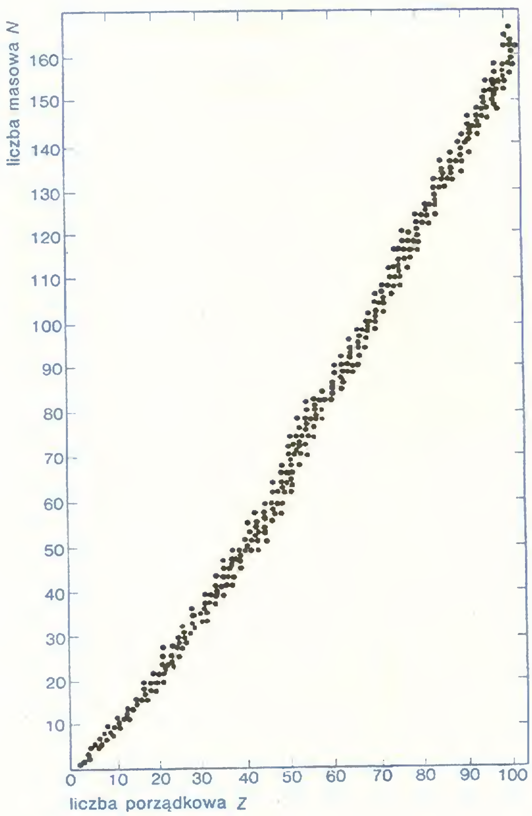
parabola mas

ścieżka trwałości

Spin jądra mierzony w jednostkach $\hbar$ ($\hbar=h/2\pi$, h — stała Plancka) jest zawsze liczbą całkowitą dla



Rys. 4. Parabola mas



Rys. 5. Ścieżka trwałości

spin jądra

jąder o parzystej liczbie nukleonów i liczbą połówkową dla jąder o A nieparzystym. Pochodzi on od wektorowego złożenia się momentów pędu ruchu nukleonów w jądrze oraz ich własnych spinów równych $\frac{1}{2}\hbar$. Wszystkie niewzbudzone jądra o parzystej liczbie zarówno protonów, jak i neutronów mają spin równy zeru. Jest to konsekwencją wspomnianej już tendencji do łączenia się nukleonów jednego rodzaju w pary. Całkowity moment pędu takiej pary znika, gdyż najchętniej łączą się w nią nukleony o przeciwnie skierowanych momentach pędu. Wartości spinów wszystkich jąder w stanie podstawowym, o nieparzystej liczbie nukleonów jednego lub obu rodzajów, wyjaśniono na podstawie obecnych danych o budowie jądra. Jest to istotny sukces teorii jądra atomowego.

moment magnetyczny

Jądro atomowe ma, obok ładunku elektrycznego Ze, pewne własności magnetyczne, opisywane przez dipolowy moment magnetyczny. Wiąże się on ze spinem jądra, lecz w odróżnieniu od niego nie jest wielkością skwantowaną — może przyjmować dowolne wartości. Moment magnetyczny wyraża się za pomocą jednostki zwanej magnetonem jądrowym:

$$\mu_n = \frac{|e|\hbar}{2M_p c} = 0{,}50509 \cdot 10^{-23}\,\mathrm{erg}\cdot\mathrm{Gs}^{-1} = 0{,}50509 \times \times 10^{-26}\ \mathrm{J}\cdot\mathrm{T}^{-1}$$

(M_p — masa protonu, c — prędkość światła). Maksymalna wartość rzutu momentu magnetycznego na wybrany kierunek w przestrzeni wiąże się z maksymalną wartością I rzutu spinu na ten kierunek:

$$\mu_I = I g_I \mu_n = \gamma_I \hbar I.$$

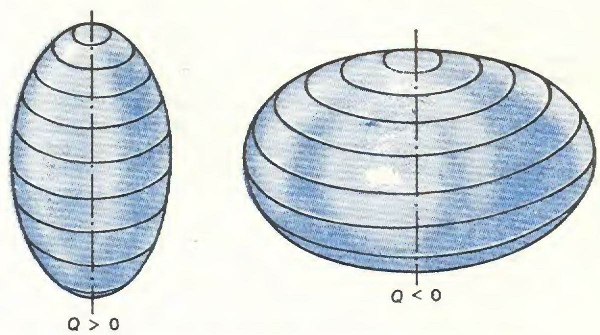
Wprowadzone we wzorze wielkości g_I i γ_I noszą nazwę jądrowego czynnika g oraz stosunku giromagnetycznego. Moment magnetyczny wyznacza się doświadczalnie, badając rozszczepienie linii w widmach atomowych spowodowane jego oddziaływaniem z polem magnetycznym powłok elektronowych atomu. Wyznaczone eksperymentalnie wielkości g lub γ są czułym sprawdzianem poprawności różnych teoretycznych modeli budowy jądra atomowego.

moment elektryczny

Pole elektryczne wokół jądra, będącego rozciągłym naładowanym obiektem, zależy od jego kształtu. Zależność tę opisuje się rozkładając potencjał elektrostatyczny na szereg multipolowy, w którym o wielkości kolejnych członów, zależnych od $(1/r)$ w coraz wyższych potęgach, decydują momenty elektryczne. Elektryczny moment monopolowy równa się ładunkowi jądra. Elektryczny moment dipolowy jest równy zeru ze względu na równomierne rozłożenie ładunku w jądrze. Moment kwadrupolowy pojawia się wtedy, gdy kształt jądra odbiega od kształtu kulistego. Wyższe momenty są z reguły bardzo małe i zwykle można je pominąć.

kształty jąder

Pomiar momentu kwadrupolowego jądra (np. przez obserwację wpływu pola elektrycznego jądra na ruch elektronów poruszających się po orbitach



Rys. 6. Znak i wartość elektrycznego momentu kwadrupolowego dostarczają informacji o kształcie jądra

atomowych — moment kwadrupolowy wpływa na nadsubtelną strukturę widm optycznych) najczęściej jest źródłem informacji o kształcie jądra (rys. 6). Okazuje się, że istnieje wiele jąder silnie zdeformowanych w stanie podstawowym, o dużej wartości momentu kwadrupolowego, których kształt przypomina w większości wypadków wydłużone cygaro. Skupiają się one w pewnych obszarach tablicy pierwiastków (np. zdeformowane są jądra z obszaru pierwiastków ziem rzadkich oraz z obszaru aktynowców), poprzedzielanych obszarami silniej związanych jąder sferycznych skupionych wokół jąder magicznych (tj. jąder o tzw. magicznych liczbach protonów i neutronów równych jednej z liczb 2, 8, 20, 50, 82, 126 — cechują się one szczególnie silnym wiązaniem i trwałością). Kształt jądra wpływa na widmo jego wzbudzeń. Jądra zdeformowane mogą np. szczególnie łatwo rotować wokół osi prostopadłej do osi ich symetrii. Prowadzi to do pojawienia się w ich energiach wzbudzenia regularnych sekwencji zwanych pasmami rotacyjnymi.

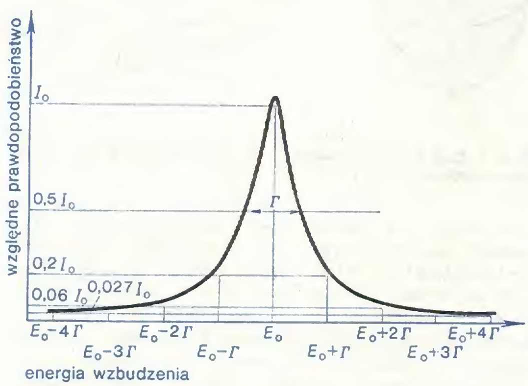
Wzbudzenia jąder atomowych

Zgodnie z prawami mechaniki kwantowej jądro atomowe może znajdować się tylko w pewnych stanach o określonych cechach kwantowych, m.in. spinie, parzystości, energii. Stwierdzenie powyższe jest ścisłe tylko dla energii wzbudzenia jądra, która jest za mała, aby mogła nastąpić emisja jakiegokolwiek nukleonu z jądra, czyli dla energii mniejszych od energii wiązania najsłabiej związanego nukleonu (zwanej energią wiązania ostatniego nukleonu). Dla energii większych każda energia, spin i parzystość są dopuszczalne. Jest to zrozumiałe, gdyż każdy stan, z którego może nastąpić emisja cząstki o pewnej energii kinetycznej, można utworzyć przez wprowadzenie cząstki o takiej samej energii do jądra. Cząstce możemy jednak nadać dowolną energię kinetyczną. W ten sposób możemy tworzyć stany jądra o dowolnej energii. Okazuje się jednak, że tylko przy niektórych wyróżnionych wartościach energii jądro tworzy układ trwający pewien, odpowiednio długi czas (rozumie się przez to czas co najmniej kilkakrotnie dłuższy od tzw. charakterystycznego czasu jądrowego potrzebnego na przejście najszybszego nukleonu w jądrze przez odcinek równy średnicy jądra, wynosi on w zależności od jądra i jego energii wzbudzenia 10^{-22}–10^{-23} s). W odróżnieniu od stanów leżących poniżej progu na emisję nukleonów, zwanych stanami związanymi, wzbudzone stany leżące w obszarze continuum energii noszą nazwę stanów rezonansowych. Istnienie bariery kulombowskiej na powierzchni jądra może być przyczyną znacznego utrudnienia emisji cząstek naładowanych ze stanów niezwiązanych o małej energii. Gdy emisja neutronów będących cząstkami obojętnymi jest niemożliwa (gdy np energia wiązania neutronu jest większa niż energia wzbudzenia), stany takie wykazują cechy stanów związanych i dlatego nazwano je „kwazizwiązanymi”.

stany związane i rezonansowe

Wszystkie stany wzbudzone jąder, zarówno związane jak i rezonansowe, istnieją tylko przez pewien określony czas, którego średnia wartość zwana jest czasem życia stanu. Stany związane rozpadają się przez wysłanie promieniowania γ lub niekiedy ulegają rozpadowi β (→ Rozpady jąder atomowych). Stany rezonansowe rozpadają się najczęściej przez emisję nukleonu lub grupy nukleonów.

czas życia stanu



Rys. 7. Krzywa względnego prawdopodobieństwa wzbudzenia stanu o szerokości Γ przy różnych wartościach energii

Prawdopodobieństwo wzbudzenia stanu rezonansowego jest funkcją energii wzbudzenia E wyrażoną wzorem (rys. 7)

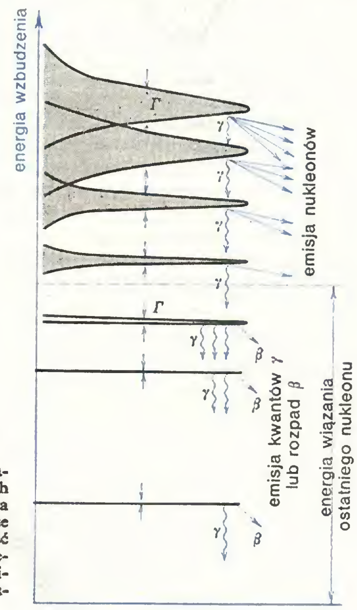
$$P(E) \sim \frac{1}{(E-E_0)^2+\Gamma^2/4}.$$

Występująca we wzorze wielkość Γ, równa zakresowi energii, dla którego gęstość prawdopodobieństwa wzbudzenia stanu jest większa od połowy wartości maksymalnej (przy $E = E_0$), nazywa się szerokością energetyczną stanu. Stanowi wzbudzonemu nie odpowiada więc dokładnie jedna wartość energii. Pod pojęciem „energia stanu” rozumie się odpowiadającą mu wartość E_0. Szerokość energetyczna Γ wiąże się z jego czasem życia τ przez zasadę nieoznaczoności Heisenberga:

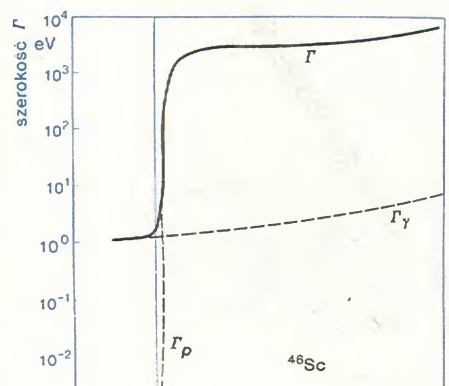
szerokość energetyczna stanu

$$\Gamma\tau \approx \hbar.$$

Stany rezonansowe mają z reguły znacznie większą szerokość energetyczną od stanów związanych (krócej żyją; rys. 8 i 9). Jest to konsekwencją faktu, że na ogół emisja nukleonów następująca pod wpływem niezwykle silnych oddziaływań jądrowych jest znacznie bardziej prawdopodobna niż emisja promieniowania elektromagnetycznego lub emisja elektronu i neutrina w rozpadzie β powodowane znacznie słabszymi oddziaływaniami. Im wyższa jest energia stanu, tym większa jest jego szerokość, gdyż rośnie liczba możliwych sposobów emisji nukleonów (np. mogą



Rys. 8. Szerokość stanów niezwiązanych jest znacznie większa niż związanych ze względu na możliwość rozpadu stanów niezwiązanych przez emisję nukleonów (rysunek schematyczny)

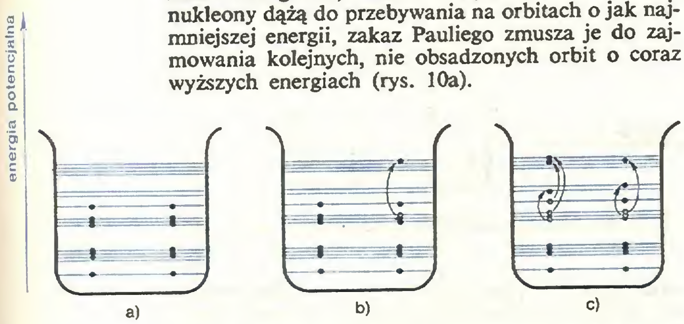


Rys. 9. Zależność od energii wzbudzenia szerokości Γ poziomów o spinie 0 w jądrze ^{46}Sc. Linie przerywane oznaczają szerokości poziomów w przypadku gdyby był możliwy ich rozpad tylko przez emisję kwantów γ (Γ_γ) lub nukleonów (Γ_p). Poniżej energii wiązania ostatniego neutronu emisja nukleonów jest niemożliwa. Powyżej energii wiązania Γ_p decyduje o szerokości stanu

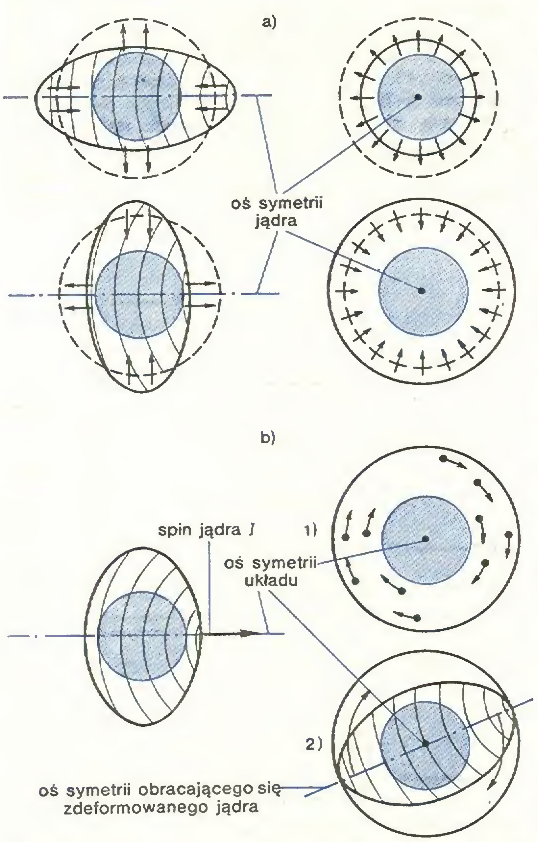
być wysyłane z coraz bardziej różnymi wartościami energii), co zwiększa prawdopodobieństwo rozpadu stanu.

wzbudzenie jądra

Aby zrozumieć proces wzbudzenia jądra, wyobraźmy je sobie jako twór składający się z nukleonów będących w ciągłym ruchu. Każdy nukleon, podobnie jak elektrony na orbitach atomowych, porusza się z pewną określoną energią kinetyczną po swej orbicie. Nukleony są fermionami, czyli cząstkami o spinie połówkowym, muszą więc podporządkowywać się zakazowi Pauliego mówiącemu, że dwie identyczne cząstki nie mogą przebywać w identycznych stanach. Dlatego też, mimo że w jądrze niewzbudzonym nukleony dążą do przebywania na orbitach o jak najmniejszej energii, zakaz Pauliego zmusza je do zajmowania kolejnych, nie obsadzonych orbit o coraz wyższych energiach (rys. 10a).



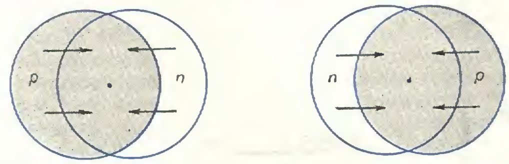
Rys. 10. Schematyczny rysunek przedstawiający różne rodzaje wzbudzeń jądra atomowego: a) stan podstawowy — nukleony zajmują orbity położone najniżej w skali energetycznej, b) wzbudzenie jednocząstkowe — jeden z nukleonów jest przeniesiony na orbitę o większej energii, c) wzbudzenie wielocząstkowe — wiele nukleonów jest przeniesionych na orbity o wyższych energiach



Rys. 11. Przykłady wbudzeń kolektywnych: a) przykład wibracji powierzchni jądrowej (tzw. wibracje β) wokół kształtu sferycznego (linia przerywana), b) rotacja jądra o spinie I; 1) można ją opisać poprzez ruch wielu nukleonów, 2) znacznie wygodniej jest je opisać za pomocą obrotu jądra zdeformowanego, o kształcie cygara (lub dysku), wokół osi prostopadłej do osi swojej symetrii. Nukleony w zacieniowanych obszarach jądra nie biorą udziału we wzbudzeniu

wzbudzenia jednocząstkowe i kolektywne

Wzbudzenie jądra atomowego polega na zmianie ruchu pewnej liczby nukleonów. Na przykład gdy cała energia wzbudzenia jest skupiona na jednym nukleonie — tylko on jest przeniesiony na orbitę o większej energii — mówimy, że wzbudzenie ma charakter jednocząstkowy (rys. 10b). Jest to najprostsza konfiguracja stanu wzbudzonego. We wzbudzeniu może brać również udział wiele nukleonów (rys. 10c). Jeśli wzbudzenie przenosi się w sposób skorelowany z jednych nukleonów na inne, możemy mieć do czynienia



Rys. 12. Gigantyczny elektryczny rezonans dipolowy. Protony i neutrony drgają wokół środka masy jądra

rezonans gigantyczny

z wzbudzeniem kolektywnym, które objawia się np. w postaci wibracji powierzchni jądrowej lub rotacji całego jądra (rys. 11). We wzbudzeniach mogą brać udział w sposób spójny niemal wszystkie nukleony jądra — prowadzi to do wzbudzeń typu rezonansów gigantycznych (rys. 12).

Modele i własności jąder

Do opsiu stanów stosuje się różne metody zależnie od typu wzbudzenia. Wzbudzenia o prostej konfiguracji najłatwiej przedstawić opisując indywidualne stany biorących w nim udział nukleonów. Dla wzbudzeń kolektywnych taki opis jest już zbyt skomplikowany, wymaga uwzględnienia dużej liczby stopni swobody układu wielu nukleonów biorących udział we wzbudzeniu. Sytuacja znacznie się upraszcza po wprowadzeniu współrzędnych kolektywnych związanych z kształtem i orientacją w przestrzeni jądra jako całości i rozpatrywaniu wzbudzeń niewielu stopni swobody związanych z tymi współrzędnymi. Przykładem mogą tu być wzbudzenia rotacyjne opisywane bardzo prosto przez wprowadzenie obrotu odpowiednio zdeformowanego jądra, natomiast bardzo skomplikowane w opisie wielocząstkowym.

współrzędne kolektywne

Aby opisać własności stanów wzbudzonych należy przyjąć pewne założenia dotyczące oddziaływań, jakim ulegają nukleony, czy też jakie wpływają na kolektywne stopnie swobody. W tym celu stosuje się modele budowy jądra atomowego (→Modele jądrowe). Proste konfiguracje opisuje się za pomocą modelu powłokowego, operującego wzbudzeniami i oddziaływaniami poszczególnych nukleonów. Bardziej skomplikowane konfiguracje opisuje się za pomocą odpowiednich modeli kolektywnych.

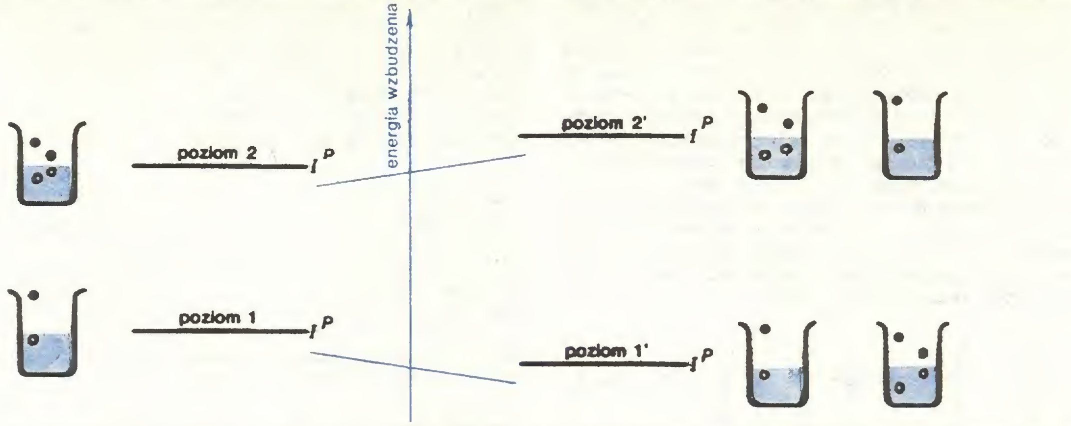
jądrowy poziom energetyczny

Na ogół konfiguracja stanu wzbudzonego opisywana za pomocą danego modelu nie jest konfiguracją „czystą", tzn. będącą określonym wzbudzeniem nukleonu czy grupy nukleonów, czy też związaną ze wzbudzeniem określonego kolektywnego stopnia swobody. Wszystkie stany różniące się tylko magnetyczną liczbą kwantową — rzutem spinu na wybraną oś kwantyzacji określa się pojęciem „jądrowego poziomu energetycznego". (Gdy na jądro nie działają czynniki zewnętrzne, np. pole magnetyczne, energia stanu nie zależy od kierunku ustawienia spinu w przestrzeni — wszystkie stany należące do danego poziomu mają wówczas tę samą energię). Sąsiedztwo dwu poziomów o innej konfiguracji a tym samym spinie i parzystości powoduje, że następuje „zmieszanie" konfiguracji, tzn. jeden i drugi stan stają się mieszaniną obu konfiguracji różniących się np. liczbą wzbudzonych nukleonów (rys. 13; mieszające się stany muszą mieć ten sam spin i parzystość, gdyż oddziaływania jądrowe zachowują te wielkości). Jest to

mieszanie konfiguracji

znowu jedna z charakterystycznych kwantowych cech mikroświata: stany o identycznych cechach kwantowych są w zasadzie nierozróżnialne. Stany o nieco

Przy dalszym dodawaniu nukleonów do jądra podwójnie magicznego wzbudzenia stają się coraz bardziej kolektywne. Oddziaływania między nukleonami



Rys. 13. Mieszanie konfiguracji. Pod wpływem oddziaływań międzynukleonowych konfiguracje blisko położonych poziomów o tym samym spinie I i parzystości P ulegają zmieszaniu

różnych cechach kwantowych, np. różnej energii a tym samym spinie i parzystości starają się jak najbardziej upodobnić do siebie. Im mniejsza jest różnica energii między nimi, tym większy jest stopień zmieszania. Mieszanie się konfiguracji można interpretować w ten sposób, że jądro wzbudzone do danego stanu pod wpływem oddziaływań międzynukleonowych wielokrotnie przechodzi od jednej konfiguracji do drugiej. Zmieszanie zmienia energię stanów — poziomy odsuwają się od siebie. Oczywiście, może nastąpić wzajemne mieszanie wielu sąsiadujących poziomów mających podobne liczby kwantowe. W takiej sytuacji funkcja falowa danego poziomu ma wiele składowych odpowiadających różnym konfiguracjom. Poziomy jądrowe o niewielkich energiach wzbudzenia są stanami stosunkowo „czystymi" — w ramach odpowiednio dobranego modelu dają się opisać za pomocą funkcji falowej zawierającej niewielką liczbę składowych o łatwej interpretacji fizycznej. Wynika to stąd, że przy małych energiach wzbudzenia poziomy o podobnych liczbach kwantowych pojawiają się rzadko.

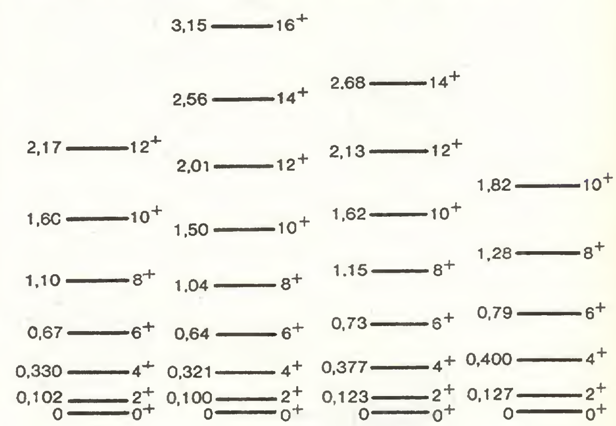
jądra magiczne

Jądra, w których liczba protonów i neutronów równa jest jednej z liczb magicznych: 2, 8, 20, 50, 82, 126, są szczególnie trwałe (np. $^{16}_{8}O$, $^{40}_{20}Ca$, $^{208}_{82}Pb$) i nazywane są jądrami magicznymi. Ze względu na dużą trwałość jąder magicznych do ich wzbudzenia potrzebna jest znacznie większa energia niż do wzbudzenia jąder sąsiednich (np. energia pierwszych stanów wzbudzonych w jądrach $^{16}_{8}O$ i $^{18}_{8}O$ wynoszą odpowiednio: ok. 6 MeV i ok. 2 MeV, w jądrach $^{206}_{82}Pb$ i $^{208}_{82}Pb$ — odpowiednio 1 MeV i 2,6 MeV). Dlatego też najniżej leżące poziomy wzbudzone jąder bliskich jądrom podwójnie magicznym są wzbudzeniami stosunkowo prostymi. Bierze w nich udział niewielka liczba nukleonów znajdujących się poza rdzeniem — jądrem podwójnie magicznym. Na przykład wzbudzenia jądra różniącego się od jądra podwójnie magicznego liczbą nukleonów o jeden odzwierciedlają sekwencję kolejnych poziomów jednocząstkowych w danym obszarze liczb masowych. W sekwencji tej nie widać jakiejkolwiek regularności tak w energii, jak i w spinach kolejnych poziomów wzbudzonych. Stąd też widma wzbudzeń jąder bliskich jądrom magicznym również nie wykazują specjalnych cech regularności. Mogą one być zasadniczo różne w sąsiadujących jądrach, gdyż przy niewielkiej liczbie wzbudzonych nukleonów dodanie jednego nukleonu zmienia w sposób istotny sytuację. Przy wyższych energiach wzbudzenia konfiguracje przestają być tak proste, ponieważ możliwe stają się również wzbudzenia rdzenia, których konfiguracje mogą się mieszać z konfiguracjami pozostałych nukleonów.

wzbudzenia jąder bliskich magicznym

znajdującymi się poza rdzeniem prowadzą do powstania między nimi korelacji (również przestrzennych, np. może powstać trwała deformacja jądra). Wzbudzenia są teraz wzbudzeniami całych zespołów nukleonów. Ich cechami charakterystycznymi są małe wartości energii, występowanie regularności w sekwencji energii i spinów oraz podobieństwo u wielu jąder (rys. 14). Wzbudzenia mają charakter najbardziej kolektywny w jądrach, w których liczby protonów i neutronów Z i N leżą w przybliżeniu w połowie między dwoma kolejnymi liczbami magicznymi (np. w jądrach izotopów pierwiastków ziem rzadkich lub aktynowców). Najbardziej wyróżniającą się cechą wzbudzeń tych jąder są pasma rotacyjne wynikające z istnienia obrotów zdeformowanych jąder wokół osi prostopadłej do osi ich symetrii. Dalsze zwiększanie liczby nukleonów prowadzi do tworzenia jąder coraz bliższych następnemu jądru magicznemu.

pasma rotacyjne

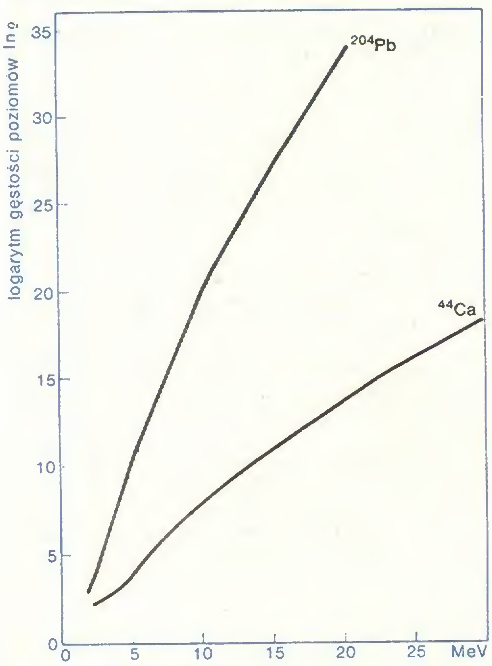


Rys. 14. Widma wzbudzeń poziomów rotacyjnych jąder zdeformowanych. Liczby podają energię wzbudzenia poziomu (w MeV) oraz jego spin, znak (+) dotyczy parzystości. Dla różnych jąder widma te wyglądają bardzo podobnie

Wzbudzenia tych jąder stopniowo tracą charakter kolektywny.

duże energie wzbudzenia

Przy dużych energiach wzbudzenia (od 20 MeV dla jąder lekkich, od 10 MeV dla jąder ciężkich) gęstość poziomów staje się bardzo wielka, bowiem podobną energię wzbudzenia można osiągnąć różnymi sposobami, np. przez wzbudzenie bardzo różnej liczby nukleonów. Rośnie szerokość energetyczna poziomów. W rezultacie poziomy zaczynają się pokrywać. Duża gęstość poziomów powoduje, że wszystkie stany w danym obszarze energii wzbudzenia są bardzo silnie zmieszane — ich funkcje falowe mogą zawierać nawet miliony składowych! W tej sytuacji do opisu własności poziomów stosuje się na ogół metody statystyczne. Między innymi zakłada się, że pojawienie się każdej z tych wielu konfiguracji jest równie prawdopodobne. Obliczone przy tych założeniach prawdopodobieństwo rozpadu stanu, np. przez emisję nukleonu, jest uwarunkowane wyłącznie prawdopodobieństwem przejścia nukleonu przez skok potencjału na powierzchni jądra i nie wymaga znajomości struktury stanu rozpadającego się i stanu, w jakim jądro pozostaje po rozpadzie (tzn. nie trzeba rozpatrywać możliwości tworzenia takiej konfiguracji stanu wzbudzonego, w której nukleon jest wzbudzony do energii odpowiadającej energii jego emisji). W większości wypadków konfrontacja z eksperymentem potwierdza słuszność przyjętych założeń.



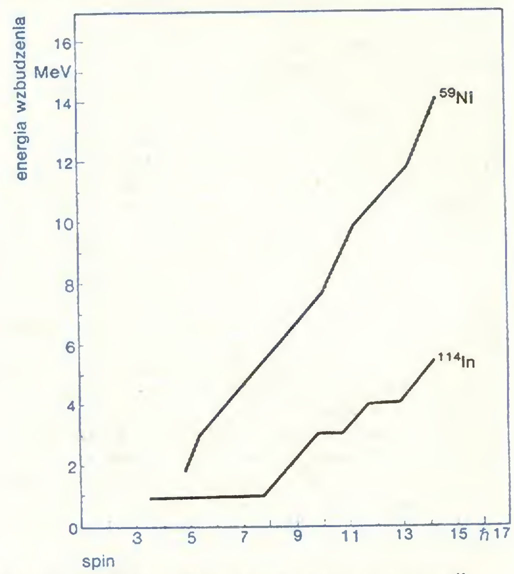
Rys. 15. Zależność gęstości poziomów od energii wzbudzenia dla dwóch różnych jąder

gęstość poziomów a energia wzbudzenia

Gęstość poziomów bardzo silnie zależy od energii wzbudzenia. Zależność ta jest w przybliżeniu wykładnicza — zwiększenie energii wzbudzenia o 2–3 MeV powoduje już ok. 10-krotny wzrost gęstości! (rys. 15). Do szacowania gęstości poziomów stosuje się metody statystyczne z termodynamiki. Jądro traktuje się wówczas jako gaz fermionów w równowadze termodynamicznej o pewnej temperaturze odpowiadającej energii wzbudzenia. Istnienie skwantowanych, stosunkowo rzadko rozmieszczonych stanów energetycznych nukleonów oraz zakazu Pauliego powoduje, że spośród wszystkich nukleonów tylko te mogą brać udział we wzbudzeniach, które są najbliżej nie obsadzonych stanów nukleonowych (najbliżej tzw. powierzchni Fermiego). Średnia liczba wzbudzonych nukleonów jest liniową funkcją temperatury. Zatem, w odróżnieniu od gazu klasycznego, zależność całkowitej energii od temperatury nie jest liniowa lecz kwadratowa. Gwałtowny wzrost gęstości poziomów w zależności od energii wzbudzenia jest wynikiem nie tylko zwiększenia się ilości możliwych sposobów rozkładu energii wśród nukleonów biorących udział we wzbudzeniu, lecz również wynikiem wzrostu średniej liczby wzbudzonych nukleonów.

gęstość poziomów a spin

Przy ustalonej energii wzbudzenia gęstość poziomów zależy od ich spinów. Jest to zrozumiałe, gdy się weźmie pod uwagę fakt, że aby utworzyć stan o wysokim spinie, nukleony muszą obsadzać poziomy w taki sposób, by spiny możliwie licznej grupy nukleonów układały się w tym samym kierunku dając odpowiednio dużą wypadkową wartość. Ogranicza to swobodę rozdziału energii na poszczególne nukleony (a więc swobodę obsadzania różnych orbit) i obniża gęstość stanów. Dla każdej energii wzbudzenia istnieje pewna optymalna wartość spinu, przy której gęstość jest maksymalna, oraz pewna maksymalna wartość spinu. Osiąga się ją tylko przy szczególnej, jedynej konfiguracji, w której spiny nukleonów ułożone są w sposób najbardziej równoległy. Nie ma sensu wówczas mówić o temperaturze związanej ze wzbudzeniem — mimo dużej energii wzbudzenia jądro jest „zamrożone", podobnie jak we wzbudzeniach bliskich stanowi podstawowemu. Innymi słowy, cała energia wzbudzenia jest skupiona w określonym przez spin obrocie jądra, nic



Rys. 16. Przykłady teoretycznie oszacowanych linii „yrast"

linia yrast

nie pozostaje na jego wzbudzenie wewnętrzne. Zależność maksymalnej wartości spinu od energii wzbudzenia (rys. 16) nosi nazwę linii „yrast" (szwedz. „najbardziej zakręcony"). Istnienie linii yrast ma poważny wpływ na rozpad stanów wysokospinowych, m.in. utrudnia emisję nukleonów, które unosząc z jądra znaczną energię wzbudzenia (równą ich energii kinetycznej zwiększonej o energię wiązania), muszą również wynieść duży spin, gdyż maksymalny spin jądra po emisji jest znacznie mniejszy od spinu wyjściowego.

stany wysokospinowe

Ze względu na małą gęstość poziomów o dużym spinie stopień ich zmieszania jest stosunkowo mały. Mimo znacznej energii wzbudzenia poziomy o dużym spinie są więc, podobnie jak poziomy w pobliżu stanu podstawowego, stosunkowo „czyste". Utworzenie jądra w stanie o wysokim spinie nie jest łatwe, gdyż wnikająca do jądra cząstka, niosąca duży moment pędu, ma energię na tyle wysoką, że mało prawdopodobne jest jej pozostanie w jądrze. Nie dotyczy to ciężkich jonów, które z powodu dużej masy nawet przy małych prędkościach mają w zderzeniu z jądrem duże momenty pędu. Właśnie dzięki reakcjom przez

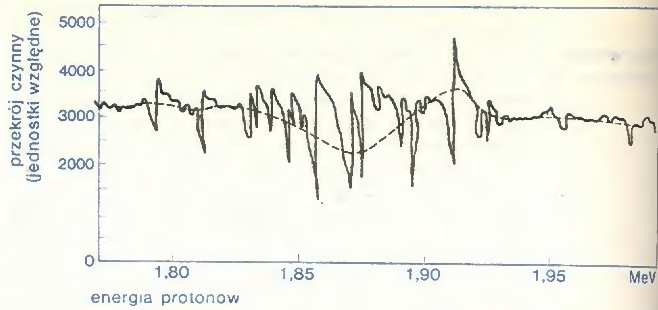
nie wywoływanym stany wysokospinowe są obecnie intensywnie badane.

Wspomniane uprzednio założenie o równoprawności każdej konfiguracji w obszarze gęsto leżących stanów nie jest jednakże zawsze spełnione. Okazuje się, że przy pewnych energiach wzbudzanie niektórych konfiguracji jest szczególnie łatwe, występują one w stanach wzbudzonych z wyjątkowo dużą amplitudą. Przykładem mogą być tu wszelkie rezonanse gigantyczne, które zawsze występują przy stosunkowo dużych energiach. Na przykład maksimum elektrycznego dipolowego rezonansu gigantycznego, polegającego na oscylacjach protonów i neutronów względem środka masy jądra, występuje przy energiach wzbudzenia wynoszących kilkanaście MeV. Rezonans ten ma spin różniący się od spinu stanu podstawowego o dodaną wektorowo wartość równą $1\hbar$ i ma przeciwną parzystość. We wszystkich stanach w tym obszarze energii wzbudzenia, mających spiny utworzone z wektorowego dodania do spinu stanu podstawowego wartości $1\hbar$ i mających przeciwną parzystość niż stan podstawowy, prawdopodobieństwo wystąpienia konfiguracji rezonansu dipolowego jest znacznie zwiększone, co objawia się przez uprzywilejowanie emisji z tych stanów dipolowych kwantów γ.

elektryczny dipolowy rezonans gigantyczny

analogowy rezonans izobaryczny

Podobną sytuację mamy w wypadku analogowych rezonansów izobarycznych. Znacznie upraszczając zagadnienie, naturę tych rezonansów można wyjaśnić następująco. Z niezależności ładunkowej sił jądrowych wynika, że stany jądrowe nie powinny zależeć od zamiany neutronu na proton znajdujący się na takiej samej orbicie (i odwrotnie — protonu na neutron; nie dotyczy to jąder ciężkich, w których istnieje nadmiar neutronów, więc wszystkie orbity neutronowe odpowiadające protonom są już zajęte), lecz powinny być prawie identyczne. Jednakże naładowany proton ma w polu kulombowskim jądra energię potencjalną wynoszącą od kilku do kilkunastu MeV (w zależności od ładunku Ze jądra), energia stanu musi być więc o taką wartość zwiększona. Zatem stan znajduje się teraz w obszarze gęsto położonych poziomów wzbudzonych jądra powstałego po zamianie neutronu na proton. Jest on zwany analogiem odpowiedniego niskoleżącego stanu w jądrze o tym samym A (a więc w izobarze), lecz o Z protonach i N neutronach, nazywanego stanem macierzystym. We wszystkich stanach wzbudzonych jądra $(Z+1, N-1)$ o spinie i parzystości takiej samej jak spin i parzystość stanu macierzystego, a znajdujących się w bezpośrednim sąsiedztwie miejsca, w którym powinien wystąpić stan analogowy, znacznie wzmocniona jest konfiguracja taka, jaką ma stan macierzysty (szczególnie jeśli jest to konfiguracja prosta). Uśrednienie widma stanów w pewnym przedziale energii uwidocznia pozornie istniejący w tym obszarze energii oddzielny stan rezonansowy o dużym wkładzie struktury stanu macierzystego zwany analogowym rezonansem izobarycznym (rys. 17). Do opisu własności tych stanów wprowadza się formalizm izospinu (zwanego również spinem izotopowym —



Rys. 17. Przekrój czynny reakcji rozpraszania elastycznego protonów przez jądra argonu: $^{40}A(p, p)^{40}A$. W obszarze, w którym powinien wystąpić stan analogowy (dla wzbudzeń odpowiadających energii wychwytywanych protonów ok. 1,87 MeV) szczególnie łatwo wzbudzają się poziomy o liczbach kwantowych takich samych jak stan analogowy — duży jest w nich udział konfiguracji stanu analogowego. Uśredniony przekrój czynny (linia przerywana) zachowuje się w tym obszarze energii protonów tak, jak gdyby istniał rezonans w wychwycie protonów

w formalizmie tym zakłada się, że proton i neutron są dwoma stanami tej samej cząstki różniącymi się ustawieniem izospinu w wyimaginowanej izoprzestrzeni).

metody wzbudzania jąder

Jądra można wzbudzać różnymi metodami. Wniknięcie nukleonu lub złożonej cząstki do jądra prowadzi na ogół do powstania stanów wysokowzbudzonych, gdyż poza energią kinetyczną wniesioną do jądra wydziela się energia wiązania nukleonu czy też grupy nukleonów. Tworzenie stanów o niższej energii w wyniku rozpadu stanów wysokowzbudzanych jest na ogół niezależne od ich struktury, która przy małych energiach wzbudzenia może być bardzo różnorodna (jest to wynikiem zmieszania różnych konfiguracji w stanie wysokowzbudzonym). Reakcje przekazu nukleonu lub grupy nukleonów między jądrem a cząstką przelatującą w bezpośrednim sąsiedztwie jego powierzchni preferują tworzenie stanów o konfiguracjach prostych, odpowiadających pojawieniu się nukleonów na określonych orbitach wokół niewzbudzonego jądra tarczy (lub wyrwaniu nukleonów z określonych jego orbit). W reakcjach rozpraszania nieelastycznego preferowane są wzbudzenia kolektywne (np. pobudzenie jądra do oscylacji lub rotacji przez pole jądrowe lub kulombowskie przelatującej w jego pobliżu cząstki).

Badania wzbudzeń jądrowych, ich konfiguracji (najczęściej za pomocą reakcji jądrowych) oraz rozpadów dostarczają podstawowych danych do poznania budowy jądra i natury sił jądrowych. Mimo olbrzymiego postępu w ostatnich latach, w dziedzinie tej jest jeszcze wiele zagadnień niedokładnie poznanych i nie rozwiązanych.

A. Bohr, B.R. Mottelson *Struktura jądra atomowego*, t. 1. Warszawa 1975; B.L. Cohen *Concepts of Nuclear Physics*, New York 1971; P. Marmier, E. Sheldon *Physics of Nuclei and Particles*, vol. 1, New York 1971; E. Segrè *Nuclei and Particles*, New York 1964.

Siły jądrowe

Adam Sobiczewski

oddziaływania silne

Siłami jądrowymi nazywamy specyficzne siły, które działają między nukleonami (neutronami i protonami), tj. składnikami jądra atomowego i powodują ich wiązanie. Stanowią one szczególny i najlepiej zbadany przypadek oddziaływań silnych występujących pomiędzy hadronami (barionami i mezonami, → Cząstki elementarne i ich oddziaływania). Oddziaływania silne tworzą odrębną klasę wśród czterech znanych w fizyce oddziaływań (silne, elektromagnetyczne, słabe i grawitacyjne) i są najsilniejszymi z nich.

W powyższym, tradycyjnym określeniu sił jądrowych, które tu przyjmujemy, ograniczamy się do oddziaływań występujących w jądrach zwykłych. W hiperjądrach występuje jeszcze dodatkowo oddziaływanie między nukleonami a cząstką Λ (→ Hiperjądra).

Chociaż często używa się intuicyjnego pojęcia siły, to jednak przy ilościowym opisie oddziaływania pomiędzy nukleonami stosuje się pojęcia energii oddziaływania lub potencjału, tj. te pojęcia, których się używa w mechanice kwantowej opisującej mikroobiekty, jakimi są nukleony.

W porównaniu z oddziaływaniami elektromagnetycznymi i grawitacyjnymi, wiedza nasza o oddziaływaniach jądrowych jest bardzo niepełna. Jednym z powodów tego jest ich bardzo mały zasięg (promień działania). Objawiają się one dopiero przy bardzo dużych zbliżeniach nukleonów (jak w jądrze atomowym lub przy zderzeniu jąder), a więc w sytuacjach, których wytwarzanie i badanie można było rozpocząć stosunkowo niedawno (w obecnym wieku, a intensywniej — dopiero od lat trzydziestych, gdy zbudowano pierwsze akceleratory). W oddziaływaniach między dużymi obiektami siły jądrowe nie odgrywają żadnej roli i dlatego nie można było wynieść o nich żadnej wiedzy z fizyki klasycznej opisującej takie obiekty. W klasycznej fizyce ważne są tylko oddziaływania dalekiego zasięgu, jak grawitacyjne i elektromagnetyczne. Drugim powodem niepełnej znajomości oddziaływań jądrowych jest skomplikowany ich charakter.

Głównym źródłem wiedzy o siłach jądrowych jest badanie najprostszych układów nukleonowych: neutron-proton (*n-p*) i proton-proton (*p-p*). Jest to badanie rozpraszania dwu nukleonów na sobie. Najdogodniejszy do analizy teoretycznej jest układ *n-p*, ponieważ nie zaburza go odpychanie elektryczne (kulombowskie). W odróżnieniu od układów *n-n* i *p-p*, układ *n-p* ma stan związany, który również dostarcza informacji o siłach jądrowych. Czystego układu *n-n* nie można badać, gdyż nie ma tarcz neutronowych.

Ważnym źródłem wiedzy są również własności jąder, np. zależność objętości (lub promienia) jądra od liczby masowej *A*, czy też średnia energia wiązania przypadająca na jeden nukleon w jądrze (→ Jądra atomowe i ich wzbudzenia)

Dodatkowym źródłem jest badanie oddziaływania mezonów (głównie mezonów π) z nukleonami. Przyjmuje się bowiem, że oddziaływanie jądrowe jest wynikiem wymiany mezonów pomiędzy nukleonami, podobnie jak oddziaływanie elektromagnetyczne jest wynikiem wymiany fotonów między cząstkami naładowanymi.

Własności sił jądrowych

Omówimy kilka charakterystycznych własności sił jądrowych. Jak pokazują wyniki rozpraszania nukleonów na sobie, a także własności jąder, zasięg sił jądrowych jest bardzo krótki; wynosi ok. 1–2 fm. Jest więc niewiele większy od promienia samego nukleonu i jednocześnie dosyć bliski średniej odległości między nukleonami w jądrze. Oznacza to z jednej strony, że nukleony w jądrze są dosyć ściśle „upakowane", a z drugiej, że każdy nukleon w jądrze oddziałuje tylko z nukleonami znajdującymi się najbliżej niego.

krótki zasięg

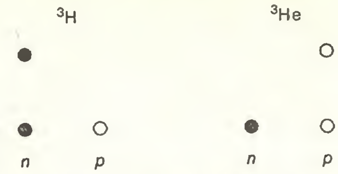
charakter przyciągający

Siły jądrowe są na ogół przyciągające. Inaczej nie mogłyby istnieć układy związane nukleonów, czyli jądra atomowe. Szczegółowy charakter tych sił jest jednak skomplikowany. Na bardzo małych odległościach, do ok. 0,4–0,5 fm, są one odpychające (tzw. rdzeń odpychający). Na dalszych odległościach mogą być zarówno przyciągające jak i odpychające, zależnie od stanu, w jakim znajduje się układ dwu nukleonów. Przeważa jednak charakter przyciągający.

Badanie własności jąder oraz rozpraszania nukleonów wskazuje, że siły jądrowe między dwoma neutronami (*n-n*), neutronem i protonem (*n-p*) oraz dwoma protonami (*p-p*) są w przybliżeniu takie same, jeśli tylko te pary cząstek znajdują się w identycznych stanach. Własność ta nazywa się ładunkową niezależnością sił jądrowych. Naturalnie, aby to stwierdzić, w badaniu oddziaływania *p-p* należy uwzględnić (odjąć) efekt odpychania kulombowskiego.

niezależność ładunkowa

Jedną z dróg sprawdzania ładunkowej niezależności jest porównanie energii wiązania dwóch jąder zwierciadlanych, tj. pary jąder, których liczba protonów różni się o jeden, a liczba masowa jest taka sama: np. pary ^3H–^3He (rys. 1) lub ^{11}B–^{11}C. Jądra takie przechodzą w siebie nawzajem przy zamianie wszystkich neutronów na protony i odwrotnie. Okazuje się, że z dużą dokładnością różnica między energiami wiązania tych jąder równa jest różnicy między

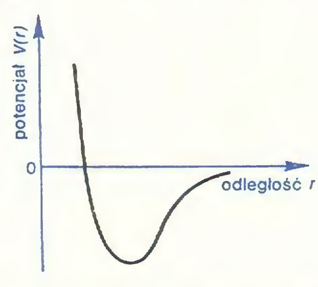


Rys. 1. Przykład jąder zwierciadlanych

ich energiami kulombowskimi, wywołanej różną liczbą zawartych w nich protonów. Nie pojawienie się żadnej innej różnicy świadczy o tym, że siły jądrowe *n-n* i *p-p* są takie same. Własność ta, zwana symetrią ładunkową sił jądrowych, jest węższa od własności niezależności ładunkowej. Aby stwierdzić niezależność ładunkową, należy jeszcze wykazać, że siły *n-p* są identyczne z *n-n* i *p-p*. Do tego celu nie wystarcza porównywanie jąder zwierciadlanych, bowiem liczba wiązań *n-p* jest w nich taka sama. Należy rozważyć szersze układy, np. trójki jąder jak: ^6He–^6Li–^6Be czy ^8Li–^8Be–^8B. Porównanie ich energii wiązania, a także widm poziomów wzbudzonych, wskazuje na ładunkową niezależność sił. Na własność tę wskazuje także porównanie wyników rozpraszań *n-p* i *p-p*. Gdyby siły jądrowe były tylko przyciągające, to gęstość jądra rosłaby szybko ze wzrostem liczby masowej *A*. Każdy nukleon dążyłby bowiem do pozostawania w zasięgu oddziaływania wszystkich pozostałych i promień jądra nie przekraczałby tego zasięgu. Energia wiązania byłaby wtedy proporcjonalna do liczby wszystkich możliwych wiązań, tj. liczby oddziałujących par, tzn. do $A(A-1)/2$, czyli w przybliżeniu do A^2. Tymczasem z doświadczenia wiadomo, że tak nie jest; gęstość jądra jest w przybliżeniu stała, a energia wiązania — proporcjonalna do *A*. Fakty te świadczą o własności tzw. wysycania (nasycania się) sił jądrowych. Mianowicie, siły te muszą być silnie odpychające przy małych odległościach (rdzeń odpychający), by zapobiec wzrostowi gęstości przy wzroście *A*, oraz muszą mieć taki zasięg, by każdy nukleon oddziaływał tylko z najbliższymi sąsiadami, a nie ze wszystkimi nukleonami w jądrze. Ten ostatni warunek jest po to, by energia wiązania była proporcjonalna do *A*, a nie A^2. Sytuacja jest podobna do tej, jaka istnieje w cieczy i w ciele stałym. Na przykład siły występujące między molekułami w cieczy są odpychające na małych odległościach, a przyciągające na większych, przy czym zasięg sił jest krótki, tak że cząstka może oddziaływać tylko z cząstkami najbliższymi. Wykres potencjału takich sił podany jest na rys. 2.

symetria ładunkowa

własność wysycania



Rys. 2. Przykład potencjału sił o własności wysycania

Oprócz rdzenia odpychającego, do własności wysycania sił wnosi także swój wkład składowa wymienna tych sił (patrz niżej). Składowa ta jest bowiem zależna od stanu i powoduje, że w pewnych stanach siły jądrowe są odpychające na każdej odległości.

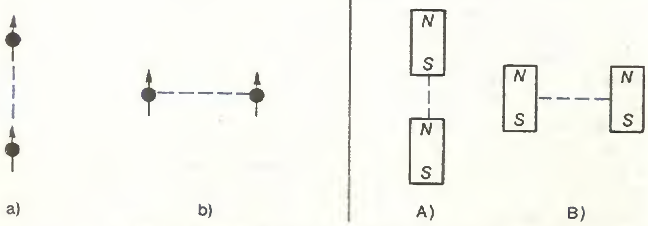
zależność od spinu

Siły jądrowe zależą od względnego ustawienia spinów oddziałujących nukleonów. O zależności tej świadczy np. fakt, że stan 1*s* (tj. najniższy energetycznie stan o orbitalnym momencie pędu $L = 0$) układu *n-p* jest związany, gdy spiny neutronu i protonu ustawione są równolegle (stan trypletowy o całkowi-

tym spinie $S = 1$), a nie jest związany, gdy spiny są ustawione antyrównolegle (stan singletowy o całkowitym spinie $S = 0$).

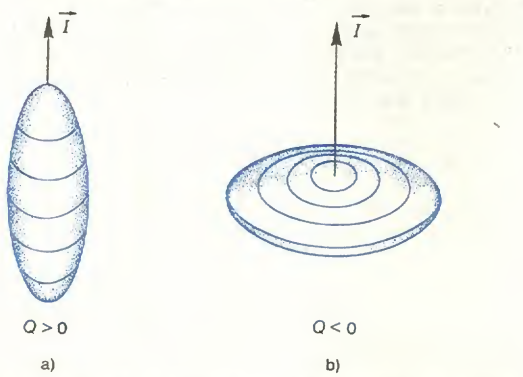
charakter niecentralny

Siła między dwoma ciałami (punktowymi) działająca w kierunku linii łączącej te ciała nazywa się siłą centralną. Przykładem jest siła przyciągania lub odpychania elektrostatycznego (kulombowskiego) pomiędzy elektrycznymi ładunkami punktowymi. Wartość bezwzględna siły centralnej (a także jej potencjał) zależy tylko od odległości między ciałami. Oprócz odległości, siła może zależeć także od orientacji wektora wzajemnego położenia względem kierunku wyróżnionego w przestrzeni. W przypadku nukleonów może to być kierunek ich spinów. Siła taka nie jest już centralna i nazywa się siłą tensorową. Na przykład przy ustawieniu nukleonów takim, jak na rys. 3a (wektor wzajemnego położenia współliniowy ze spinami) siła tensorowa jest inna niż przy ustawieniu jak na rys. 3b (wektor położenia prostopadły do spinów). Przypomina to sytuację, jaką mamy przy znanym, klasycznym oddziaływaniu dwu magnesów. Przy ustawieniu A przyciągają się one, a przy ustawieniu B — odpychają. Dla ilustracji wzięliśmy tu tylko ustawienia skrajne; przy ustawieniach pośrednich siła przyjmuje również wartości i kierunki pośrednie.



Rys. 3. Przykład różnych ustawień wektora wzajemnego położenia dwu nukleonów względem kierunku ich spinów. Porównanie z różnymi ustawieniami dwu magnesów

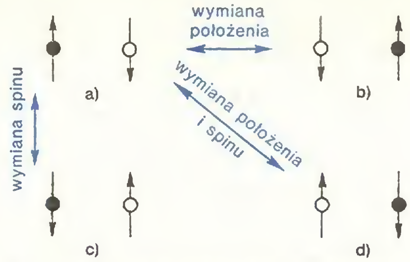
Jednym z argumentów, który świadczy o niecentralnym charakterze sił jądrowych, jest niezerowy moment kwadrupolowy (elektryczny) Q deuteronu. Z pomiarów wynika, że moment ten jest dodatni. Rysunek 4 ilustruje, że dodatniemu momentowi Q odpowiada kształt wydłużony jądra, typu cygara (rys. 4a), a ujemnemu — kształt spłaszczony, typu dysku (rys. 4b). Porównanie rysunków 4 i 3 sugeruje, że wydłużony kształt deuteronu przemawia za uprzywilejowaniem ustawienia (a) z rys. 3, tj. za przyciągającym charakterem sił jądrowych przy tym ustawieniu. Analogia z siłami między dwoma magnesami jest więc bardzo duża. Jednakże natura tych dwu rodzajów sił jest inna. Oddziaływanie jądrowe jest silniejsze i ma inną zależność od odległości, identyczna jest tylko zależność kątowa.



Rys. 4. Przykład wydłużonego (a) i spłaszczonego (b) kształtu jądra. $\vec{I}$ jest momentem pędu, a Q momentem kwadrupolowym jądra

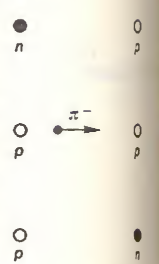
charakter wymienny

Część oddziaływania jądrowego stanowią siły wymienne, które są efektem czysto kwantowym, nie mającym żadnej analogii klasycznej. Działanie ich polega na wymianie między nukleonami niektórych lub wszystkich wielkości opisujących te nukleony, np.: położenia (tzw. siły Majorany), spinu (siły Bartletta) oraz zarówno położenia jak i spinu (siły Heisenberga). Dla pełności terminologii dodajmy, że zwykłe siły jądrowe, które nie prowadzą do wymiany żadnej wielkości, nazywają się siłami Wignera. Różne możliwości wymiany dla układu n-p są zilustrowane na rys. 5. Biorąc stan a za wyjściowy, stan b otrzymuje się z niego przez wymianę położenia, stan c — przez wymianę spinu, a stan d — przez wymianę obu tych wielkości. To znaczy stan b otrzymywany jest ze stanu a (i odwrotnie), jeśli działają tylko siły Majorany, stan c — jeśli tylko siły Bartletta, a d — jeśli tylko siły Heisenberga. Siły Wignera pozostawiają stan a (jak również pozostałe stany b, c, d) bez zmiany.



Rys. 5. Procesy wymiany układu neutron-proton. Neutron oznaczony jest kółkiem wypełnionym, proton — pustym, a spin — strzałką

Siły wymiany znajdują wyjaśnienie w mezonowej teorii sił jądrowych (patrz niżej). Na przykład procesy prowadzące od stanu a do b lub d z rys. 5 mogą być zrealizowane drogą wymiany między neutronem i protonem mezonu naładowanego π^- lub π^+ (rys. 6).



Rys. 6. Wymiana mezonu π^- między neutronem a protonem

Najbardziej bezpośredniego dowodu doświadczalnego występowania sił wymiany dostarcza badanie rozpraszania neutronów na protonach. Jeśli energia neutronu padającego na tarczę protonową jest duża, znacznie większa od energii oddziaływania (przyciągania czy odpychania) między nim a protonem, to neutron zdoła przekazać protonowi w trakcie oddziaływania („zderzenia") tylko małą część swojego pędu. Po zderzeniu będzie więc nadal poruszał się z dużą prędkością i w kierunku mało zmienionym w stosunku do początkowego (rys. 7c i d). Rozkład kątowy takich neutronów po zderzeniu będzie miał zatem silne maksimum przy małych kątach (rys. 7e). Maksimum przy dużych kątach (tzw. rozpraszanie do tyłu) możemy otrzymać jedynie w przypadku wymiany neutronu na proton i odwrotnie (rys. 7h). Fakt, że w doświadczeniu otrzymujemy maksimum zarówno przy małych, jak i przy dużych kątach (rys. 7i), świadczy o obecności zarówno zwykłej jak i wymiennej składowej w oddziaływaniu jądrowym. Na rys. 7 proces rozpraszania zilustrowany został w układzie laboratoryjnym (a, c, f), w którym proton tarczy spoczywa przed zderzeniem, oraz w układzie środka masy (b, d, g), w którym sumaryczny pęd obu nukleonów jest równy zeru.

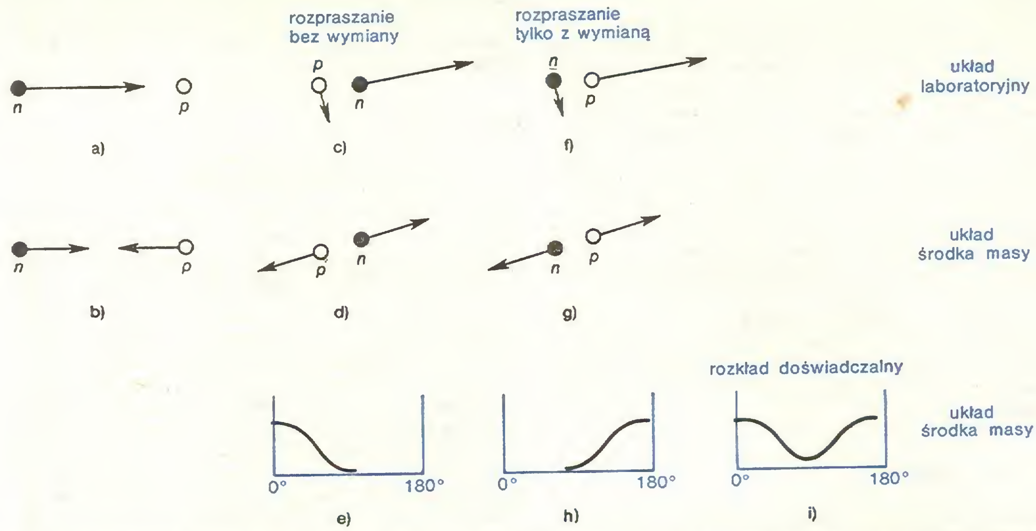
siły spin-orbita

Siły jądrowe zależą nie tylko od wzajemnego ustawienia spinów oddziałujących nukleonów, ale i od ustawienia ich względem wektora orbitalnego momentu pędu $\vec{L}$. Część sił powodującą tę zależność nazywamy siłami spin-orbita. O istnieniu ich świadczą bezpośrednio doświadczenia polaryzacyjne, tzn. takie, w których bądź tarcza, bądź wiązka padająca, bądź też produkty rozpraszania są spolaryzowane, czyli mają więcej cząstek ze spinem skierowanym do góry, niż w dół lub na odwrót.

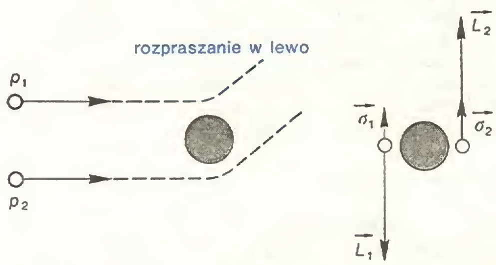
Rozważmy dla ilustracji (rys. 8) rozpraszanie spolaryzowanej wiązki protonów ze spinami ustawio-

nymi do góry na jądrach o spinie równym zeru. Oddziaływanie protonu wiązki z jądrem jest naturalnie sumą oddziaływań ze wszystkimi nukleonami jądra.

tego zasada zachowania energii. Cząstka wirtualna może być jednak wyemitowana bez spełnienia tej zasady, byle na dostatecznie krótki czas, na jaki pozwala



Rys. 7. Rozpraszanie neutronu o dużej energii na protonie w układzie laboratoryjnym i układzie środka masy. U dołu pokazane są trzy rozkłady kątowe rozproszonych neutronów: (e) — rozkład jaki otrzymalibyśmy gdyby nie było w ogóle wymiany, (h) — gdyby była wyłącznie wymiana oraz (i) — rozkład jaki otrzymuje się w doświadczeniu



Rys. 8. Rozpraszanie spolaryzowanych protonów na jądrze bezspinowym. Obok zilustrowano orientację spinu protonów względem ich orbitalnego momentu pędu

Załóżmy dla uproszczenia, że działają tylko siły spin-orbita i że są one przyciągające, gdy spin $\vec{\sigma}$ jest skierowany zgodnie z momentem orbitalnym $\vec{L}$ oraz odpychające w przypadku przeciwnym. Widać z rys. 8, że protony będą wówczas rozpraszane tylko w lewo.

Ponieważ w rzeczywistości siły spin-orbita są jedną z kilku a nie jedyną składową sił jądrowych, rozpraszanie występuje w obie strony. Obserwowana jednak różnica ilości cząstek rozproszonych w lewo i w prawo (tzw. asymetria lewo-prawo) świadczy o istnieniu tych sił

Mezonowa teoria sił jądrowych

Oddziaływanie między obiektami fizycznymi realizowane jest za pośrednictwem pola. W opisie kwantowym (→ Teoria pola) oznacza to wymianę kwantów — cząstek tego pola. Wymiana ta prowadzi do przekazania pędu od jednego obiektu do drugiego, a więc do występowania sił między nimi. Na przykład oddziaływanie elektromagnetyczne polega na wymianie kwantów pola elektromagnetycznego, czyli fotonów.

cząstki wirtualne Cząstki przenoszące oddziaływanie są cząstkami wirtualnymi, tj. cząstkami nieobserwowalnymi bezpośrednio w procesie oddziaływania. Możliwość ich obserwacji oznaczałaby bowiem zwykłą ich emisję, dla której trzeba by dostarczyć emitującemu je obiektowi energię potrzebną do ich wytworzenia, a więc równą co najmniej ich energii spoczynkowej. Wymaga

zasada nieokreśloności Heisenberga, czyli co najwyżej na czas:

$$\Delta t = \hbar/\Delta E \lessapprox \hbar/m_0c^2,$$

gdzie ΔE — energia cząstki, m_0 — jej masa spoczynkowa, c — prędkość światła. W czasie tym cząstka wirtualna może się oddalić co najwyżej na odległość:

$$r = c\Delta t \lessapprox \hbar c/m_0c^2.$$

Odległość:

$$r_z \approx \hbar c/m_0c^2 \qquad (1)$$

jest zatem zasięgiem sił, jaki otrzymuje się za pośrednictwem pola, którego kwanty mają masę spoczynkową m_0. Widać stąd, że siły elektromagnetyczne mają zasięg nieskończony (siły długiego zasięgu), ponieważ dla fotonu $m_0 = 0$. Zasięg zaś sił jądrowych jest: **zasięg sił jądrowych**

$$r_z \approx \hbar c/(140 \text{ MeV}) \approx 1{,}4 \text{ fm},$$

ponieważ masa spoczynkowa mezonu π (pionu), przenoszącego oddziaływanie jądrowe, wynosi ok. 140 MeV/c^2.

Zasięg sił jądrowych obliczyliśmy tutaj mając zmierzoną masę dobrze obecnie znanego mezonu π. Historycznie było jednak odwrotnie. Dla objaśnienia sił jądrowych wysunięto, śmiałą na owe czasy, hipotezę istnienia cząstek silnie oddziałujących z nukleonami i przenoszącymi oddziaływanie jądrowe. Znając z doświadczenia zasięg (krótki) tego oddziaływania można było oszacować na podstawie wzoru (1) ich masę m_0. Dokonał tego fizyk japoński H. Yukawa w 1934 r., na długo przed doświadczalnym odkryciem samych cząstek (mezonów π), które nastąpiło dopiero w 1947 r. Historia ta jest jednym z ładniejszych przykładów, jak teoria może dać objaśnienie zjawiska na długo przed odkryciem jego podstaw doświadczalnych i w efekcie odkrycie to przyśpieszyć.

Ze wzoru (1) wynika, że zasięg sił jest tym mniejszy, im większa jest masa spoczynkowa przenoszących je cząstek. Na przykład wymiana dwu pionów na raz prowadzić będzie do sił o dwukrotnie mniejszym zasięgu niż wymiana pojedynczego pionu.

potencjał Yukawy W bardziej ilościowym ujęciu wymiana jednej cząstki prowadzi do oddziaływania opisanego potencjałem Yukawy:

$$V(r) = -g^2 e^{-\mu r}/r,$$

gdzie g — stała, $\mu = m_0c^2/\hbar c$, przy czym m_0 jest masą spoczynkową tej cząstki. Krótki zasięg opisany wzorem (1) przy $m_0 \neq 0$ znajduje tu swoje odbicie w szybkim, wykładniczym spadku oddziaływania z odległością.

Mówiąc dokładniej, potencjał Yukawy odpowiada wymianie cząstki skalarnej, tj. cząstki o spinie zero i parzystości dodatniej. Wymiana cząstki pseudoskalarnej (spin zero i parzystość ujemna), jaką jest pion, prowadzi do potencjału tensorowego. Wreszcie, wymiana cząstki wektorowej (spin równy jeden) pozwala odtworzyć rdzeń odpychający oraz oddziaływanie spin-orbita. Ponieważ obecnie znamy mezony

wymiana mezonów

o wszystkich wymienionych cechach, można oczekiwać, że wymiana ich pozwoli odtworzyć wszystkie omówione własności oddziaływania jądrowego. Rzeczywiście, z pomocą wymiany jednomezonowej znanych mezonów o masie do 1 GeV/c^2 możemy obecnie zadowalająco odtworzyć znane z doświadczenia i omówione w poprzednim paragrafie oddziaływanie między nukleonami. Nie oznacza to jednak jeszcze, że stan mezonowej teorii sił jądrowych jest już obecnie zadowalający.

Zapis sił jądrowych

Dodajmy jeszcze kilka uwag o sposobie zapisu różnych składowych sił jądrowych.

Siły zależne tylko od odległości (siły centralne) możemy opisać potencjałem:

$$V_c(r),$$

gdzie r — odległość między nukleonami.

Siły zależne od wzajemnego ustawienia spinów dają się opisać wyrażeniem:

$$V_S(r)\vec{\sigma}_1\cdot\vec{\sigma}_2,$$

gdzie $\vec{\sigma}_1$ i $\vec{\sigma}_2$ — wektory spinu cząstek 1 i 2. Siły te są różne przy dwu możliwych ustawieniach spinów: równoległym (stan trypletowy) i antyrównoległym (stan singletowy), gdyż:

$$\vec{\sigma}_1\cdot\vec{\sigma}_2 = \begin{cases} 1 \text{ dla stanu trypletowego,} \\ -3 \text{ dla stanu singletowego.} \end{cases} \tag{2}$$

Siły tensorowe można opisać potencjałem:

$$V_T(r)\, S_{12},$$

gdzie:

$$S_{12} = 3\frac{(\vec{\sigma}_1\cdot\vec{r})(\vec{\sigma}_2\cdot\vec{r})}{r^2} - \vec{\sigma}_1\cdot\vec{\sigma}_2,$$

a siły spin-orbita wyrażeniem:

$$\tfrac{1}{2}V_{LS}(r)\vec{L}\cdot(\vec{\sigma}_1+\vec{\sigma}_2).$$

We wzorach powyższych $\vec{r} = \vec{r}_1 - \vec{r}_2$ jest wektorem względnego położenia nukleonów, a $\vec{L} = \vec{r}\times\vec{p}$ jest wektorem względnego orbitalnego momentu pędu, przy czym $\vec{p}$ jest pędem względnym.

potencjał oddziaływania 2 nukleonów

Zatem całkowity potencjał oddziaływania między dwoma nukleonami ma postać:

$$V_{NN} = V_c(r) + V_S(r)\vec{\sigma}_1\cdot\vec{\sigma}_2 + V_T(r)S_{12} + \tfrac{1}{2}V_{LS}(r)\vec{L}\cdot(\vec{\sigma}_1+\vec{\sigma}_2). \tag{3}$$

Jest on zbudowany tak, aby był skalarem (niezmienniczość względem obrotu w przestrzeni zwykłej), a także aby zapewniał zachowanie parzystości (niezmienniczość względem odbicia w początku układu współrzędnych) obowiązujące w oddziaływaniach silnych. Aby zapewnić niezależność ładunkową (niezmienniczość względem obrotu w przestrzeni izospinowej), każda z czterech funkcji $V_c(r)$, $V_S(r)$, $V_T(r)$ i $V_{LS}(r)$ powinna być postaci:

$$V(r) = V_1(r) + V_2(r)\vec{\tau}_1\cdot\vec{\tau}_2,$$

gdzie $\vec{\tau}$ — wektor izospinu nukleonu.

Można wykazać, że wzór (3) jest najogólniejszą postacią potencjału nukleon-nukleon, jeśli uwzględnić zależność jego od prędkości (czy pędu $\vec{p}$, czy też momentu pędu $\vec{L}$) tylko w pierwszej potędze. Między innymi opisuje on wszystkie siły wymienne. Na przykład wymianę spinu otrzymuje się działając na funkcję falową wyrażeniem:

$$P_\sigma = \tfrac{1}{2}(1+\vec{\sigma}_1\cdot\vec{\sigma}_2), \tag{4}$$

które zgodnie z (2) daje:

$$P_\sigma = \begin{cases} 1 \text{ dla stanu trypletowego,} \\ -1 \text{ dla stanu singletowego,} \end{cases}$$

tzn. nie zmienia znaku funkcji falowej w stanie trypletowym, a zmienia znak w stanie singletowym; prowadzi więc dokładnie do tego samego, co operacja zamiany spinów między cząstkami 1 i 2. Obecność wyrażenia $\vec{\sigma}_1\cdot\vec{\sigma}_2$ w potencjale (3) oznacza zatem, zgodnie ze wzorem (4), obecność operatora wymiany spinu P_σ, opisującego spinowe siły wymienne.

Potencjał (3) zawiera ostatecznie osiem funkcji odległości $V_i(r)$, które należy wyznaczyć z analizy wyników rozpraszania N-N (tzn. nukleonu na nukleonie). Wykonano wiele takich analiz; wyniki ich nie są jednoznaczne. Wszystkie one prowadzą jednak do kilku wspólnych wniosków, z których ważniejsze to: 1) wszystkie człony w oddziaływaniu (3) są potrzebne, 2) na małych odległościach, mniejszych od ok. 0,5 fm, oddziaływanie jest odpychające (rdzeń odpychający), 3) oprócz zależności od pełnego spinu, oddziaływanie zależy także od parzystości stanu, w którym zachodzi; w stanach o dodatniej parzystości (L — parzyste) jest ono silniejsze niż w stanach o parzystości ujemnej (L — nieparzyste) i jest przyciągające, 4) na dużych odległościach oddziaływanie jest dobrze opisywane przez potencjał Yukawy odpowiadający wymianie jednego pionu.

Siły wielociałowe

Dotychczas zakładaliśmy, że siły jądrowe mają naturę dwuciałową. Oznacza to, że jeśli mamy np. trzy znajdujące się blisko siebie nukleony A, B i C (rys. 9), to siła działająca np. na nukleon A jest: $F_{AB}+F_{AC}$, gdzie F_{AB} jest siłą, która działa między nukleonami A i B w nieobecności nukleonu C, a F_{AC} — siłą działającą między A i C w nieobecności B. Inaczej mówiąc, siły dwuciałowe działają tylko pomiędzy parą cząstek i nie są modyfikowane przez obecność innych cząstek. Na przykład, w bardzo dobrym przybliżeniu siły elektromagnetyczne i grawitacyjne mają taki właśnie charakter.



Rys. 9. Trzy wzajemnie oddziałujące nukleony

Istnieje jednakże możliwość występowania także trój- i więcejciałowych sił jądrowych. Możemy to przedstawić następująco. Przy opisie sił jądrowych za pomocą wymiany mezonów wymiana jednego mezonu odpowiada siłom dwuciałowym, bo tylko między dwoma nukleonami wymiana taka może zachodzić i nie jest ona przy tym zależna od obecności trzeciego nukleonu. Natomiast przy wymianie dwu mezonów na raz jest już i inna możliwość niż wymiana obu między dwoma nukleonami. Mianowicie jeden z mezonów wyemitowanych przez nukleon A może być pochłonięty przez nukleon B, a drugi — przez C. Odpowiada to już siłom trójciałowym. Analogicznie można przedstawić sobie cztero-, pięcio- i ogólnie wielociałowe siły. Widać, że przy takiej naturze tych sił, siły trójciałowe będą miały zasięg dwa razy mniejszy niż dwuciałowe, a czterociałowe — trzy razy mniejszy itd. Wynika to ze wzoru (1), gdyż m_0 jest w tym wypadku masą dwu, trzech i odpowiednio więcej mezonów. Skoro zasięg sił dwuciałowych wynosi ok. 1,4 fm, to trójciałowych wyniesie ok. 0,7 fm, a czterociałowych — ok. 0,47 fm. Ponieważ średnia odległość między nukleonami w jądrach wynosi ok. 2,1 fm — widać, że rola sił maleje ze wzrostem stopnia ich wielociałowości. Ponadto, ponieważ

siły dwuciałowe wykazują silne odpychanie na odległościach mniejszych od ok. 0,5 fm i praktycznie nie dopuszczają do zbliżenia dwu nukleonów większego niż na tę odległość — należy oczekiwać, że rola sił cztero- i więcejciałowych jest już do pominięcia. Warto zatem badać jedynie siły trójciałowe, w uzupełnieniu do dwuciałowych. Poświęcono już wiele uwagi różnym efektom sił trójciałowych. Nadal jednak wnioski z tych badań, co do roli sił trójciałowych w jądrach, nie są zdecydowane i rozstrzygające.

D.M. BRINK *Nuclear forces*, Oxford 1965; B.L. COHEN *Concepts of nuclear physics*, New York 1971; *Encyklopedia fizyki*, t. 3, str. 318, Warszawa 1974; H. FRAUENFELDER AND E.M. HENLEY *Subatomic physics*, Englewood Cliffs 1974; R. SACHS *Fizyka teoretyczna jądra atomowego*, Warszawa 1957; A. STRZAŁKOWSKI *Wstęp do fizyki jądra atomowego*, Warszawa 1973: Sz. SZCZENIOWSKI *Fizyka doświadczalna*, cz. 6, Warszawa 1974; Z. WILHELMI *Fizyka reakcji jądrowych*, Warszawa 1976.

Modele jądrowe

Adam Sobiczewski

zagadnienie wielu ciał

Jądra poznanych dotychczas pierwiastków są układami od jednego do ok. 260 nukleonów; poza pierwiastkami najlżejszymi są więc one układami wielu ciał. Opis takiego układu (czyli rozwiązanie zagadnienia wielu ciał) polegałby na pełnym określeniu ruchu wszystkich nukleonów powstającym pod wpływem ich wzajemnego oddziaływania. Wiemy jednak, że w ogólnym wypadku zagadnienia takiego nie potrafimy rozwiązać ściśle już dla trzech ciał, ani w mechanice kwantowej ani klasycznej, nawet przy stosunkowo prostym oddziaływaniu. Tymczasem oddziaływanie między nukleonami jest skomplikowane i nie w pełni jeszcze poznane (→ Siły jądrowe).

Wyjątkowo trudne zadanie, jakie mamy przy opisie układu wielu nukleonów (jądro atomowe), można zilustrować porównując je z opisem powłoki elektronowej atomu (tzn. układu wielu elektronów), który to opis jest już bardzo złożony. Przede wszystkim, elektrony oddziałują ze sobą za pomocą dobrze znanych i stosunkowo prostych sił elektromagnetycznych, których główną składową jest odpychanie kulombowskie. Oddziaływanie między nukleonami zaś, jak wspomnieliśmy, jest skomplikowane i nie w pełni znane. Poza tym, elektrony są w polu silnie naładowanego jądra (o ładunku równym sumie ładunków wszystkich elektronów — w atomie obojętnym), którego masa jest znacznie (ok. 1840 razy) większa od masy elektronu i które zatem prawie spoczywa w układzie środka masy. W tych warunkach oddziaływanie między elektronami jest tylko poprawką (sprowadzającą się w dużej mierze do częściowego ekranowania od pola jądra elektronów zewnętrznych przez wewnętrzne, a więc do zmniejszania ładunku jądra „widzianego" przez elektrony zewnętrzne) do oddziaływania ich z niemal nieruchomym centrum pola, jakim jest jądro. W jądrze zaś nie ma takiego centrum; wszystkie nukleony poruszają się „równoprawnie" w polu jedynie wzajemnego oddziaływania. Pokazuje to, że opis ruchu nukleonów jest znacznie jeszcze trudniejszy od i tak trudnego zadania, jakim jest opis ruchu elektronów w atomie.

Wszystko to powoduje, że w opisie jąder musimy uciekać się do znacznych uproszczeń, przybliżeń. Modele jądrowe są takimi właśnie uproszczeniami. Stosuje się je zarówno przy opisie struktury jądra, jak i reakcji jądrowych. W tym ostatnim wypadku, zamiast o modelu, mówi się często o mechanizmie reakcji (→ Reakcje jądrowe). W artykule niniejszym omówimy modele struktury jądra.

Model powinna cechować przede wszystkim prostota oraz łatwość przewidywania na jego podstawie własności jądra. Prostota ta jednak przyczynia się do organiczenia modelu oraz sprawia, że za jego pomocą można opisać tylko niektóre własności jądra. Do opisu innych własności jądra trzeba użyć innego modelu, który często różni się znacznie od poprzedniego, a czasem wydaje się wręcz z nim sprzeczny. Zachęcające do używania modeli jest jednak to, że niektóre z nich mają bardzo szeroki zakres stosowalności, co świadczy o trafności zastosowanych przybliżeń. Prostota modelu łączy się na ogół z jego poglądowością, która jest pomocna w wyrobieniu pewnej intuicji ułatwiającej jego stosowanie. Cechę tę mają zwłaszcza modele oparte na analogiach klasycznych (np. model kroplowy jądra).

cząstki niezależne lub silnie skorelowane

Obecnie stosuje się ok. dziesięciu modeli struktury jądra, które są pewną odmianą lub połączeniem dwu zasadniczych modeli: modelu cząstek niezależnych oraz modelu cząstek silnie skorelowanych. W pierwszym z nich nukleony lub ich grupy poruszają się we wspólnym potencjale niezależnie lub prawie niezależnie od siebie. W drugim — nukleony oddziałują ze sobą tak silnie, że ruch jednego jest ściśle skorelowany z ruchem innych. Różne odmiany i połączenia tych dwu bardzo odbiegających od siebie wyobrażeń o ruchu nukleonów w jądrze pozwalają opisać bardzo różne własności jąder: od własności, w których korelacje między nukleonami są nieważne (efekty jednocząstkowe) do takich, w których są one podstawowe (efekty kolektywne).

Modele podstawowe

Omówimy kilka podstawowych modeli struktury jądra, które są najczęściej stosowane, bądź bezpośrednio do opisu własności jądra, bądź do konstrukcji bardziej złożonych modeli.

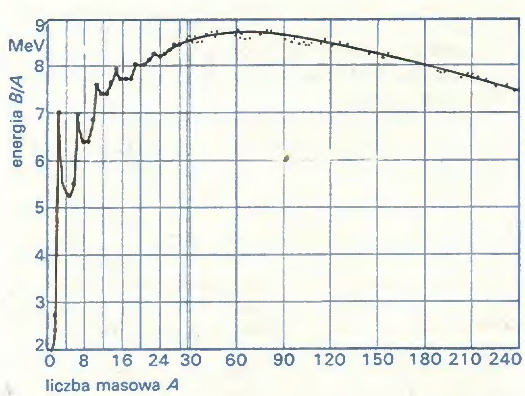
Model kroplowy

Model kroplowy, który wykorzystuje analogię między jądrem a kroplą cieczy, jest najprostszą wersją modelu silnych korelacji. Podstawą tej analogii są dwa fakty doświadczalne: a) stała gęstość materii w jądrze, niezależna od jego wielkości. Wynosi ona

$$\varrho \approx 0{,}17 \text{ nukleonu/fm}^3, \tag{1}$$

energia wiązania

b) niemal stała wartość energii wiązania jądra w prze-



Rys. 1. Zależność doświadczalnej wartości energii wiązania jądra na jeden nukleon B/A od liczby masowej A. Aby lepiej zilustrować strukturę tej zależności dla jąder lekkich, rozciągnięto dla nich skalę A

liczeniu na jeden nukleon B/A (rys. 1); wynosi ona ok. 8 MeV. Fakt, że B/A jest prawie stałe, oznacza, że energia wiązania B jest prawie proporcjonalna do A, lub inaczej, że część B, która jest proporcjonalna do A (energia objętościowa) czyli

energia objętościowa

$$B_v = a_v A, \tag{2}$$

przy czym a_v — współczynnik proporcjonalności, jest jej częścią główną.

Obie podane własności jądra są charakterystyczne dla cieczy; gęstość jej jest stała, niezależnie od objętości (a nawet, w dużej mierze od ciśnienia zewnętrznego). Także ciepło parowania — które jest odpowiednikiem energii wiązania — w przeliczeniu na jednostkę objętości, a zatem i na jedną cząstkę, jest stałe. Wyparowanie jednej cząstki (molekuły), a więc oswobodzenie jej z więzów łączących ją z innymi cząstkami, wymaga stałej energii określonej dla danej cieczy. Obie własności są wyrazem wysycania sił działających między molekułami cieczy, czy nukleonami w jądrze; siły takie działają tylko między najbliższymi sąsiadami.

Analogię z cieczą można rozszerzyć i przyjąć, że nukleony znajdujące się na powierzchni jądra są wiązane słabiej od tych, które są wewnątrz. Mają one bowiem ok. dwa razy mniej nukleonów w swoim najbliższym sąsiedztwie. Powoduje to zmniejszenie energii wiązania o wielkość proporcjonalną do liczby nukleonów powierzchniowych, a zatem do pola powierzchni jądra, a więc do $A^{2/3}$. W rezultacie zmniejszenie to jest

energia powierzchniowa

$$B_s = -a_s A^{2/3}. \tag{3}$$

model kropelkowy

Idąc jeszcze dalej, można uwzględnić wpływ zakrzywienia powierzchni jądra i dodać wyrażenie $-a_k A^{1/3}$. Dokonuje się tego w modelu kropelkowym jądra będącym ulepszeniem modelu kroplowego.

Przy obliczaniu energii wiązania jest niezbędne uwzględnienie faktu, że „ciecz" jądrowa jest naładowana dodatnio. Trzeba więc dodać (ujemną) energię odpychania kulombowskiego

energia kulombowska

$$B_{kul} = -\frac{3}{5}\frac{(Ze)^2}{R} = -\frac{3}{5}\frac{(Ze)^2}{rA^{1/3}} = -a_{kul}\frac{Z^2}{A^{1/3}}, \tag{4}$$

gdzie $R = rA^{1/3}$ — promień jądra, e — ładunek protonu, a Z — liczba protonów w jądrze (liczba atomowa).

Jest to już wszystko, co można osiągnąć stosując model kroplowy do opisu energii wiązania. Dalszy istotny fakt, że w niezbyt lekkich jądrach jest więcej neutronów niż protonów, uwzględnia się już za pomocą modelu gazu Fermiego (patrz niżej). Daje on poprawkę postaci

energia symetrii

$$B_{sym} = -a_{sym}\frac{(N-Z)^2}{A}, \tag{5}$$

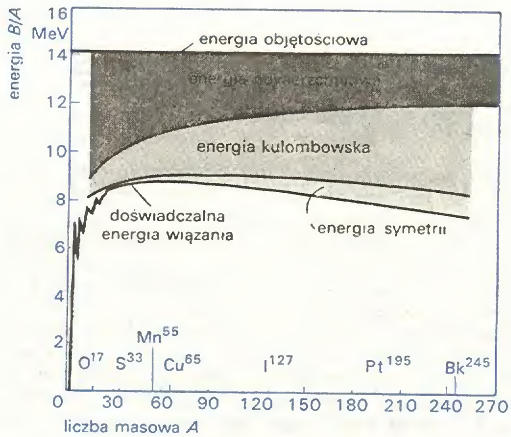
która mówi, że wiązanie jądra jest najsilniejsze, gdy jest ono symetryczne względem zamiany neutronów na protony i odwrotnie, czyli gdy liczba neutronów N równa jest liczbie protonów Z. Energia (5) nazywa się energią symetrii.

Udział poszczególnych członów: energii objętościowej B_v, powierzchniowej B_s, kulombowskiej B_{kul} oraz energii symetrii B_{sym} w energii wiązania B jest zilustrowany na rys. 2. Widać z niego, że rola członu powierzchniowego B_s jest największa dla jąder lekkich (bo dla nich największy jest stosunek pola powierzchni do objętości) i maleje ze wzrostem A. Rola członu kulombowskiego zaś szybko rośnie z A, bo razem z A wzrasta ładunek jądra Ze. Także rola członu symetrii B_{sym} rośnie z A, gdyż razem z A wzrasta nadmiar neutronów w jądrze $N-Z$.

Model kroplowy, oprócz energii symetrii, nie opisuje także dwu mniejszych efektów wpływających na energię wiązania. Jeden — to różnica między energią wiązania jąder parzysto-parzystych (N i Z parzyste), nieparzystych (nieparzyste A) i nieparzysto-nieparzystych (N i Z nieparzyste). Wkład tego efektu do energii wiązania można zapisać:

$$B_{pa} = \begin{cases} \Delta & \text{dla jąder parzysto-parzystych,} \\ 0 & \text{—„— nieparzystych,} \\ -\Delta & \text{—„— nieparzysto-nieparzystych.} \end{cases} \tag{6}$$

Δ jest tutaj połową energii rozerwania pary neutronów lub protonów.



Rys. 2. Wkład energii objętościowej, powierzchniowej, kulombowskiej i symetrii do całkowitej (doświadczalnej) energii wiązania na jeden nukleon: B/A

Drugi efekt — to efekt struktury powłokowej, którego przejawem jest fakt, że niektóre jądra (o zamkniętych powłokach) są szczególnie silnie związane. (Na rys. 1 jest to dobrze widoczne jedynie dla najlżejszego jądra o podwójnie zamkniętych powłokach: ^4He (cząstka α), które jest znacznie silniej związane od jąder sąsiednich — lokalne maksimum na krzywej B/A. Dalsze, słabsze maksima przy $A = 8, 12, 16, \ldots$ świadczą o tendencji jąder do „struktury alfowej", tzn. do tego, by nukleony ich wiązały się w szczególnie trwałe grupy — cząstki α). Opisu struktury powłokowej jądra dostarcza oddzielny model — model powłokowy (patrz niżej).

Dopiero suma wszystkich wymienionych efektów

całkowita energia wiązania

$$B = B_v + B_s + B_{kul} + B_{sym} + B_{pa} + \text{ efekty powłokowe} \tag{7}$$

dobrze odtwarza doświadczalną energię wiązania B.

Rysunek 1, choć uproszczony, pozwala jednak zrozumieć podstawę energetyki jądrowej. Zarówno reakcja rozszczepienia jądra ciężkiego, jak i reakcja syntezy jąder lekkich powodują wyzwolenie energii. Obie prowadzą bowiem od jąder słabiej związanych — do silniej. Najsilniej związane są jądra o $A \approx 60$.

Model kroplowy jest także pomocny przy opisie kolektywnych drgań jąder (model kolektywny — patrz niżej) oraz przy opisie samorzutnego rozszczepiania (→ Rozpady jąder atomowych).

Model gazu Fermiego

Model gazu Fermiego jest najprostszą wersją modelu cząstek niezależnych. Przyjmuje się w nim, że nukleony poruszają się niezależnie od siebie i w nieograniczenie dużym obszarze (materia jądrowa — brak efektów powierzchniowych). Jedyna relacja między nimi jest narzucona przez zakaz Pauliego, który mówi, że cząstki identyczne nie mogą zajmować tego samego stanu. Ponieważ stan w modelu gazu Fermiego jest określony przez pęd lub energię kinetyczną cząstki e (poziom energetyczny) i wartość rzutu jej spinu na wybraną oś, to każdy poziom energetyczny może być zajmowany przez dwa protony (o przeciwnych wartościach rzutu spinu) i dwa neutrony.

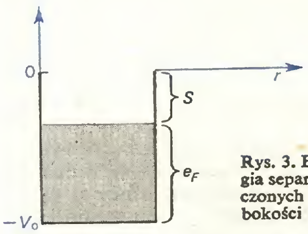
Energia poziomu przebiega ciągłe widmo wartości. W stanie podstawowym cząstki zapełniają poziomy najniższe, od energii $e = 0$ do energii maksymalnej $e = e_F$ (jest to tzw. energia Fermiego, a odpowiadające jej poziom i pęd są odpowiednio poziomem i pędem Fermiego). Wartość energii Fermiego zależy wyłącznie od gęstości cząstek w materii jądrowej, przy czym neutrony i protony traktuje się jako oddzielne gazy Fermiego. Przy równej liczbie neutronów N i protonów Z w jądrze (jądra lekkie), ich gęstości są identyczne, a więc i energie Fermiego są takie same i wynoszą

energia Fermiego

$$e_F^{(n)} = e_F^{(p)} = e_F = \frac{\hbar^2}{2m}(^3/_2\pi^2\varrho)^{2/3}, \tag{8}$$

gdzie m — masa nukleonu, ϱ — gęstość materii jądrowej. Podstawiając wartość doświadczalną $\varrho = 0{,}17\ \text{fm}^{-3}$ dostajemy $e_F \approx 39$ MeV.

Zatem największa energia kinetyczna nukleonu w jądrze e_F jest znacznie mniejsza od jego energii spoczynkowej (ok. 940 MeV). Oznacza to, że poprawki relatywistyczne związane z ruchem są małe i nierelatywistyczna fizyka jądrowa jest dobrym przybliżeniem (z wyjątkiem reakcji jądrowych wysokich energii). Wartość energii Fermiego e_F pozwala nam także ocenić głębokość potencjału jądrowego (rys. 3). Ponieważ wartość doświadczalna energii S



Rys. 3. Energia Fermiego e_F i energia separacji S nukleonów umieszczonych w jamie potencjału o głębokości V_0

potrzebnej do oderwania jednego nukleonu od jądra (energia separacji) wynosi ok. 8 MeV, to głębokość powinna wynosić:

$$V_0 = e_F + S \approx 47 \text{ MeV}.$$

Rzeczywiście, potencjał o takiej mniej więcej głębokości pozwala na odtworzenie wielu wyników doświadczalnych.

W ogólnym wypadku, gdy $N \neq Z$, energie Fermiego neutronów i protonów są różne i wynoszą:

$$e_F^{(n)} = e_F\left(\frac{N}{A/2}\right)^{2/3}; \quad e_F^{(p)} = e_F\left(\frac{Z}{A/2}\right)^{2/3}. \tag{9}$$

Wówczas w całkowitej energii kinetycznej gazu Fermiego

$$E_{\text{kin}} = {}^3/_5[Ne_F^{(n)} + Ze_F^{(p)}] \approx {}^3/_5 e_F A \times \left[1 + {}^5/_9\left(\frac{N-Z}{A}\right)^2\right] \tag{10}$$

pojawia się człon symetrii (5) zwiększający energię kinetyczną, a więc zmniejszający energię wiązania jądra B przy wzroście nadmiaru neutronów $(N-Z)$. Członu tego nie można było otrzymać z modelu kroplowego.

Model powłokowy

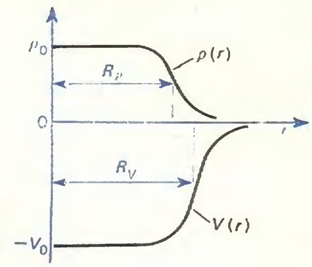
Model kroplowy oraz model gazu Fermiego opisują tylko pewne wartości uśrednione, które zmieniają się w sposób gładki przy zmianie takich parametrów jądra jak liczba nukleonów czy deformacja. Model kroplowy opisuje np. gładką zależność energii wiązania od liczby protonów i neutronów oraz od deformacji. Wszystkie jednak wielkości jądrowe wykazują pewną mikrostrukturę — mniej lub bardziej wyraźne fluktuacje nałożone na te gładkie zależności. Świadczą one o powłokowej strukturze jądra, podobnej do struktury powłok elektronowych. Model powłokowy ma na celu opis tej struktury.

Jednym z pierwszych i ważniejszych argumentów przemawiających za strukturą powłokową jąder było istnienie tzw. liczb magicznych. Stwierdzono mianowicie, że jądra o liczbie protonów lub neutronów równej 2, 8, 20, 28, 50 i 82 oraz liczbie neutronów 126 są szczególnie silnie związane i trwałe. Trudno jest spowodować ich rozpad, a nawet wzbudzenie. Mówimy, że mają one zamkniętą powłokę protonową albo neutronową (lub obie), podobnie jak atomy gazów szlachetnych mają zamkniętą powłokę elektronową.

liczby magiczne

Model powłokowy jest modelem jednocząstkowym. Zakłada on, że nukleony poruszają się niezależnie od siebie w potencjale będącym wynikiem oddziaływania jednego nukleonu ze wszystkimi pozostałymi. Wobec krótkiego zasięgu sił jądrowych, przebieg tego potencjału V powinien być bardzo zbliżony do przebiegu gęstości materii ϱ (i jednocześnie ładunku, gdyż rozkład protonów i neutronów w jądrze jest w przybliżeniu taki sam).

Z danych doświadczalnych wynika, że zależność gęstości ϱ od odległości od środka jądra r ma w przybliżeniu przebieg przedstawiony na rys. 4, tzn wewnątrz jądra gęstość jest stała i szybko spada w re-



Rys. 4. Zależność gęstości materii ρ oraz potencjału V od odległości od środka jądra r

jonie jego powierzchni. Zatem potencjał $V(r)$ powinien mieć przebieg taki jak przedstawiony na tym samym rysunku u dołu, tzn. być proporcjonalny do $\varrho(r)$, tylko z przeciwnym znakiem (potencjał przyciągający) i mieć jedynie promień R nieco większy (mniej więcej o zasięg sił jądrowych) niż $\varrho(r)$.

Rozpatrzymy oddzielnie jądra kuliste i zdeformowane.

jądra kuliste

Potencjałem, którego zależność od odległości r jest podobna do przedstawionej na rys. 4 jest potencjał Woodsa–Saxona. Ma on postać

potencjał Woodsa–Saxona

$$V(r) = \frac{-V_0}{1 + e^{(r-R)/a}}, \tag{11}$$

gdzie R — promień potencjału, a — parametr rozmycia (grubości) powierzchni potencjału.

Stan własny nukleonu umieszczonego w tym potencjale jest scharakteryzowany przez cztery liczby kwantowe: — radialną liczbę kwantową n ($n = 1, 2, 3, \ldots$), która opisuje liczbę oscylacji funkcji falowej w funkcji r (im większa jest ta liczba, przy tych samych wartościach pozostałych liczb kwantowych, tym większa jest energia nukleonu), — orbitalny moment pędu l ($l = 0, 1, 2, \ldots$), — jego rzut na oś kwantyzacji m_l oraz — rzut spinu na tę oś m_s. Ponieważ potencjał jest sferyczny i żaden kierunek kwantyzacji nie jest wyróżniony, to energia stanu nie zależy od m_l. Nie zależy ona także od m_s, gdyż potencjał jest niezależny od kierunku rzutu spinu. Każdy poziom energetyczny jest zatem $2(2l+1)$-krotnie zdegenerowany (ponieważ są dwie możliwe wartości $m_s = -^1/_2, ^1/_2$ i $2l+1$ możliwych wartości $m_l = -l, -l+1, \ldots +l$) i wystarcza oznaczać go liczbami n i l. Dla stanów o $l = 0, 1, 2, 3, 4, 5, 6, 7, \ldots$ używa się powszechnie symboli literowych $s, p, d, f, g, h, i, j, \ldots$

Układ poziomów otrzymany w potencjale Woodsa–Saxona jest przedstawiony na rys. 5b. Są to po-

ziomy neutronowe. Układ poziomów protonowych, przy którego obliczaniu należy uwzględnić oddziaływanie kulombowskie, różni się od niego znacznie tylko w górnej części (poziomy odpowiadające ciężkim jądrom — oddziaływanie kulombowskie większe). Sumując liczby cząstek, które obsadzają kolejne poziomy, łatwo możemy się przekonać, że potencjał Woodsa-Saxona, podobnie jak i prostsze od niego potencjały: — oscylatora harmonicznego (którego szczególnie prosty układ poziomów jest podany dla porównania na rys. 5a)

potencjały oscylatora harmonicznego i jamy prostokątnej

$$V(r) = {}^1/_2 m\omega^2 r^2, \qquad (12)$$

gdzie m — masa nukleonu, a ω — parametr „sztywności" potencjału oraz — jamy prostokątnej

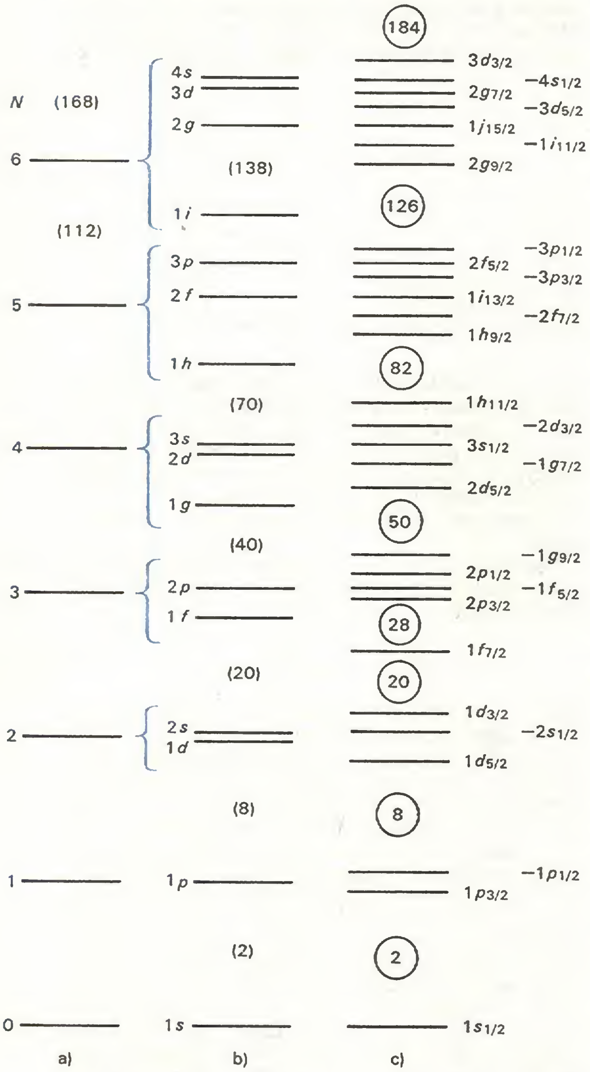
$$V(r) = \begin{cases} -V_0 & \text{dla } r \leq R, \\ 0 & \text{dla } r > R, \end{cases} \qquad (13)$$

gdzie R — promień jamy, odtwarzają jedynie trzy pierwsze liczby magiczne: 2, 8 i 20. Żaden z nich (jak i inne jeszcze potencjały) nie pozwala na odtworzenie liczb dalszych. Aby je odtworzyć należy włączyć do potencjału oddziaływanie spin-orbita, tj. przyjąć potencjał w postaci

potencjał oddziaływania spin-orbita

$$V(r) + V_{so}(r)\vec{l}\cdot\vec{s}, \qquad (14)$$

gdzie $\vec{l}$ — wektor orbitalnego momentu pędu nukleonu, $\vec{s}$ — wektor jego spinu. Spinowo-orbitalna część potencjału V_{so} pochodzi ze spinowo-orbitalnej składowej sił jądrowych i jest największa w obszarze powierzchni jądra, w którym zachodzi najszybsza zmiana gęstości materii ϱ. Gdy przyjmie się taki potencjał, liczby m_l i m_s nie są już dobrymi liczbami kwantowymi; liczbami takimi pozostają jednak: pełny moment pędu nukleonu $j = l \pm {}^1/_2$ oraz jego rzut na oś kwantyzacji $m = m_l + m_s$. Wymienienie tych liczb, obok liczb n i l, w pełni określa stan. Sama energia stanu nie zależy od m ($m = -j, -j+1, \ldots, j$) i wobec tego degeneracja każdego poziomu energetycznego wynosi $(2j+1)$. Układ poziomów jest podany na rys. 5c. Widać z niego, że każdy poziom (z wyjątkiem poziomów o $l = 0$) zostaje rozszczepiony na dwa: $j = l + {}^1/_2$ i $j = l - {}^1/_2$. Dane doświadczalne wykazują, że poziom o większym j leży niżej, co oznacza, że oddziaływanie spin-orbita ma charakter przyciągający. Z rys. 5c wynika także, że wzięcie pod uwagę tego oddziaływania pozwala na odtworzenie wszystkich znanych doświadczalnie liczb magicznych.



Rys. 5. Układ poziomów neutronowych otrzymanych w potencjale: oscylatora harmonicznego (a), Woodsa-Saxona bez sprzężenia spin-orbita (b) oraz ze sprzężeniem spin-orbita (c). Liczby z lewej strony oznaczają numer powłoki oscylatorowej N. Nawiasy klamrowe pokazują, które poziomy potencjału Woodsa-Saxona powstają przez rozszczepienie poziomu oscylatora harmonicznego. W nawiasach lub w kółkach podane są liczby wszystkich cząstek zapełniających poziomy położone poniżej

Stosując model powłokowy możemy odtworzyć własności jąder w stanie podstawowym oraz w stanach niskowzbudzonych. Odtworzenie to jest najbardziej bezpośrednie dla jąder nieparzystych, których wiele własności (spin, parzystość, moment magnetyczny, energia wzbudzenia) jest z reguły własnościami stanu zajętego przez nieparzysty nukleon. Opis tych własności jest najlepszy dla jądra podwójnie magicznego plus lub minus jeden nukleon. Jądro podwójnie magiczne może być uważane wtedy za szczególnie sztywny, niełatwo ulegający wzbudzeniu rdzeń i o własności jądra nieparzystego decyduje nukleon „nadwyżkowy" lub brakujący (tzw. „dziura") w zamkniętej powłoce. Na przykład, w jądrze ^{209}Pb rdzeń stanowią 82 protony oraz 126 neutronów i ostatni 127. neutron decyduje o własnościach jądra. Zgodnie z rys. 5c, stan podstawowy tego jądra powinien mieć spin $^9/_2$ i parzystość dodatnią (+), gdyż wówczas ostatni neutron obsadza najniższy dostępny poziom $2g_{9/2}$ ($l = 4$, tzn. parzyste). Najniższe zaś stany wzbudzone powinny być: $^{11}/_2+$, $^{15}/_2-$, $^5/_2+$. Z doświadczenia rzeczywiście otrzymuje się zarówno stan podstawowy, jak i najniższe wzbudzone zgodne z tymi przewidywaniami.

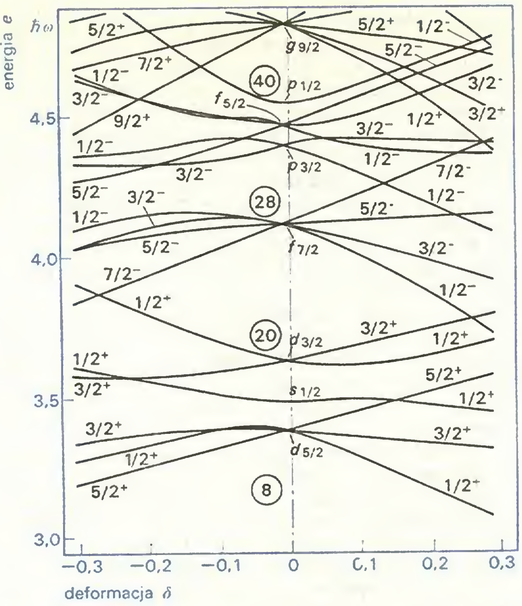
jądra zdeformowane

Oprócz jąder kulistych istnieją jądra zdeformowane. Wykazują one trwałą, statyczną deformację, a kształt ich jest zbliżony do kształtu wydłużonej elipsoidy obrotowej (kształt cygara — por. rys. 7). Oczekuje się ponadto, że niektóre jądra nietrwałe mogą mieć kształt elipsoidy spłaszczonej (kształt dysku).

Do opisu ruchu nukleonów w jądrze zdeformowanym, jest potrzebny potencjał zdeformowany. Może to być np. potencjał Woodsa-Saxona, ale o powierzchniach ekwipotencjalnych odpowiednio zdeformowanych. Potencjał taki jest często stosowany w obliczeniach. Znacznie prostsze jednak są obliczenia za pomocą potencjału Nilssona, który jest nieco zmodyfikowanym potencjałem oscylatora harmonicznego, zawierającym oddziaływanie spin-orbita. Modyfikacja jest na tyle zręczna, że mimo uproszczenia potencjał jest realistyczny. Model powłokowy oparty na potencjale Nilssona nazywa się modelem Nilssona i jest szczególnym, wciąż bardzo powszechnie stosowanym modelem powłokowym dla jąder zdeformowanych.

model Nilssona

Przy potencjale zdeformowanym, całkowity moment pędu nukleonu j nie jest już dobrą liczbą kwantową. Liczbami takimi pozostają jedynie parzystość oraz rzut całkowitego momentu pędu na oś symetrii jądra. Energia stanu zależy teraz od deformacji jądra.



Rys. 6. Przykład zależności energii jednocząstkowej do deformacji jądra δ dla kilkunastu poziomów (fragment wykresu Nilssona). Przy każdym poziomie podany jest rzut pełnego momentu pędu tego poziomu na oś symetrii jądra oraz jego parzystość

Przykład takiej zależności jest podany na rys. 6, na którym wartościom dodatnim parametru deformacji odpowiada wydłużony kształt jądra, a ujemnym — spłaszczony. Widać, że poziomy przecinają się. Ich układ, kolejność zmieniają się wraz z deformacją. Jeśli znamy zatem deformację jądra, to za pomocą

wykres Nilssona

wykresu poziomów (zwanego wykresem lub diagramem Nilssona) i — ogólnie — za pomocą modelu Nilssona możemy odtworzyć lub przewidzieć własności jądra w jego stanie podstawowym i stanach niskowzbudzonych, podobnie jak dla jąder sferycznych. Jeśli nie wiemy jaka jest deformacja, ale znamy inne własności jądra, to możemy postąpić odwrotnie: z własności tych wnosić o jego deformacji.

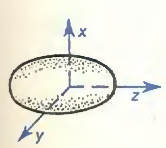
Przy kształcie kulistym (deformacja $\delta = 0$) poziomy wykresu z rys. 6 pokrywają się z poziomami z rys. 5c. Deformacja usuwa częściowo degenerację poziomów, pozostaje jeszcze tylko dwukrotna degeneracja każdego poziomu.

Model kolektywny

W jądrze atomowym poza zjawiskami jednocząstkowymi występują zjawiska kolektywne, w których bierze udział, w sposób skorelowany, wiele cząstek. Model kolektywny ma za zadanie opisanie tych zjawisk; jest on modelem cząstek silnie skorelowanych. Przykładami zjawisk kolektywnych są: ruch obrotowy (rotacja) i drgania (wibracje) jądra. Modele opisujące oddzielnie rotację i wibracje nazywane są odpowiednio rotacyjnym i wibracyjnym.

model rotacyjny

Jak wspomniano wyżej, niektóre jądra atomowe są trwale zdeformowane. Przykładem są jądra pierwiastków ziem rzadkich oraz aktynowców. Kształt ich jest (w dobrym przybliżeniu) elipsoidą wydłużoną, osiowo symetryczną (rys. 7).

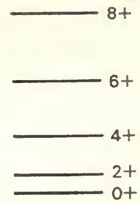


Rys. 7. Jądro zdeformowane

Obrót tak zdeformowanego jądra względem osi prostopadłej (np. osi $0x$) do osi symetrii ($0z$ na rys. 7) prowadzi w mechanice kwantowej do następującego widma energii:

$$E = \frac{\hbar^2}{2J} I(I+1), \qquad (15)$$

gdzie I — moment pędu jądra, a J — jego moment bezwładności względem osi obrotu. Moment pędu I może przybierać wartości: $I = 0, 2, 4, 6, \ldots$; widmo (15) (zwane widmem rotacyjnym) jest pokazane na rys. 8. Takie charakterystyczne widma zaobserwowano u parzysto-parzystych jąder zdeformowanych. Model rotacyjny został właśnie wynaleziony do ich wyjaśnienia.

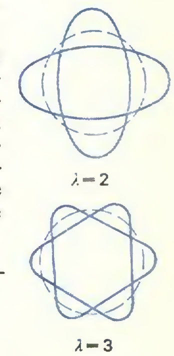


Rys. 8. Widmo rotacyjne jądra zdeformowanego

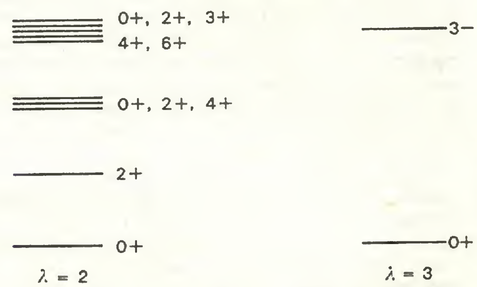
model wibracyjny

Wśród różnych wzbudzeń jąder obserwujemy takie, które możemy powiązać z ich drganiami. Szczególnie prosty obraz dostajemy przyjmując, że jądro jest kroplą nieściśliwej cieczy, której drgania możemy opisać drganiami jej powierzchni. Rysunek 9 podaje kształty powierzchni jądra sferycznego przy dwóch rodzajach drgań: kwadrupolowym (kształty elipsoidalne) i oktupolowym (kształty typu gruszki), prowadzących do najniższych energii wzbudzenia.

W modelu kroplowym widmo wzbudzeń odpowiadające drganiom ma postać jak na rys. 10. Przy drganiach kwadrupolowych najniżej położony jest poziom 2+; na wysokości dwukrotnie wyższej znajdują się trzy poziomy (tryplet) — 0+, 2+, 4+; na potrójnej wysokości — pięć poziomów, a wyżej — układy coraz bardziej złożone. Na wysokości tych pięciu poziomów znajduje się dopiero pierwszy poziom wzbudzony oktupolowy 3−, a poziom następny — dopiero na wysokości podwójnej. Rzeczywiste widma niektórych jąder sferycznych wykazują pewne



Rys. 9. Kształty powierzchni jądra wykonującego drgania kwadrupolowe (multipolowość $\lambda = 2$) i oktupolowe ($\lambda = 3$) wokół kształtu kulistego odpowiadającego stanowi równowagi



Rys. 10. Widmo wibracyjne najniższych wzbudzeń kwadrupolowych ($\lambda = 2$) i oktupolowych ($\lambda = 3$), otrzymane w modelu kroplowym

cechy widm z rys. 10. Na ogół jednak obraz jest bardziej złożony.

Podobnie można opisywać wzbudzenia wibracyjne jąder zdeformowanych.

Model uogólniony

Model uogólniony (zwany też połączonym, scalonym lub zunifikowanym) jest połączeniem modelu jednocząstkowego (powłokowego) z modelem kolektywnym. Dopiero w ramach tego modelu można opisać w sposób stosunkowo pełny i konsekwentny własności jąder.

Za najprostszy model uogólniony można uważać omówiony wcześniej model Nilssona. By opisać bowiem strukturę jednocząstkową jądra zdeformowanego musi on zawierać deformację jądra, która jest wielkością kolektywną. Z tego punktu widzenia, opis tego modelu można by umieścić na początku niniejszego rozdziału, a nie w rozdziale poświęconym modelowi powłokowemu. Jednak w modelu Nilssona rola parametru kolektywnego — deformacji — jest na tyle tylko pomocnicza, że bardziej naturalne wydaje się traktowanie go jako modelu powłokowego niż uogólnionego. Od typowego modelu uogólnionego wymagamy bowiem więcej, mianowicie — by opisywał on zarówno własności jednocząstkowe, jak i kolektywne lub też by opisywał własności kolektywne w sposób pełny (mikroskopowy — jak często się mówi), tj. oparty o wewnętrzną, jednocząstkową, mikroskopową strukturę jądra.

Możliwości i rolę modelu uogólnionego zilustrujemy trzema przykładami. Jako pierwszy weźmiemy

opis jąder nieparzystych

jądro nieparzyste, które jest zwykle traktowane jako układ jednej cząstki dodanej do parzysto-parzystego rdzenia. Najniższe wzbudzenia takiego układu są pewną superpozycją wzbudzeń kolektywnych rdzenia i wzbudzeń cząstkowych cząstki nieparzystej umieszczonej w polu rdzenia. Widma tych wzbudzeń są więc pewną kombinacją widma kolektywnego (typu przedstawionego na rys. 8 lub 10) i jednocząstkowego (typu przedstawionego na rys. 5 lub 6); mogą zatem być opisane w pełni dopiero przez model uogólniony.

rozwinięcie modelu rotacyjnego

Drugiego przykładu może dostarczyć model rotacyjny, który jest szczególnym przypadkiem modelu kolektywnego. We wzorze (15), który opisuje energie ruchu obrotowego jądra zdeformowanego występuje parametr J — moment bezwładności jądra. Dopóki pozostajemy w ramach samego modelu rotacyjnego, parametr ten jest czysto fenomenologiczny. Możemy wyznaczyć jego wartość z doświadczenia porównując zmierzoną energię stanu rotacyjnego o danym momencie pędu I z energią obliczoną ze wzoru (15). Nie możemy jednak powiązać tej wartości z wewnętrzną strukturą jądra, o której w czystym modelu rotacyjnym nic nie wiemy. Powiązanie takie jest jednakże bardzo ważne, gdyż pozwala wykorzystać wiedzę doświadczalną o widmach jąder obracających się (w szczególności, zdobywaną ostatnio wiedzę o widmach jąder szybko obracających się → Jądra atomowe w stanach ekstremalnych) do sprawdzenia i wzbogacenia wiedzy o strukturze jądra. Dokonać tego możemy dopiero w modelu uogólnionym, który obejmuje wewnętrzną strukturę jądra. Korzystając z tego modelu możemy wykonać mikroskopowe obliczenia momentu bezwładności, który w sposób istotny zależy od szczegółów tej struktury.

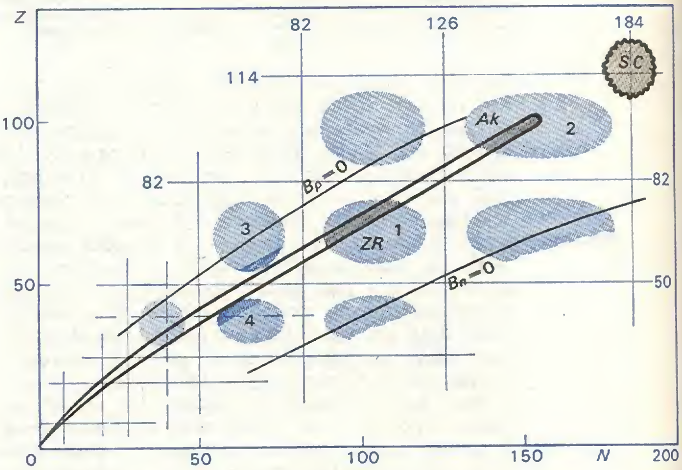
energia całkowita jądra

Trzecią ilustracją potrzeby i możliwości modelu uogólnionego jest stosowany najczęściej sposób obliczania energii całkowitej jądra, a zwłaszcza jej zależności od deformacji. Znajomość tej zależności pozwala wyznaczyć statyczną deformację (deformację równowagi) jądra, obliczyć barierę potencjału na rozszczepienie, energię wzbudzeń kolektywnych itd. Obecnie energię jądra oblicza się najczęściej jako sumę energii kropli cieczy (model kroplowy jądra, a więc jedna z najprostszych wersji modelu kolektywnego) oraz poprawki powłokowej. Poprawka ta uwzględnia niejednorodny, powłokowy charakter wewnętrznej struktury. Do jej wyznaczenia jest potrzebny jednocząstkowy, powłokowy model tej struktury (np. model Nilssona), przy interesującej nas deformacji jądra. Widzimy więc, że obliczenie energii jest dopiero możliwe w pełnym modelu uogólnionym.

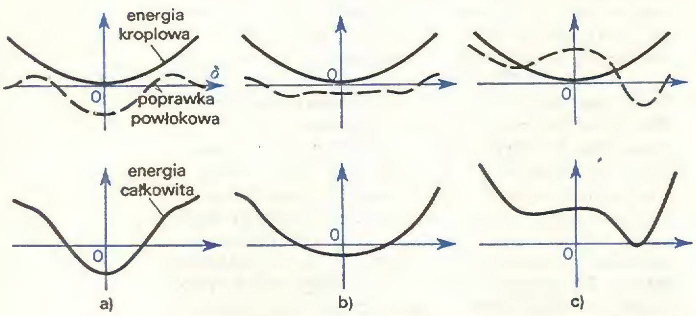
Przyjrzyjmy się nieco wynikom takich obliczeń. Są one pokazane na rys. 11, który przedstawia orientacyjnie zależność energii kroplowej, poprawki powłokowej oraz energii całkowitej od deformacji kwadrupolowej δ, opisującej elipsoidalny kształt jądra (tej samej co na rys. 6). Rozważone są trzy wypadki: jądra o zamkniętych powłokach (a), jądra o niewielu nukleonach poza zamkniętymi powłokami (b) oraz jądra o dużej liczbie nukleonów poza zamkniętymi powłokami — równej w przybliżeniu połowie liczby wszystkich nukleonów w powłokach (neutronowej i protonowej) niezapełnionych (c). Widoczne jest, że w wypadku a) poprawka powłokowa ma głębokie minimum przy kształcie kulistym ($\delta = 0$), co powoduje, że całkowita energia ma także minimum przy $\delta = 0$ i szybko rośnie ze wzrostem deformacji (wydłużaniem lub spłaszczaniem kształtu elipsoidalnego). Oznacza to, że w stanie równowagi jądro takie jest kuliste i trudno jest je zdeformować; jest ono mało podatne („sztywne") na deformację. W wypadku b) jądro jest nadal kuliste w stanie równowagi, ale jest już bardziej podatne („miękkie") na deformację, jest mu łatwo oscylować wokół kształtu kulistego. W wypadku c) jądro jest zdeformowane w stanie równowagi i znowu stosunkowo mało podatne na odkształcanie od kształtu równowagi.

obszary występowania jąder zdeformowanych

Wynika stąd, gdzie na mapie nuklidów możemy oczekiwać jąder zdeformowanych. Jądra te będą grupowały się wokół środków prostokątów utworzonych przez linie $Z = Z_{mag}$ i $N = N_{mag}$, odpowiadające protonowym i neutronowym liczbom magicznym (rys. 12). Obszary jąder zdeformowanych będą tym większe, im większe są powłoki; przy małych powłokach mogą wcale nie wystąpić. Na rys. 12 cztery największe obszary jąder zdeformowanych są oznaczone numerami 1–4. Tylko dwa z nich są tak usytuowane, że niemal przez środek ich przechodzi ścieżka trwałości β. Są to: obszar ziem rzadkich *1* i aktynowców *2*. W tych obszarach występują zatem jądra zdeformowane trwałe (ziemie rzadkie) lub długożyciowe (część aktynowców) i one właśnie zostały zbadane jako pierwsze spośród jąder zdeformowanych.



Rys. 12. Położenie znanych i przewidywanych obszarów jąder zdeformowanych na mapie nuklidów. Przez obszary *1* (ziemie rzadkie — *ZR*) i *2* (aktynowce — *Ak*) przechodzi ścieżka trwałości β. Na skraj obszaru *3* (jądra neutronodeficytowe) prowadzą reakcje z ciężkimi jonami (obszar ciemnoniebieski) a do obszaru *4* (jądra neutrononadmiarowe) — reakcje rozszczepienia jąder ciężkich. Jądra położone poza liniami zerowej energii wiązania neutronu ($B_n = 0$) lub protonu ($B_p = 0$) nie mogą istnieć; rozpadałyby się natychmiast lub prawie natychmiast po ich utworzeniu. Zaznaczone jest także położenie hipotetycznej wyspy (kulistych) jąder superciężkich (*SC*), usytuowanej na przecięciu linii przewidywanych liczb magicznych $Z_{mag} = 114$ i $N_{mag} = 184$



Rys. 11. Zależność gładkiej (kroplowej) części energii, poprawki powłokowej oraz energii całkowitej od deformacji kwadrupolowej δ dla jądra o zamkniętych powłokach (a), niewielkiej liczbie nukleonów poza zamkniętymi powłokami (b) oraz znacznej ich liczbie (c)

Rozwój modeli jądrowych

Modele podstawowe rzadko są stosowane oddzielnie. Widzieliśmy np., że już do bardzo przybliżonego opisu energii wiązania jąder, oprócz modelu kroplowego był także potrzebny model gazu Fermiego. W celu

dokładniejszego zaś opisu trzeba było jeszcze dodać (por. wzór 7) model przedstawiający nukleony w jądrze jako powiązane w pary oraz model powłokowy. Drugim przykładem potrzeby kojarzenia modeli jest przedstawiony powyżej i często stosowany model uogólniony, który jest połączeniem modelu kolektywnego i powłokowego. Ogólnie, im więcej chcemy objąć faktów lub dokładniej je opisać, tym bardziej złożonego modelu musimy używać.

Już same modele podstawowe, opisane w rozdziale pierwszym, z reguły nie mają obecnie tak prostej postaci, w jakiej je przedstawiliśmy wcześniej. Na przykład, przy stosowaniu modelu powłokowego musimy na ogół uwzględnić oddziaływanie resztkowe (szczątkowe). Jest to część oddziaływania między nukleonami, która nie została włączona do potencjału modelu powłokowego. W oddziaływaniu resztkowym występuje m.in. składowa o stosunkowo krótkim zasięgu działania, która łączy nukleony w pary. Uwzględnienie jej pozwala objąć modelem powłokowym wspomniany już model przedstawiający nukleony w jądrze jako powiązane parami. Taki model nukleonów (połączonych w pary) jest stosowany zarówno do dokładniejszego opisu energii wiązania (por. wzór 6), jak też do ulepszenia opisu wielu innych wielkości jądrowych, np. momentu bezwładności.

potencjał samozgodny

Nawet jeśli nie uwzględnia się oddziaływań resztkowych, to jednak takiemu „czystemu" modelowi powłokowemu stawia się obecnie często wyższe wymagania niż dawniej. Wymaga się mianowicie, by potencjał stosowany w tym modelu był samozgodny, tj. zgodny z rozkładem gęstości nukleonów, które go wytwarzają. Wyjaśnimy to dokładniej. Jeśli weźmiemy np. potencjał Woodsa–Saxona, to rozwiązując równanie Schrödingera opisujące ruch nukleonów w tym potencjale, możemy znaleźć rozkład gęstości umieszczonych w nim nukleonów. Następnie, znając już tę gęstość i znając oddziaływanie między nukleonami, możemy wyznaczyć potencjał, który one wytwarzają. Przekonamy się, że potencjał ten różni się trochę od wyjściowego potencjału Woodsa–Saxona. Po poprawieniu go o tę różnicę możemy znaleźć nowy rozkład gęstości, a z niego nowy potencjał itd. Dopiero po kilku lub kilkunastu takich krokach (zależnie od tego, jak dobrze był odgadnięty potencjał wyjściowy) rozkład gęstości obliczony w danym potencjale będzie odtwarzał go na powrót; oznacza to właśnie, że znaleziony został potencjał samozgodny. Postępowanie według takiego schematu nazywa się metodą Hartree'ego–Focka.

metoda Hartree'ego–Focka

Powyższy przykład pokazuje, jak ulepszenie modelu, nie naruszając podstawowych jego założeń (w tym wypadku — niezależność ruchu nukleonów), czyni go bardziej złożonym. W pierwszym modelu powłokowym, potencjał przyjmowało się (odgadywało) w możliwie prostej i poglądowej postaci. W metodzie zaś Hartree'ego–Focka odnajduje się go za pomocą dosyć złożonej procedury obliczeniowej. Obecnie coraz częściej prosty dawniej model zostaje zastąpiony dość złożonym postępowaniem obliczeniowym. Model traci wtedy swoją poglądowość na tyle, że przestaje być nazywany modelem. Niektórzy fizycy nie mówią już np. o modelu powłokowym, a tylko o strukturze powłokowej jądra, o opisie lub o teorii tej struktury.

wpływ zakazu Pauliego

Ważnym etapem rozwoju każdego modelu było uzasadnienie jego sukcesu, szczególnie, gdy sukces ten był nieoczekiwany. Przykładem tego jest znowu model powłokowy. Ponieważ siły jądrowe mają krótki zasięg (niewiele przekraczający rozmiary nukleonu), i wobec tego jądro może zostać związane tylko przy dość ścisłym „upakowaniu" nukleonów, sukces modelu powłokowego wydał się dziwny. Model ten przyjmuje bowiem swobodny ruch nukleonów w jądrze; wymaga zatem, by średnia droga swobodna nukleonu była równa lub większa od rozmiarów jądra, a więc większa niż średnia odległość między nukleonami. Zagadkę jego sukcesu rozwiązało wykrycie wpływu zakazu Pauliego na ruch nukleonów. Otóż zakaz ten wzbraniając nukleonom zajmowania poziomów zajętych już przez inne, sprawia, że nukleony w jądrze „odczuwają" tylko niektóre ze zderzeń między sobą; mianowicie tylko zderzenia silne, które mogą „wybić" je ponad „morze" stanów zajętych. Inne, słabsze zderzenia nie mogą zmienić stanu nukleonów i nie są zatem przez nie odczuwalne. Jest to równoważne efektowi wydłużenia drogi swobodnej.

Encyklopedia fizyki, t. 1, str. 843, Warszawa 1972; B.L. Cohen *Concepts of nuclear physics*, New Delhi 1975; H. Frauenfelder, E.M. Henley *Subatomic physics*, Englewood Cliffs 1974; A. Sobiczewski *Potencjał jednocząstkowy dla jąder zdeformowanych*, Post. Fiz. 20, 649 (1969); A. Strzałkowski *Wstęp do fizyki jądra atomowego*, Warszawa 1978; Sz. Szczeniowski *Fizyka doświadczalna*, cz. 6, Warszawa 1974; Z. Wilhelmi *Fizyka reakcji jądrowych*, Warszawa 1976.

Rozpady jąder atomowych

Adam Sobiczewski

Jądro atomowe może znajdować się w stanie podstawowym, tj. w stanie, w którym ma najniższą energię całkowitą (a więc i najmniejszą masę), lub w stanie wzbudzonym.

jądro wzbudzone

Jądro w stanie wzbudzonym ulega rozpadowi. Może to być rozpad, który zmienia skład jądra (np. rozpad β, α, rozszczepienie), jak i rozpad, który składu tego nie zmienia (rozpad γ, konwersja wewnętrzna). W ostatnim przypadku jądro przechodzi ze stanu o wyższej energii wzbudzenia do stanu o energii niższej (deekscytacja), w szczególności do stanu podstawowego. W stanie podstawowym jądro bądź w ogóle nie rozpada się (jądra trwałe), bądź ulega rozpadowi zmieniającemu jego skład, czyli prowadzącemu do innego już jądra.

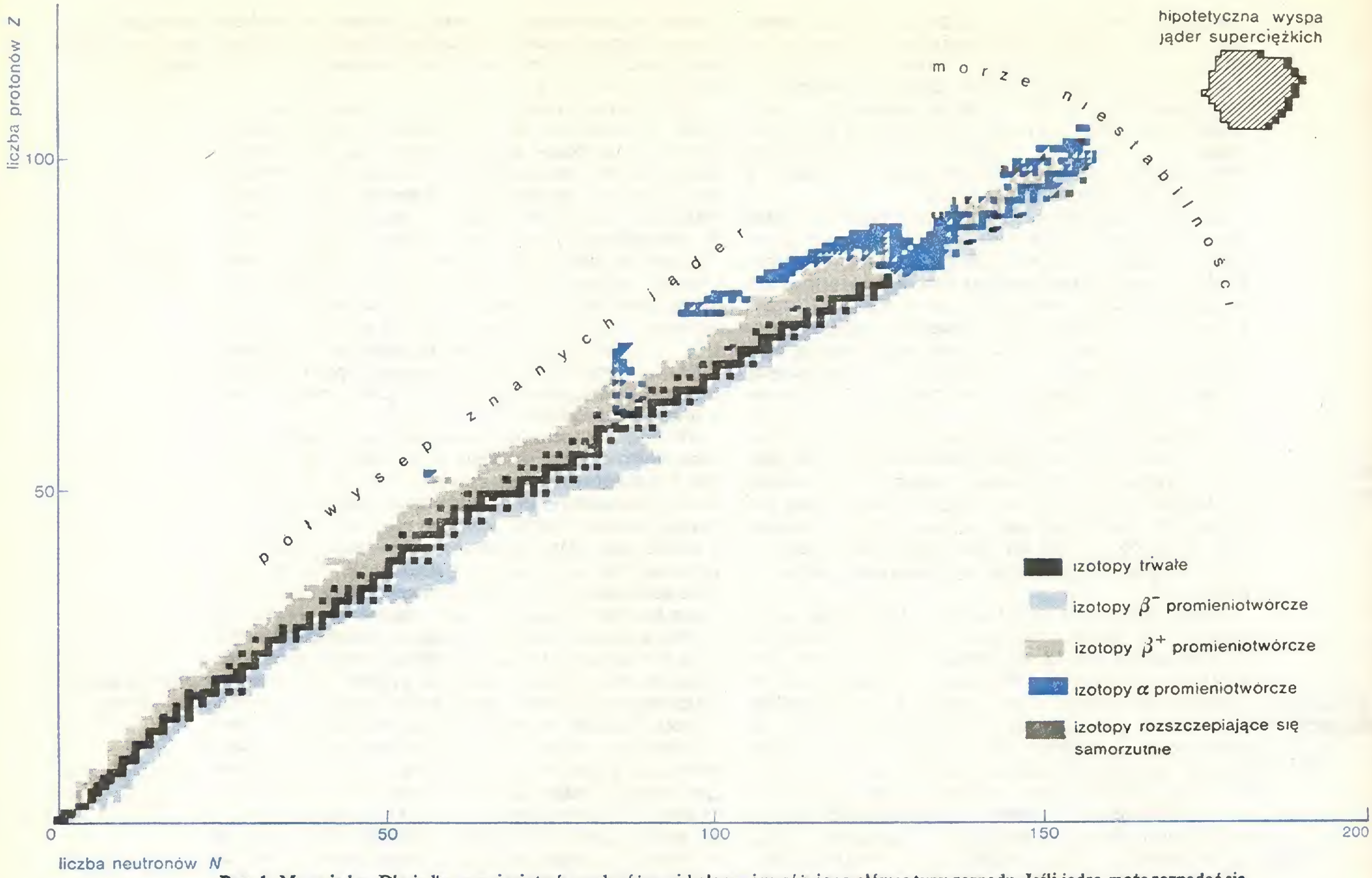
izomery jądrowe

Niektóre stany wzbudzone są szczególnie wolno rozpadające się (stany długożyciowe, metatrwałe). Nazywamy je stanami izomerycznymi jądra, a samo jądro znajdujące się w tym stanie — izomerem. Izomer jądrowy różni się zatem od jądra w zwykłym stanie nie składem jądra, lecz jego strukturą (podobnie jak w przypadku izomeru chemicznego), co oznaczamy przez dodanie litery *m* (od metatrwały) przy liczbie masowej. Na przykład ^{115m}In oznacza izomer jądra ^{115}In.

egzoenergetyczność rozpadu

Warunkiem koniecznym każdego rozpadu jest, by wydzielona w nim była pewna ilość energii (proces egzoenergetyczny), tzn., by masa wszystkich produktów rozpadu była mniejsza od masy układu wyjściowego. Wśród rozpadów jądrowych najważniejsze są: rozpady β i α, rozszczepienie oraz rozpad γ i konwersja wewnętrzna. Rozpad β zachodzi wskutek oddziaływań słabych, rozpad γ i konwersja wewnętrzna wskutek oddziaływań elektromagnetycznych, a w rozpadzie α i rozszczepieniu istotną rolę odgrywają oddziaływania silne (→ Oddziaływania słabe, Oddziaływania elektromagnetyczne, Oddziaływania silne).

Obecnie znamy około 2000 różnych jąder. Położenie ich na mapie jąder, tj. na wykresie, na którym na jednej osi odłożona jest liczba protonów Z, a na drugiej — liczba neutronów N jądra, ilustruje rys. 1. Są to w większości jądra nietrwałe, rozpadające się (promieniotwórcze), wytworzone sztucznie w różnego rodzaju reakcjach jądrowych. Na przykład w drodze reakcji rozszczepienia jąder powstaje wiele izotopów bogatych w neutrony (neutrono-nadmiarowych),



Rys. 1. Mapa jąder. Dla jąder promieniotwórczych różnymi kolorami wyróżniono główne typy rozpadu. Jeśli jądro może rozpadać się na różne sposoby, np. na drodze rozpadu α, β^+, konwersji wewnętrznej i rozszczepienia samorzutnego, ale rozpad α występuje najczęściej, to oznaczone jest ono na mapie jako jądro α-promieniotwórcze. Widoczne jest, że dla większości jąder ciężkich dominuje rozpad α, chociaż rozszczepienie odgrywa dla tych jąder także ważną rolę (wg Delta nr 5, 1975)

a w drodze reakcji z ciężkimi jonami (→ Fizyka ciężkich jonów) — wiele izotopów ubogich w neutrony (neutrono-deficytowych). Wraz z postępem technicznym ciągle wzrasta możliwość przeprowadzania różnych reakcji, a tym samym możliwość poznania nowych jąder nietrwałych.

jądra trwałe i promieniotwórcze

Trwałych jąder jest jedynie 264. W przyrodzie oprócz tych 264 jąder trwałych występuje pewna liczba jąder promieniotwórczych (promieniotwórczość naturalna). Przykładami naturalnych jąder promieniotwórczych są jądra: izotop toru ^{232}Th i izotopy uranu ^{238}U i ^{235}U. Mają one długie okresy rozpadu, porównywalne z wiekiem Ziemi (dokładniej, z wiekiem najstarszych skał skorupy ziemskiej, ocenianym na ok. 5 miliardów lat) i dlatego mogły przetrwać w skorupie ziemskiej w ilościach obserwowalnych. Rozpadając się, dają początek całym szeregom (rodzinom, rys. 1, str. 245) promieniotwórczym. Każdy szereg składa się z jąder promieniotwórczych, które powstają w drodze kolejnych rozpadów α lub β jednego z wymienionych jąder wyjściowych, i kończy się na jądrze trwałym. Szeregi powstające z jąder: ^{232}Th, ^{238}U i ^{235}U nazywają się szeregami: torowym, uranowym i aktynowo-uranowym i kończą się odpowiednio na jądrach trwałych: izotopach ołowiu ^{208}Pb, ^{206}Pb i ^{207}Pb.

Ogólne własności rozpadu jąder

Wytwarzamy obecnie bardzo wiele jąder i w bardzo różnych stanach. Dobierając odpowiednio rodzaj i parametry reakcji jądrowej możemy wytwarzać jądra w określonym stanie. Z chwilą jednak, gdy jądro jest już utworzone, nie mamy praktycznie wpływu na jego rozpad. Rozpad odbywa się samorzutnie, spontanicznie. Istnieje określone prawdopodobieństwo λ, właściwe dla danego jądra i stanu, w którym się ono znajduje, że rozpadnie się ono w jednostce czasu. Ilość wszystkich rozpadów w tej jednostce czasu równa jest zatem liczbie wszystkich jąder pomnożonej przez to prawdopodobieństwo:

prawo rozpadu promieniotwórczego

$$dN/dt = -\lambda N. \qquad (1)$$

Znak minus we wzorze (1) oznacza, że wskutek rozpadu liczba jąder maleje. Wzór (1) nazywa się prawem rozpadu promieniotwórczego. Prawo to określa zależność liczby jąder N od czasu t; ma ona postać wykładniczą:

$$N(t) = N(0)e^{-\lambda t}. \qquad (2)$$

$N(0)$ oznacza liczbę jąder w chwili początkowej $t = 0$, w której zaczynamy obserwację. Prawo rozpadu (1) jest równaniem różniczkowym, którego rozwiązaniem jest funkcja $N(t)$ podana wzorem (2).

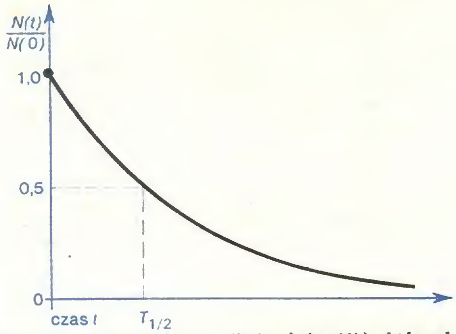
Prawo (1) opisuje rozpad nie tylko jąder, ale wszelkich obiektów, których prawdopodobieństwo rozpadu na jednostkę czasu jest stałe, niezależne ani od chwili, w której zachodzi, ani od liczby obiektów. Podlega mu np. także rozpad nietrwałych cząstek elementarnych.

Prawdopodobieństwo λ, zwane także stałą rozpadu, określa szybkość rozpadu. Szybkość tę często charakteryzuje się także czasem (okresem) połowicznego zaniku (połowicznego rozpadu) $T_{1/2}$, tj. czasem, w którym liczba jąder maleje, wskutek rozpadu, do połowy. Wyraża się on przez λ następująco:

czas połowicznego zaniku

$$T_{1/2} = \frac{\ln 2}{\lambda}. \qquad (3)$$

Rysunek 2 ilustruje zależność (2) liczby jąder $N(t)$ od czasu. Zaznaczony jest także na nim czas połowicznego zaniku $T_{1/2}$.



Rys. 2. Zależność od czasu liczby jąder $N(t)$, które do chwili t jeszcze się nie rozpadły, odniesionej do liczby jąder w chwili początkowej $N(0)$. Zaznaczony jest okres połowicznego zaniku $T_{1/2}$

aktywność próbki

Liczbę rozpadających się jąder w danej próbce promieniotwórczej w jednostce czasu nazywa się jej aktywnością. Zgodnie z (1) liczba ta jest równa λN. Zatem aktywność każdej próbki spada, zgodnie z (2), wykładniczo z czasem.

Jeśli jądro będące w danym stanie ulega kilku różnym rozpadom, np. rozpadowi β, α i in., to całkowite prawdopodobieństwo λ równe jest sumie prawdopodobieństw λ_i poszczególnych rozpadów, tzn.

$$\lambda = \sum_{i=1}^{n} \lambda_i, \tag{4}$$

gdzie n jest liczbą możliwych rozpadów. Wobec relacji (3) wzór ten możemy przepisać jako związek pomiędzy całkowitym czasem połowicznego zaniku (zapisanym krótko jako T, zamiast $T_{1/2}$), a czasami połowicznego zaniku T_i ze względu na poszczególne rozpady:

$$\frac{1}{T} = \sum_{i=1}^{n} \frac{1}{T_i}. \tag{5}$$

Ze wzoru (5) wynika, że najważniejszy jest rozpad o najmniejszym T_i, który decyduje o całkowitym czasie T. Zgodnie z (4) aktywność λN jest sumą aktywności odpowiadających poszczególnym rodzajom rozpadu:

$$\lambda N = \sum_{i=1}^{n} \lambda_i N. \tag{6}$$

Rozpad α

Rozpad α polega na wyrzuceniu z jądra cząstki α, tj. jądra ${}^4_2He$, które wśród jąder lekkich jest jądrem szczególnie silnie związanym. Rozpad α możemy zapisać symbolicznie jako

$${}^{A}_{Z}X_N \rightarrow {}^{A-4}_{Z-2}Y_{N-2} + \alpha.$$

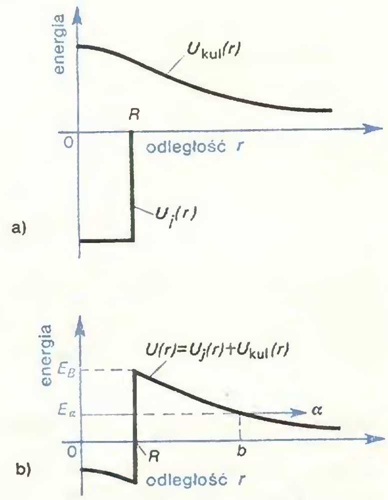
Obecnie znamy ok. 370 jąder ulegających rozpadowi α ze stanu podstawowego, w tym dla ok. 260 z nich jest to rozpad główny, tzn. rozpad, którego prawdopodobieństwo λ_α jest większe od prawdopodobieństw rozpadu innego rodzaju.

prawdopodobieństwo rozpadu α

Prawdopodobieństwo rozpadu α bardzo silnie zależy od energii tego rozpadu Q_α — jest tym większe, im większa jest energia Q_α. Energia rozpadu (równa praktycznie energii kinetycznej emitowanej cząstki α, E_α) mierzona we wszystkich obserwowanych rozpadach mieści się w granicach ok. 2–9 MeV. Odpowiadający tej energii czas połowicznego zaniku T_α jest w granicach od ok. $8 \cdot 10^{15}$ lat (${}^{142}_{62}Sm$) do ok. $3 \cdot 10^{-7}$ s (${}^{212}_{84}Po$). Oznacza to, że różnicy energii rozpadu Q_α, wynoszącej ok. 7 MeV, odpowiada ogromna różnica czasów T_α sięgająca ok. 29 rzędów wielkości.

bariera potencjału dla cząstki α

Rozpad α jest efektem czysto kwantowym. Polega on na przeniknięciu cząstki α przez barierę potencjału (efekt tunelowy), które nie jest możliwe w fizyce klasycznej. Bariera potencjału pojawia się jako rezultat nałożenia się dwu oddziaływań pomiędzy cząstką α i pozostałą częścią jądra: oddziaływania jądrowego, które ma charakter krótkozasięgowy i jest przyciągające, oraz oddziaływania kulombowskiego, które ma charakter długozasięgowy i jest odpychające. Zilustrowane jest to na rys. 3. Rysunek 3a pokazuje schematyczny przebieg potencjału jądrowego $U_j(r)$ i potencjału kulombowskiego $U_{kul}(r)$ w funkcji odległości r cząstki α od środka jądra. Z rys. 3b widać, że superpozycja tych oddziaływań daje potencjał z barierą, którą cząstka α o energii E_α musi przeniknąć na odległości od R (promień jądra) do b (punkt wyjścia z bariery). Im energia E_α jest większa, tym grubość (i jednocześnie względna wysokość) bariery do pokonania jest mniejsza, a zatem większe prawdopodobieństwo przeniknięcia przez nią.



Rys. 3. Mechanizm powstawania bariery potencjału dla rozpadu α. a) Schematyczny przebieg potencjału jądrowego $U_j(r)$ i potencjału kulombowskiego $U_{kul}(r)$ dla cząstki α w funkcji jej odległości r od środka jądra. b) Superpozycja tych dwu potencjałów dająca barierę potencjału, którą cząstka α o energii E_α musi pokonać na odległości od R do b

prawdopodobieństwo emisji α

Prawdopodobieństwo emisji cząstki α można w przybliżeniu zapisać jako:

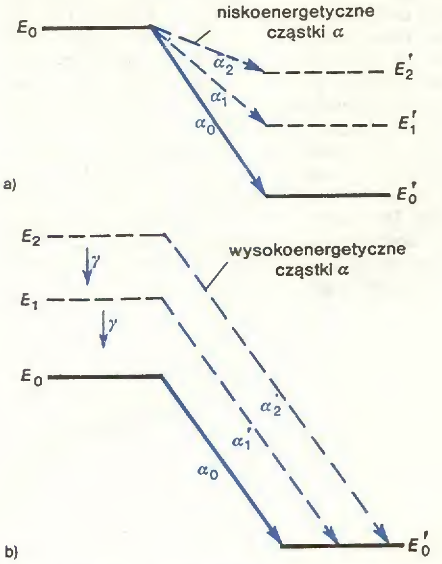
$$\lambda_\alpha \approx \frac{v_\alpha}{2R} P_\alpha = \nu_\alpha P_\alpha, \tag{7}$$

gdzie v_α jest prędkością cząstki α w jądrze, a R — promieniem jądra. Wielkość $\nu_\alpha = v_\alpha/2R$ jest częstością „uderzeń" cząstki α o barierę potencjału (tj. liczbą przebiegnięć cząstki od jednego do drugiego brzegu jądra w jednostce czasu), zatem całkowite prawdopodobieństwo emisji α w jednostce czasu jest iloczynem liczby uderzeń o barierę w tym czasie i prawdopodobieństwa P_α przeniknięcia przez nią przy każdym uderzeniu.

Rola rozpadu α jest tym większa, im większa jest liczba atomowa Z jądra. Tym większe bowiem jest wtedy prawdopodobieństwo rozpadu λ_α, czyli tym mniejszy czas połowicznego zaniku T_α. Najmniejsze czasy T_α mają więc jądra najcięższych pierwiastków lub pierwiastków bogatych w protony (lub, co na jedno wychodzi, ubogich w neutrony). Czasy T_α jąder pierwiastków ziem rzadkich są ogromne (np. ok. $8 \cdot 10^{15}$ lat wynosi T_α izotopu samaru ${}^{142}_{62}Sm$, jak podaliśmy wyżej). Dla najcięższych zaś znanych obecnie pierwiastków (Z = 104, 105, 106) — już tylko rzędu sekund. Na przykład dla odkrytego w 1974 roku izotopu ${}^{263}106$ wynosi on 0,9 s.

struktura subtelna widma α

Przy precyzyjnych badaniach rozpadu α określonego jądra można na ogół zaobserwować cząstki α o kilku różnych energiach (tzw. struktura subtelna widma α). Jest to wynikiem tego, że jądro początkowe może się rozpadać nie tylko do stanu podstawowego jądra końcowego, lecz także do jego stanów wzbudzonych (→ Jądra atomowe i ich wzbudzenia); ilustruje to rys. 4a. Natężenia jednak rozpadów do stanów wzbu-



Rys. 4. Ilustracja pochodzenia nisko- i wysokoenergetycznych cząstek α: a) Rozpad α ze stanu podstawowego jądra wyjściowego do stanu podstawowego oraz do stanów wzbudzonych (niskoenergetyczne cząstki α) jądra końcowego. b) Rozpad α ze stanu podstawowego oraz ze stanów wzbudzonych (wysokoenergetyczne cząstki α) jądra wyjściowego do stanu podstawowego jądra końcowego

nisko- i wysokoenergetyczne cząstki α

dzonych (rozpady $\alpha_1, \alpha_2, ...$) są z reguły bardzo małe w stosunku do natężenia rozpadu do stanu podstawowego (rozpad α_0), gdyż energie $E_{\alpha_1}, E_{\alpha_2} ...$ są mniejsze od E_{α_0}. Można obserwować także rozpady ze stanów wzbudzonych jądra początkowego do stanu podstawowego jądra końcowego (rys. 4b). Tutaj, ze względu na wysokie energie cząstek α, $E_{\alpha_1'}, E_{\alpha_2'}$ .., można by oczekiwać dużych natężeń linii α_1', α_2' ...,. W rzeczywistości jednak natężenia te są bardzo małe, gdyż stany wzbudzone „chętniej" ulegają rozpadowi γ niż α i tylko w bardzo niewielu przypadkach jądro wzbudzone zdąży wyemitować cząstkę α przed emisją kwantu γ.

Przy dokładnym opisie rozpadu α należy brać pod uwagę kilka dodatkowych czynników. Należą tu:
— stopień różnicy w strukturze jądra początkowego i końcowego. Znaczne różnice w tej strukturze, jak różnice w spinie, parzystości, w strukturze powłokowej, deformacji, powodują dodatkowe spowolnienia tzw. wzbronienia) w rozpadzie α;
— unoszenie przez niektóre cząstki α z jądra pewnej ilości momentu pędu. Daje to spowolnienie w emisji tych cząstek, ponieważ bariera potencjału dla nich jest wyższa (efekt sił odśrodkowych);
— wzór (7) jest słuszny przy założeniu, że cząstki α są już gotowe w jądrze, przed ich emisją. W rzeczywistości w jądrze są protony i neutrony i istnieje tylko określone, znacznie mniejsze od jedności prawdopodobieństwo tworzenia się z nich struktur złożonych, takich jak cząstki α.

Rozpad β

Pod nazwą rozpadu lub przemiany β rozumiemy trzy procesy: rozpad β^-, rozpad β^+ i wychwyt elektronu orbitalnego.

rozpad β^-

Rozpad β^- polega na przemianie jednego neutronu w jądrze na proton, której towarzyszy emisja negatonu (elektronu ujemnego) e^- i antyneutrina $\tilde{\nu}$. Symbolicznie możemy to zapisać:

$$ {}^{A}_{Z}X_N \rightarrow {}^{A}_{Z+1}Y_{N-1} + e^- + \tilde{\nu}. \quad (8)$$

rozpad β^+

Rozpad β^+ jest przemianą jednego protonu w jądrze na neutron, której towarzyszy emisja pozytonu (elektronu dodatniego) e^+ i neutrina. Symboliczny zapis:

$$ {}^{A}_{Z}X_N \rightarrow {}^{A}_{Z-1}Y_{N+1} + e^+ + \nu. \quad (9)$$

wychwyt elektronowy

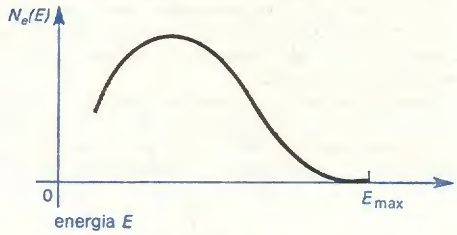
Wychwyt elektronowy polega na pochwyceniu przez jądro negatonu z jednej z powłok atomowych, czemu towarzyszy emisja neutrina. Symboliczny zapis:

$$ {}^{A}_{Z}X_N + e^- \rightarrow {}^{A}_{Z-1}Y_{N+1} + \nu. \quad (10)$$

Najbardziej prawdopodobny jest wychwyt z powłoki *K* (wychwyt *K*), znajdującej się najbliżej jądra.

widmo rozpadu β

Widmo energetyczne elektronu (negatonu e^- lub pozytonu e^+) obserwowanego w rozpadzie β jest ciągłe. Ilustruje to rys. 5 podający zależność ilości



Rys. 5. Zależność ilości elektronów N_e obserwowanych w rozpadzie β od ich energii E

obserwowanych elektronów N_e od ich energii E. Fakt ten odpowiada różnemu możliwemu podziałowi energii rozpadu pomiędzy elektron i neutrino (czy antyneutrino) i był jednym z głównych powodów wprowadzenia (przez Pauliego) hipotezy istnienia neutrina. Neutrino bowiem, jako cząstka oddziałująca słabo z materią, nie jest bezpośrednio obserwowane w rozpadzie β. Jedynie maksymalna energia widma, odpowiadająca wypadkowi, gdy elektron unosi całą dostępną energię, równa jest energii przejścia β:

$$E_{max} = E_X - E_Y.$$

W wypadku wychwytu elektronu emitowana jest tylko jedna cząstka (neutrino). Ma więc ona określoną energię, równą energii przejścia.

Ze wzorów (8–10) widać, że rozpad β nie zmienia liczby masowej jądra *A*. Jest więc przemianą, w której jądro wyjściowe i końcowe są izobarami. Rozpad β^- zwiększa ładunek jądra o 1, a rozpad β^+ i wychwyt elektronowy — zmniejszają.

izobary trwałe względem rozpadu β

Wśród wszystkich izobarów o danym *A* istnieją tylko jeden, dwa lub trzy stabilne (trwałe) względem rozpadu β. Dokładniej, wśród izobarów o nieparzystym *A* istnieje tylko jeden trwały, a wśród izobarów o parzystym *A* mogą istnieć jeden, dwa lub trzy izobary trwałe. Stanowią one jądra najsilniej związane wśród tych izobarów. Dla parzystego *A* są to jądra o parzystej liczbie zarówno protonów *Z*, jak i neutronów *N* (tzw. jądra parzysto-parzyste). Natomiast jądra o nieparzystych *Z* i *N* (jądra nieparzysto-nieparzyste) nie są trwałe względem rozpadu β (z wyjątkiem kilku jąder najlżejszych, nie wychodzących poza ${}^{14}_{7}N$). Szczególna trwałość jąder parzysto-parzystych pochodzi stąd, że oddziaływania jądrowe uprzywilejowują wiązanie się identycznych nukleonów w pary.

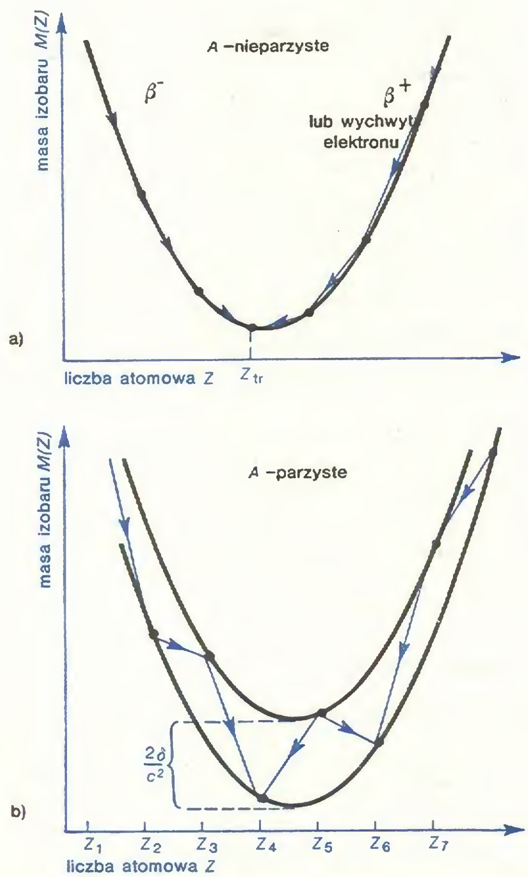
izobary trwałe o *A* nieparzystym

Zilustrowane to jest na rys. 6. Na rys. 6a podana jest zależność masy izobaru o nieparzystym *A* od jego ładunku *Z*. Zależność tę dobrze opisuje model kroplowy jądra (→ Modele jądrowe). Wykresem jej jest

parabola. Poszczególne izotopy odpowiadają na tej paraboli punktom, których współrzędna Z ma wartości całkowite. Gdyby nie było sił kulombowskich pomiędzy protonami, najniższy punkt (najmniejsza masa, najsilniej związane jądro) byłby przy $Z = A/2$. Wobec odpychania kulombowskiego dogodniejsza energetycznie jest mniejsza liczba protonów w jądrze, zatem minimum występuje przy $Z < A/2$. Widoczne jest, że parabola ma zawsze dokładnie jeden punkt najniższy, a więc istnieje dokładnie jeden izobar trwały, odpowiadający Z_{tr}.

izobary trwałe o A parzystym

W przypadku parzystego A (rys. 6b) mamy dwie parabole, oddalone od siebie o masę $2\delta/c^2$ równoważną energii rozerwania dwu par: jednej protonowej i jednej neutronowej. Przemiana β, zmieniając jądro z parzysto-parzystego na nieparzysto-nieparzyste (lub odwrotnie), prowadzi od punktu na paraboli dolnej do punktu na paraboli górnej (lub odwrotnie). Dla dowolnego punktu na paraboli górnej zawsze istnieje punkt sąsiedni, niższy, na paraboli dolnej i wobec tego jądro nieparzysto-nieparzyste zawsze (z wyjątkiem wspomnianych wyżej kilku jąder lekkich) może przejść do jądra parzysto-parzystego. Na dolnej paraboli jednak może być kilka punktów, dla których nie ma sąsiednich punktów na paraboli górnej, które leżałyby niżej. Może więc być kilka parzysto-parzystych izobarów trwałych względem rozpadu β. Na rys. 6b izobarami takimi są izobary o Z_4 i Z_6. Na rysunku tym izobary o Z_1, Z_3, Z_5 i Z_7 są nieparzysto-nieparzyste, a izobary o Z_2, Z_4 i Z_6 są parzysto-parzyste. Interesujący jest przypadek izobaru nieparzysto-nieparzystego o liczbie atomowej Z_5. Widać, że może on ulegać rozpadowi zarówno β^-, jak i β^+. Przykładem takiego izobaru jest izotop miedzi ^{64}Cu, który ulega wszystkim trzem rodzajom rozpadu β: β^- (31% wszystkich rozpadów), β^+ (15%) i wychwytowi elektronu (54%).



Rys. 6. Zależność masy izobaru od jego liczby atomowej Z: a) Izobary o nieparzystej liczbie masowej A. Widoczne jest, że wśród wszystkich izobarów o nieparzystym A istnieje tylko jeden ($Z = Z_{tr}$) trwały względem rozpadu β. Pozostałe rozpadają się bądź poprzez rozpad β^- (gdy $Z < Z_{tr}$), bądź β^+ lub wychwyt elektronu (gdy $Z > Z_{tr}$) i to tym szybciej, im dalej położony jest dany izobar od izobaru trwałego. b) Izobary o parzystym A. Dolna parabola odpowiada jądrom parzysto-parzystym, a górna — nieparzysto-nieparzystym. Widać, że w przypadku parzystego A mogą istnieć dwa (a nawet trzy) izobary trwałe względem rozpadu β. Na rysunku są to izobary o $Z = Z_4$ i $Z = Z_6$

szybkość rozpadu β

Szybkość rozpadu β rośnie na ogół ze wzrostem energii rozpadu. Rysunek 6 ilustruje więc, że przy oddalaniu się od izobaru trwałego, a więc od ścieżki trwałości (→ Jądra atomowe i ich wzbudzenia), energia a zatem i szybkość rozpadu rośnie, tzn. czasy połowicznego zaniku T_β maleją. Przy tym, jeśli oddalamy się w kierunku jąder ubogich w neutrony (np. produkty reakcji z ciężkimi jonami), to jest to wychwyt elektronu lub rozpad β^+, jeśli zaś w kierunku jąder bogatych w neutrony (np. produkty naświetlania strumieniem neutronów lub produkty rozszczepienia jąder ciężkich), to jest to rozpad β^-.

przejścia dozwolone i wzbronione

Para elektron plus neutrino (czy antyneutrino) może być emitowana z jądra w stanach o różnym orbitalnym momencie pędu $l = 0, 1, 2, \ldots$ (wyrażonym w jednostkach $\hbar$). Najbardziej prawdopodobna jest emisja z $l = 0$. Są to tzw. przejścia „dozwolone". Przejścia z wyższymi momentami orbitalnymi nazywają się, ze względów historycznych, „wzbronionymi", a samo l — stopniem wzbronienia. Każdy stopień wzbronienia powoduje spowolnienie rozpadu o ok. dwa rzędy wielkości (tzn. o ok. 100 razy). Wśród przejść dozwolonych wyróżnia się tzw. przejścia uprzywilejowane, które zachodzą między stanami jądrowymi o podobnej strukturze wewnętrznej.

Rozpad γ

Przez rozpad czy przejście γ rozumiemy przejście jądra ze stanu wzbudzonego do stanu o energii niższej, podczas którego energia przejścia unoszona jest przez promieniowanie elektromagnetyczne. Ze względu na duże energie przejść jądrowych, długości fal tego promieniowania są małe (z reguły mniejsze od 1 Å, tzn. od 0,1 nm) i ważny jest cząstkowy (korpuskularny) aspekt tego promieniowania.

Kwant promieniowania γ emitowany z jądra charakteryzuje się:

reguły wyboru dla przejść γ

— energią, jaką unosi on z jądra. Jest to energia przejścia jądrowego (deekscytacji)

$$E_\gamma = E' - E, \tag{11}$$

gdzie E jest energią stanu początkowego, a E' — energią stanu końcowego jądra,

— momentem pędu (spinem, krętem) l (mierzonym w jednostkach $\hbar$), jaki unosi on z jądra. Może on przyjmować wartości $l = 1, 2, 3, \ldots$ Mówi się, że promieniowanie (i odpowiadające mu przejście) unoszące spin l jest 2^l-polowe lub multipolowe rzędu lub stopnia l. Przy $l = 1$ jest to promieniowanie dipolowe, przy $l = 2$ — kwadrupolowe, $l = 3$ — oktupolowe. Promieniowanie określonej multipolowości l może być elektryczne (E) lub magnetyczne (M). Z zasady zachowania momentu pędu wynika, że spin unoszony przez kwant γ jest różnicą (wektorową, gdyż moment pędu jest wielkością wektorową) między spinami stanów: początkowego I i końcowego I' jądra. Daje to warunek na l

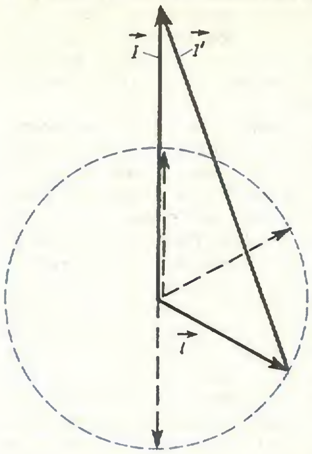
$$|I - I'| \leqslant l \leqslant I + I', \tag{12}$$

zilustrowany schematycznie na rys. 7,

— parzystością P, która może być dodatnia (+) lub ujemna (−). Promieniowanie elektryczne o multipolowości l ma parzystość $(-1)^l$, a magnetyczne — $(-1)^{l+1}$. Zasada zachowania parzystości w oddziaływaniach elektromagnetycznych daje dla przejścia γ o multipolowości l warunek

$$PP_l = P', \tag{13}$$

tzn., e między stanami jądrowymi o tej samej parzystości ($P' = P$) mogą zachodzić tylko przejścia o parzystości dodatniej (np. przejście elektryczne



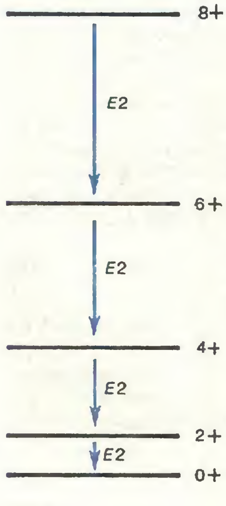
Rys. 7. Ilustracja faktu, że spin stanu początkowego $\vec{I}$ jest sumą (wektorową) spinu stanu końcowego $\vec{I'}$ oraz spinu $\vec{l}$ unoszonego przez kwant γ. Liniami kreskowanymi zaznaczone są różne możliwe orientacje spinu $\vec{l}$

kwadrupolowe lub magnetyczne dipolowe), a między stanami o parzystości przeciwnej ($P' = -P$) — przejścia o parzystości ujemnej (np. elektryczne dipolowe).

Wzory (11–13) nazywają się regułami wyboru dla przejść γ. Ponieważ są one konsekwencją zasad zachowania energii, momentu pędu i parzystości, stanowią warunki, które muszą być spełnione przy tych przejściach. Reguły wyboru pozwalają na wyznaczanie energii, spinów i parzystości stanów jądrowych (→ Spektroskopia jądrowa) za pomocą pomiaru energii i multipolowości przejść γ.

prawdopodobieństwo przejścia γ

Prawdopodobieństwo przejścia γ, λ_γ silnie zależy od energii przejścia E_γ i multipolowości l. Rośnie ono silnie ze wzrostem E_γ, a bardzo silnie maleje ze wzrostem l. Ten ostatni fakt oznacza, że najczęściej obserwuje się przejścia tylko najniższej multipolowości: $l = 1$ (dipolowe) i $l = 2$ (kwadrupolowe), a spośród wszystkich możliwych przejść między stanami o spinach I' i I, podanych przez regułę (12), obserwuje się na ogół tylko przejście o najniższym l, tj. o $l = |I'-I|$. Jeszcze innym wyrazem tej zależności jest fakt, że pomiędzy stanami o znacznie różniących się spinach jądro dokonuje najchętniej przejścia γ nie bezpośrednio (potrzebna duża multipolowość przejścia), lecz przez stany o spinach pośrednich. Przykład takich przejść podany jest na rys. 8, który pokazuje widmo rotacyjne jądra parzysto-parzystego zdeformowanego. Spiny kolejnych stanów są: 0, 2, 4, 6, 8, ... Obserwuje się tu kaskadę przejść o najniższej możliwej multipolowości ($l = 2$), a więc przejścia pomiędzy stanami o najmniejszej różnicy spinów.



Rys. 8. Przykład typowego widma rotacyjnego jądra parzysto-parzystego zdeformowanego. Jądro zdeformowane może obracać się i energie odpowiadające różnym stanom tego obrotu (rotacji) tworzą właśnie widmo rotacyjne. Poziomy tego widma leżą stosunkowo nisko i dobrze się separują od pozostałych poziomów wzbudzonych jądra. Teoretyczny opis tych poziomów podany jest przez model kolektywny jądra (→ Modele jądrowe). Kolejne stany mają spiny: 0, 2, 4, 6, 8, ... oraz parzystość dodatnią (+). Doświadczalnie obserwuje się tutaj kaskadę przejść γ typu $E2$ (elektryczne kwadrupolowe) pomiędzy sąsiednimi stanami. Przejścia pomiędzy stanami dalszymi niż sąsiednie są natomiast bardzo słabe

Interesujące są przypadki, gdy pomiędzy stanami o znacznie różniących się spinach nie występują stany o spinach pośrednich. Przykładem są jądra, w których spiny pierwszego stanu wzbudzonego i stanu podstawowego różnią się znacznie. Wtedy prawdopodobieństwo przejścia γ ze stanu wzbudzonego jest małe, bo multipolowość przejścia jest duża i czas połowicznego zaniku T_γ jest duży. Może on być rzędu nawet godzin, podczas gdy dla przeciętnych stanów jądrowych T_γ jest, orientacyjnie, w granicach od mikrosekund (10^{-6} s) do pikosekund (10^{-12} s). Takie długożyciowe (metatrwałe) stany nazywają się izomerycznymi, a jądra w tych stanach — izomerami jądrowymi. Przykładem izomeru jądrowego jest jądro izotopu indu ^{115}In w stanie o energii 335 keV. Stan ten jest stanem o najniższej energii wzbudzenia; jego spin wynosi $^1/_2$, a parzystość (−). Przejście do stanu podstawowego, który ma spin $^9/_2$ i parzystość (+), może więc, zgodnie ze wzorem (12), mieć multipolowość tylko 4 lub 5. Uwzględniając także wzór (13) może to być przejście typu $M4$ lub $E5$. W związku z wysoką multipolowością i małą energią przejścia czas połowicznego zaniku izomeru jest duży i wynosi 4,5 h.

izomery jądrowe

Konwersja wewnętrzna

Procesem konkurencyjnym do przejścia γ jest przejście, w którym energia deekscytacji jądra przekazywana jest bezpośrednio jednemu z elektronów powłoki atomowej. Przejście takie nazywa się konwersją wewnętrzną. Elektron unosi tu energię, moment pędu i parzystość przejścia jądrowego, analogicznie, jak czyni to emitowany z jądra kwant γ. Przejście ma więc określoną energię, multipolowość i parzystość. Obowiązują te same, co dla przejścia γ, reguły wyboru (11–13). Ponieważ całą energię przejścia unosi jeden elektron, to widmo elektronów konwersji wewnętrznej jest liniowe (prążkowe, dyskretne), w odróżnieniu od widma elektronów pochodzących z rozpadu β, które jest ciągłe.

Stosunek prawdopodobieństwa deekscytacji jądra w drodze konwersji wewnętrznej λ_e do prawdopodobieństwa deekscytacji w drodze przejścia γ, λ_γ, nazywa się współczynnikiem konwersji wewnętrznej α, tzn. $\alpha = \lambda_e/\lambda_\gamma$. Współczynnik α silnie zależy od energii przejścia i jego multipolowości oraz od ładunku jądra Z. Maleje on szybko ze wzrostem energii przejścia, co oznacza, że konwersja wewnętrzna odgrywa najważniejszą rolę przy stosunkowo niskich energiach. Rośnie natomiast szybko ze wzrostem stopnia multipolowości przejścia l. Fakt ten dostarcza ważnej metody określania stopnia multipolowości przejścia przez pomiar współczynnika α (→ Spektroskopia jądrowa). Współczynnik α silnie rośnie ze wzrostem liczby atomowej jądra Z, co oznacza, że rola konwersji wewnętrznej rośnie przy przejściu do pierwiastków najcięższych.

współczynnik konwersji wewnętrznej

Rozszczepienie

Rozszczepieniem nazywamy proces, w którym jądro rozpada się na dwie lub więcej porównywalnych, co do wielkości, części (fragmentów). Występuje ono dla jąder ciężkich i zachodzi z większym prawdopodobieństwem w stanie wzbudzonym jądra niż w stanie podstawowym (tzw. rozszczepienie samorzutne).

Prawdopodobieństwo rozszczepienia na dwa fragmenty (rozszczepienie podwójne) jest największe. Stosunkowo jeszcze znaczne jest prawdopodobieństwo rozszczepienia, w którym obok dwu ciężkich fragmentów powstaje także cząstka α (tzw. trypartycja). Rozszczepienie takie zachodzi z częstością ok. 1 przypadku na 400 przypadków rozszczepienia podwójnego. Prawdopodobieństwo rozszczepienia na 3 lub 4 porównywalne fragmenty jest już znikome.

trypartycja

Ponieważ rozszczepiające się jądra ciężkie są znacz-

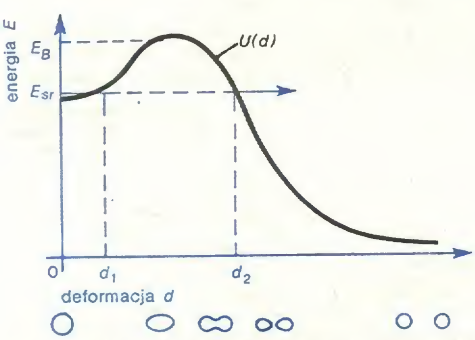
nie bogatsze w neutrony ($N/Z \approx 1{,}6$) niż jądra średnie ($N/Z \approx 1{,}3$), to i fragmenty rozszczepienia są bogate w neutrony. Fragmenty te, powstające w silnie wzbudzonych stanach, emitują neutrony bezpośrednio po utworzeniu się (neutrony natychmiastowe) w liczbie średnio ok. 2,5 na jeden akt rozszczepienia, a także po rozpadzie β (neutrony opóźnione). Emisja neutronów, które mogą z kolei powodować rozszczepienie innych jąder, stwarza możliwość zajścia reakcji łańcuchowej (→ Energia jądrowa). W każdym akcie rozszczepienia wyzwala się duża, rzędu 200 MeV, energia (energia jądrowa).

neutrony natychmiastowe i opóźnione

Rozszczepienie jąder ciężkich możliwe jest dzięki temu, że są one słabiej związane niż jądra o średniej masie, na które się rozpadają. Wiąże się to z odpychaniem kulombowskim między protonami. Energia tego odpychania rośnie ze wzrostem liczby atomowej Z jak Z^2, co powoduje silne obniżenie energii wiązania jąder ciężkich w stosunku do jąder lżejszych. Możliwe energetycznie rozszczepienie nie zachodzi jednak natychmiast, lecz z pewnym, dla niektórych jąder bardzo dużym, opóźnieniem, co spowodowane jest obecnością bariery potencjału.

przebieg rozszczepienia

Orientacyjnie proces rozszczepienia można opisać za pomocą modelu kroplowego jądra (→ Modele jądrowe). Proces ten polega na deformowaniu się jądra od kształtu kulistego lub prawie kulistego poprzez coraz bardziej wydłużony, wydłużony z przewężeniem w środku, aż do uformowania się i rozdzielenia dwu fragmentów. Zgodnie z modelem kroplowym energia potencjalna jądra może być przedstawiona jako suma energii powierzchniowej i energii kulombowskiej. Energia powierzchniowa jest proporcjonalna do pola powierzchni jądra. Ponieważ pole to rośnie ze wzrostem deformacji jądra, to i energia powierzchniowa rośnie. Energia kulombowska zaś maleje, ponieważ deformowanie jądra (a dokładniej, wydłużanie się jego) sprzyja oddalaniu się od siebie odpychających się elektrycznie protonów. Dla małych deformacji przyrost energii powierzchniowej jest większy od ubytku energii kulombowskiej, dla dużych zaś — odwrotnie. Całkowita energia zatem, jako suma ich obu, początkowo wzrasta, następnie przechodzi przez maksimum i wreszcie maleje ze wzrostem deformacji jądra. Powstaje więc bariera potencjału. Ilustruje to rys. 9, na którym energia jądra $U(d)$ przed-



Rys. 9. Zależność energii potencjalnej jądra U od jego deformacji d, jaką otrzymuje się w ramach modelu kroplowego. Dla wszystkich jąder minimum energii (stan podstawowy jądra) otrzymywane jest przy deformacji zerowej (kształt kulisty). Na dole rysunku zobrazowane są kształty odpowiadające kilku wartościom parametru deformacji d

stawiona jest w funkcji deformacji d. Jądro w stanie o energii E_{sr}, aby ulec rozszczepieniu, musi przeniknąć barierę (efekt tunelowy) od deformacji d_1 do deformacji d_2. Kształty jądra odpowiadające różnym deformacjom d ukazuje rysunek.

Proces rozszczepienia jest zatem podobny do procesu rozpadu α. Oba procesy polegają na tunelowym przeniknięciu przez barierę potencjału, utworzoną przez nałożenie się oddziaływania jądrowego, dążącego do zapobiegnięcia rozpadowi, i odpychania kulombowskiego, dążącego do rozpadu. W rozpadzie α jest to bariera ze względu na oddalanie się cząstki α od jądra, przy rozszczepieniu zaś jest to bariera ze względu na deformację jądra.

parametr rozszczepialności

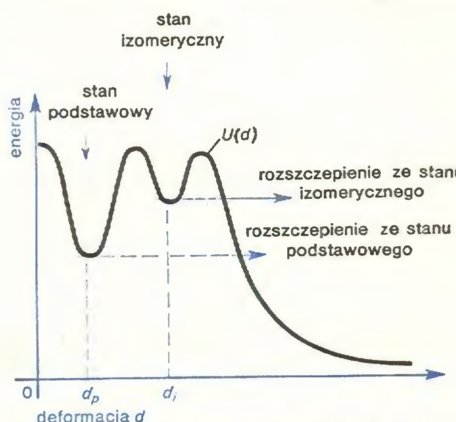
Wysokość bariery, a więc i prawdopodobieństwo rozszczepienia, a tym samym i czas połowicznego zaniku ze względu na ten proces, zależą bardzo silnie od stosunku energii odpychania kulombowskiego do energii powierzchniowej zwanego parametrem rozszczepialności, który wynosi w przybliżeniu

$$x \approx \frac{1}{50}\frac{Z^2}{A}.$$

Na przykład dla jądra izotopu uranu ^{238}U parametr rozszczepialności wynosi: $x \approx 0{,}71$, a czas połowicznego zaniku ze względu na samorzutne rozszczepienie: $T_{sr} \approx 6 \cdot 10^{15}$ lat. Dla cięższego jądra izotopu fermu ^{254}Fm: $x \approx 0{,}79$, a $T_{sr} \approx 220$ dni, tzn. jest o ok. 16 rzędów krótszy. Dla jeszcze cięższego jądra $^{258}104$: $x \approx 0{,}83$, a (zmierzony ostatnio) $T_{sr} \approx 0{,}01$ s, tzn. jest o ok. dalsze 9 rzędów krótszy.

Według modelu kroplowego jądra mające $x > 1$, tj. dla których $Z^2/A > 50$, nie mogą istnieć. Rozszczepiałyby się one natychmiast. Model kroplowy nie uwzględnia jednak ważnych dla jądra efektów powłokowych (efektów struktury powłokowej jądra), które istotnie modyfikują przewidywania tego modelu. M.in. efekty te dopuszczają istnienie jąder bardzo ciężkich z $Z^2/A > 50$ (hipoteza jąder superciężkich → Jądra atomowe w stanach ekstremalnych). Powodują one także, że dla wielu jąder w barierze na rozszczepienie pojawia się znaczne wgłębienie, lokalne minimum (tzw. drugie minimum), do którego, jeśli

hipoteza jąder superciężkich



Rys. 10. Zależność energii potencjalnej jądra U od jego deformacji d po uwzględnieniu efektów powłokowych. Efekty te powodują, że dla wielu jąder pierwsze, głębokie minimum energii potencjalnej, odpowiadające stanowi podstawowemu jądra, występuje przy kształcie zdeformowanym (jądra zdeformowane), a nie kulistym, jak przewiduje model kroplowy. Efekty powłokowe powodują także, że dla niektórych jąder występuje również drugie, płytsze minimum energii potencjalnej, odpowiadające stanowi izomerycznemu. Ze względu na mniejszą barierę, rozszczepienie ze stanu izomerycznego następuje znacznie szybciej niż ze stanu podstawowego. Deformacja jądra w stanie podstawowym oznaczona jest przez d_p, a w izomerycznym — przez d_i

jądro zostanie „schwytane", to trwa w nim stosunkowo długo, zanim się rozszczepi. Ilustruje to rys. 10, który przedstawia przebieg energii potencjalnej w funkcji deformacji dla niektórych jąder. Na rysunku widać, że jądro o deformacji d_i musi pokonać barierę zarówno po to, by się rozszczepić, jak też by przejść do stanu podstawowego. Stan jądra o tej deformacji jest więc metatrwały, a zatem jądro znajdujące się w tym stanie jest izomerem. Natura tego izomeru jest jednak inna niż izomerów dotąd omawianych. Poprzednio mówiliśmy o izomerach, których długi czas życia związany był z dużą różnicą między spinem stanu izomerycznego a spinami stanów osiąganych przez rozpad w drodze przejścia elektromagnetycznego. W omawianym wypadku zaś długi czas życia izo-

izomery rozszczepiające się meru związany jest z istnieniem bariery potencjału. Izomery tego typu nazywane są izomerami rozszczepiającymi się lub, czasami, izomerami kształtu. Pierwszy taki izomer zaobserwowany został w 1962 r. w jądrze ^{242}Am. Obecnie znamy ok. 30 takich izomerów. Występują one w jądrach uranu, plutonu, ameryku, kiuru i berkelu.

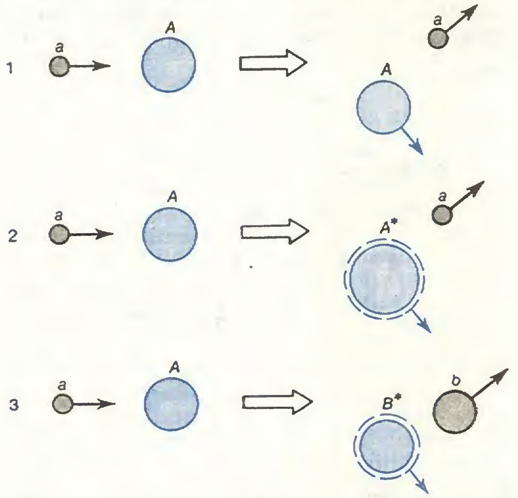
Ze względu na znacznie mniejszą barierę na rozszczepienie niż bariera jaką ma jądro w stanie podstawowym (por. rys. 10), prawdopodobieństwo rozszczepienia izomerów jest bardzo duże (stąd ich nazwa — izomery rozszczepiające się), czyli czasy ich życia ze względu na ten proces — bardzo małe. Czasy te zawierają się w przedziale od nanosekund (10^{-9} s do milisekund (10^{-3} s), są one o 21 do 31 rzędów krótsze niż odpowiednie czasy życia stanów podstawowych tych samych jąder. Na przykład czas życia ze względu na samorzutne rozszczepienie dla stanu podstawowego jądra ^{238}U wynosi ok. $6 \cdot 10^{15}$ lat, a dla rozszczepiającego się izomeru tego samego jądra — 195 ns, czyli jest o ok. 30 rzędów krótszy.

L. Groszew, I. Szapiro *Spektroskopia jąder atomowych*, Warszawa 1956; D. Halliday *Wstęp do fizyki jądrowej*, Warszawa 1957; A. Strzałkowski *Wstęp do fizyki jądra atomowego*, Warszawa 1978; Sz. Szczeniowski *Fizyka doświadczalna*, cz. 6, Warszawa 1974; E. Szpolski *Fizyka atomowa*, t. 2, cz. 2, Warszawa 1954.

Reakcje jądrowe

Piotr Decowski

Reakcją jądrową nazywamy proces wynikający z oddziaływania cząstki jądrowej (może nią być cząstka elementarna, jak np. nukleon, mezon, foton lub cząstka złożona, np. deuteron, cząstka α czy też każde inne jądro atomowe, tzw. ciężki jon) z jądrem atomowym. W wyniku zajścia reakcji jądrowej bombardowane jądro tarczy może pozostać bez zmiany, może zostać wzbudzone lub też może powstać inne jądro atomowe w stanie podstawowym czy też wzbudzonym (rys. 1). Procesowi temu towarzyszy zwykle **rozpraszanie elastyczne i nieelastyczne**



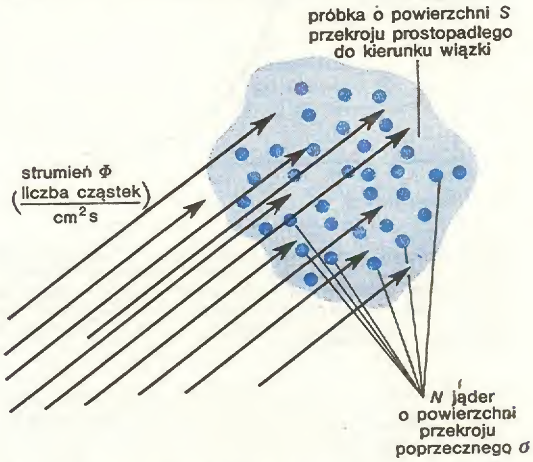
Rys. 1. Przykłady różnych reakcji jądrowych: *1.* Rozpraszanie elastyczne: cząstka *a* rozprasza się sprężyście w oddziaływaniu z jądrem *A*. Jądro *A* uzyskuje pewną energię kinetyczną, lecz jego stan wewnętrzny nie ulega zmianie. *2.* Rozpraszanie nieelastyczne. W oddziaływaniu część energii kinetycznej cząstki *a* zostaje zużyta na wzbudzenie jądra *A* (jądro wzbudzone oznaczamy jako *A**). *3.* Reakcja jądrowa sensu stricto: w wyniku oddziaływania powstaje cząstka *b* i jądro *B**

emisja jednej lub wielu cząstek elementarnych lub złożonych. Reakcje jądrowe zachodzą stosunkowo szybko — w czasie od 10^{-23} s do 10^{-15} s. Stosują się do nich wszystkie klasyczne prawa zachowania obowiązujące powszechnie w fizyce (prawo zachowania energii, pędu, momentu pędu, ładunku), ponadto kwantowy charakter procesu, wynikający z małych rozmiarów oddziałujących obiektów oraz małych wartości wymienianej energii, narzuca dodatkowe prawa zachowania (np. parzystości funkcji falowej).

Z zasady zachowania energii wynika, że ubytek lub zysk całkowitej energii kinetycznej cząstek biorących udział w reakcji może mieć źródło tylko w zysku lub ubytku łącznej masy cząstek. Różnica łącznej masy cząstek wchodzących w reakcję i łącznej masy produktów reakcji zwana jest ciepłem reakcji. **ciepło reakcji** Znajomość ciepła reakcji (np. obliczonego na podstawie dokładnych tablic mas jąder atomowych) pozwala przewidzieć kinetyczną energię produktów przy zadanej energii kinetycznej cząstki wywołującej reakcję. Do opisu reakcji jądrowych wprowadza się pojęcie „kanału reakcji". **kanał reakcji** Jest on scharakteryzowany przez rodzaj cząstek oddziałujących lub emitowanych w reakcji jądrowej, ich energię, moment pędu w ruchu względnym, przestrzenne ustawienie spinów. Tak więc mamy „kanał wejściowy" opisujący cząstkę wywołującą reakcję i jądro bombardowane oraz szereg „kanałów wyjściowych" opisujących wszystkie możliwe cząstki i jądra końcowe po zajściu reakcji.

przekrój czynny Miarą prawdopodobieństwa zajścia określonej reakcji jądrowej jest wielkość zwana przekrojem czynnym. Przekrój czynny jest liczbowo równy prawdopodobieństwu zajścia reakcji w czasie jednej sekundy dla jednostkowego strumienia cząstek wywołujących reakcję (1 cząstka/cm$^2 \cdot$s) podzielonemu przez liczbę jąder w bombardowanej próbce (rys. 2). Przekrój czynny przedstawia jakby efektywną powierzchnię przekroju jądra atomowego w badanym procesie. Przemnożony przez liczbę jąder w próbce daje powierzchnię stanowiącą efektywną przegrodę dla wiązki cząstek bombardujących (przy założeniu, że próbka jest na tyle cienka, że wzajemne „przesłanianie" jąder można pominąć).

Przekroje czynne na oddziaływanie cząstek bombardujących z jądrami są na ogół rzędu rozmiarów jąder: 10^{-24} cm^2 (choć np. dla powolnych neutronów

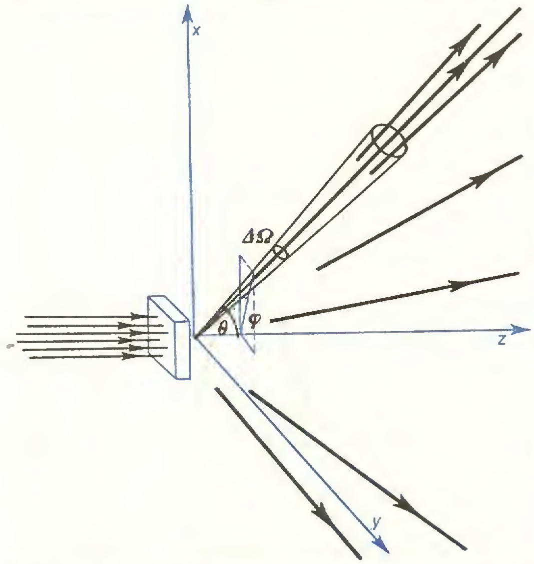


Rys. 2. Przekrój czynny w reakcji jądrowej. Przez próbkę przenika ΦS cząstek na sekundę. W próbce znajduje się N jąder, z których każde reprezentuje przekrój σ. Łączna powierzchnia czynna zajmowana przez jądra wynosi $N\sigma$ i stanowi $N\sigma/S$ całkowitej powierzchni próbki. Liczba cząstek zderzających się w ciągu sekundy z jądrami wynosi:

$$(N\sigma/S)\Phi S = N\sigma\Phi$$

mogą osiągać wartości nawet kilkadziesiąt tysięcy razy większe). Ze względu na małe wartości wprowadzono dla nich miary specjalne, jednostki zwane barnami (1 b = 10^{-24} cm²).

Prawdopodobieństwo procesu, w którym cząstki są wysyłane w określonym kierunku, wyraża się przez kątowy różniczkowy przekrój czynny: $d\sigma/d\Omega$. Jest on



Rys. 3. Przekrój czynny $\Delta\sigma$ dla procesu emisji cząstki w wąskim stożku obejmującym kąt bryłowy $\Delta\Omega$ o osi skierowanej w kierunku wyznaczonym przez kąty θ i φ jest określony przez kątowy różniczkowy przekrój czynny $d\sigma/d\Omega$:

$$\Delta\sigma = (d\sigma/d\Omega)\Delta\Omega$$

rozkład kątowy i energetyczny

graniczną wartością stosunku przekroju czynnego $\Delta\sigma$ procesu emisji cząstki w stożku o osi skierowanej w zadanym kierunku do dążącej do zera wartości obejmowanego przez stożek kąta bryłowego $\Delta\Omega$ (rys. 3). Zależność przekroju $d\sigma/d\Omega$ od kątów wyznaczających kierunek emisji nosi nazwę rozkładu kątowego reakcji. Podobnie tzw. energetyczny różniczkowy przekrój czynny $d\sigma/dE$ mówi o prawdopodobieństwie emisji cząstek o określonej energii. Wyznacza on przekrój czynny $\Delta\sigma$ na emisję cząstek o energiach mieszczących się w wąskim przedziale $(E, E+\Delta E)$: $\Delta\sigma = (d\sigma/dE)\Delta E$. Zależność energetycznego różniczkowego przekroju czynnego od energii emitowanej cząstki przy ustalonej energii cząstki padającej, nazywa się rozkładem energetycznym lub widmem emitowanych cząstek.

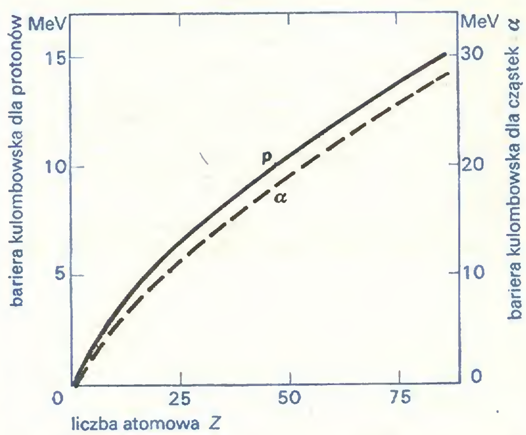
widmo energetyczne

Jądro powstające w wyniku reakcji jądrowej może być utworzone w różnych, ale na ogół ściśle określonych stanach energetycznych (jest to konsekwencją praw kwantowych rządzących mikroświatem, w myśl których dopuszczalne są tylko pewne wartości energii stanów układu). W związku z tym w widmie energetycznym cząstek towarzyszących tworzeniu jąder występują oddzielne linie odpowiadające poziomom energetycznym „jądra końcowego". Intensywność tych linii wyznacza prawdopodobieństwo wzbudzania różnych poziomów energetycznych (rys. 4).

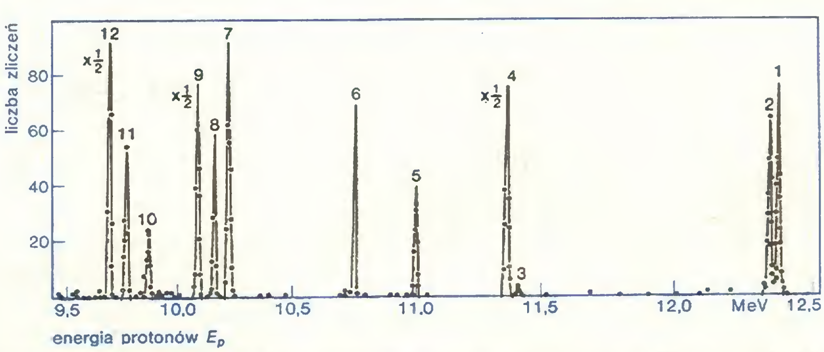
Badanie reakcji jądrowych

Badanie reakcji jądrowych zmierza do dwóch celów: poznania mechanizmu reakcji oraz poznania budowy jądra atomowego i struktury jego stanów wzbudzonych przez wpływ tych czynników na przebieg reakcji. Z reguły cząstki oddziałujące z jądrem siłami jądrowymi i wywołujące reakcje są, poza jednym wyjątkiem — neutronem, naładowane dodatnio (pomijamy tu mezony i hiperony, cząstki tworzone w oddziaływaniach o znacznej energii, ulegające bardzo szybko rozpadowi). Aby cząstka taka mogła znaleźć się w bezpośrednim sąsiedztwie jądra atomowego na tyle blisko, by krótkozasięgowe siły jądrowe mogły doprowadzić do oddziaływania, trzeba jej nadać odpowiednio dużą energię. Energia musi być wystarczająca do pokonania elektrostatycznego odpychania pochodzącego od dodatnio naładowanego jądra. Energia odpychania, zw. barierą kulombowską (rys. 5), wynosi np. dla protonów bombardujących od 2 MeV dla jąder najlżejszych do kilkunastu MeV dla jąder ciężkich (1 MeV = $1{,}6\cdot10^{-13}$ J jest energią, jaką nabywa ładunek elementarny po przebyciu różnicy potencjałów 1 miliona woltów).

bariera kulombowska



Rys. 5. Energia potrzebna na pokonanie przez protony i cząstki α odpychania elektrostatycznego różnych jąder o liczbie atomowej Z. Cząstki α — mające dwa razy większy ładunek niż proton — mają barierę w przybliżeniu dwa razy większą



Rys. 4. Widmo energetyczne protonów emitowanych pod kątem 30° podczas bombardowania tarczy aluminiowej deuteronami o energii 7 MeV. Część deuteronów ulega reakcji zdarcia — neutron zostaje włączony do jądra ^{27}Al tworząc jądro ^{28}Al, a proton wylatuje z próbki unosząc energię zwiększoną o energię wyzwoloną w reakcji (jest to reakcja egzotermiczna, ciepło reakcji wynosi w tym przypadku 5,5 MeV). Największa energia protonu — linia *1* — odpowiada powstaniu jądra ^{28}Al w stanie podstawowym, dalsze linie odpowiadają kolejnym stanom wzbudzonym (wg J. B. A. England *Metody doświadczalne fizyki jądrowej*, Warszawa 1981)

Cząstki przyspiesza się do odpowiedniej energii w różnego typu akceleratorach: elektrostatycznych, cyklicznych, liniowych (→ Akceleratory).

detekcja cząstek

Do rejestracji cząstek używa się wiele typów detektorów. Wszystkie działają na zasadzie rejestracji jonizacji wywoływanej przez cząstkę naładowaną w czasie jej przelotu przez ośrodek wypełniający objętość czynną detektora (ośrodkiem tym może być gaz, jak np. w licznikach gazowych, ciecz, jak np. w niektórych licznikach scyntylacyjnych lub w komorach pęcherzykowych, albo ciało stałe, jak w detektorach półprzewodnikowych, kliszach jądrowych itp.). Niekiedy, w celu określenia pędu cząstki analizuje się zakrzywienie jej toru w znanym polu magnetycznym, np. w spektrografie magnetycznym (→ Detekcja cząstek).

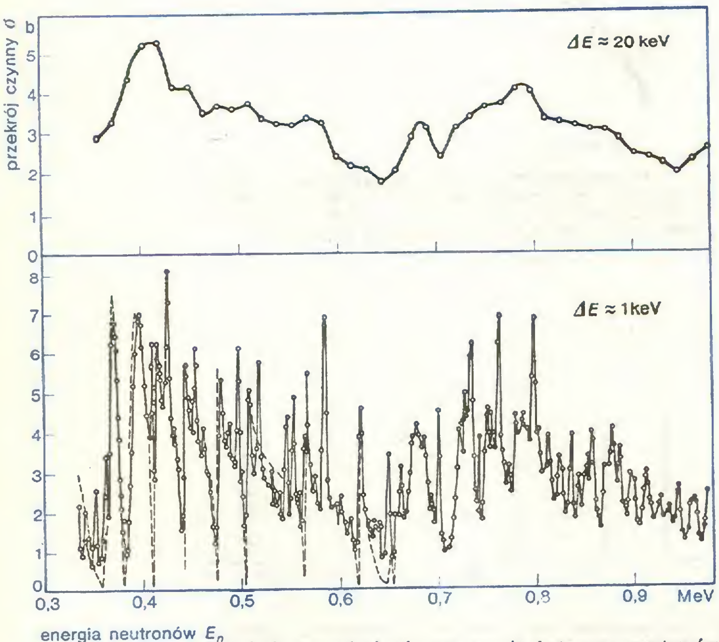
Wszystkie powyższe metody pozwalają rejestrować wyłącznie cząstki naładowane. Do detekcji cząstek obojętnych (np. neutronów lub fotonów) należy wykorzystywać wywołane przez nie reakcje wtórne, w których wytwarzane są cząstki naładowane (np. w przypadku fotonów efekt fotoelektryczny). W większości wypadków zadaniem detektorów jest wyznaczenie energii kinetycznej (lub pędu) rejestrowanej cząstki. Bardziej złożone układy detektorów (teleskopy) pozwalają określić jej masę i ładunek.

Neutrony, pozbawione ładunku elektrycznego, są szczególnym rodzajem cząstki wywołującej reakcję. Brak bariery kulombowskiej powoduje, że nawet bardzo powolne neutrony mogą oddziaływać z jądrem atomowym. Źródłem tych neutronów są reaktory jądrowe (→ Energia jądrowa).

wielkości mierzone

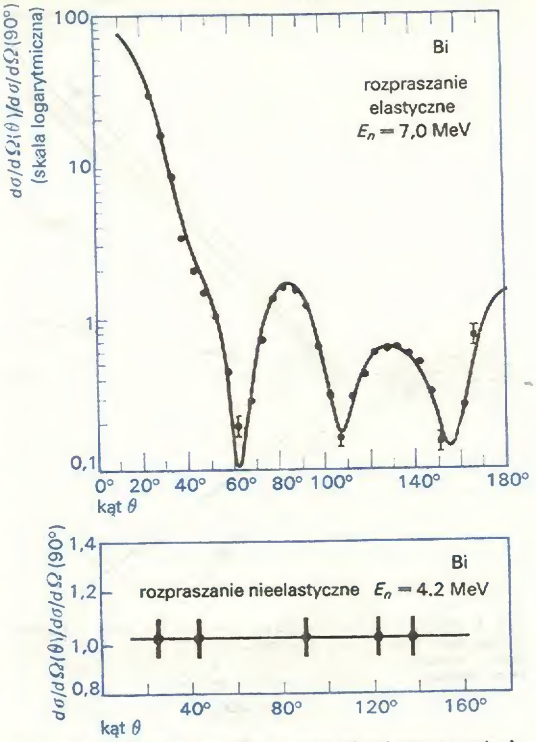
Badania reakcji jądrowych koncentrują się na pomiarach zależności przekroju czynnego od energii padających cząstek (tzw. krzywa wzbudzenia reakcji — rys. 6) oraz rozkładów energetycznych i kątowych cząstek wtórnych (rys. 7). Niekiedy cenne jest zmierzenie ukierunkowania przestrzennego spinów emitowanych cząstek (pomiar polaryzacji). Pomiar ten jest równoważny pomiarowi zależności różniczkowego przekroju czynnego od kąta azymutalnego odmierzanego względem osi wyznaczonej przez bieg wiązki cząstek o spolaryzowanych przestrzennie spinach (rys. 8).

Informacje uzyskane z takich pomiarów pozwalają wysnuć wnioski o przebiegu reakcji (jej mechanizmie).

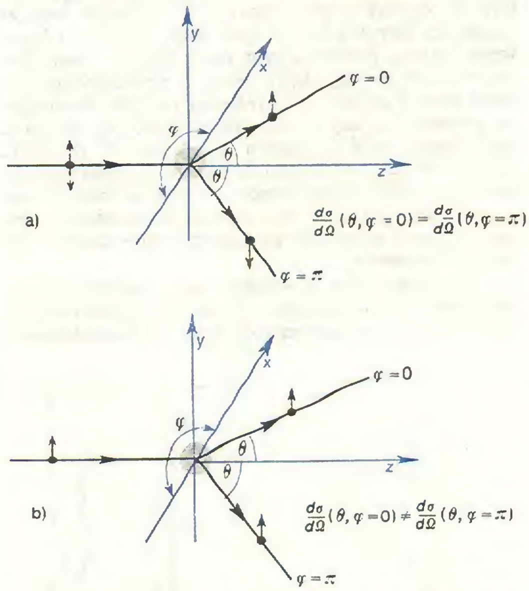


Rys. 6. Krzywa wzbudzenia rozpraszania elastycznego neutronów na jądrach ^{56}Fe. Pomiary wykonane z bardzo dobrze określoną energią padających neutronów (krzywa dolna) ujawniają jej bogatą strukturę

Gdy mechanizm reakcji jest dobrze określony, porównanie danych eksperymentalnych z przewidywaniami teoretycznymi wynikającymi z przyjętego modelu mechanizmu reakcji umożliwia poznanie struktury jądra.



Rys. 7. Rozkłady kątowe dla rozpraszania elastycznego i nieelastycznego neutronów na jądrach bizmutu. Zwraca uwagę zasadnicza różnica przebiegu obu rozkładów nasuwająca przypuszczenie, że procesy te przebiegają w całkowicie odmienny sposób



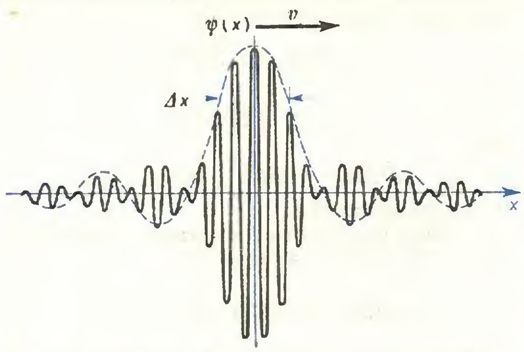
Rys. 8. Reakcja jądrowa może prowadzić do uprzywilejowania kierunku spinów emitowanych cząstek (cząstki zostają spolaryzowane) mimo, że cząstki podające są niespolaryzowane: a) Polaryzacja emitowanych cząstek jest funkcją kąta azymutalnego φ, natomiast różniczkowy przekrój czynny nie zależy od tego kąta, gdyż przy braku polaryzacji cząstek padających musi istnieć symetria procesu względem osi z. b) W przypadku istnienia polaryzacji cząstek wywołujących reakcję, różniczkowy przekrój czynny zależy od kąta azymutalnego φ.

Gdy znany jest mechanizm reakcji, badanie polaryzacji cząstek emitowanych w sytuacji a) dostarcza identycznych informacji co pomiar różniczkowego przekroju czynnego w sytuacji b)

Formalny opis reakcji jądrowych

paczka falowa

Do procesów jądrowych stosuje się opis kwantowy. Poruszająca się cząstka jest reprezentowana przez tzw. paczkę falową (rys. 9). Kwadrat amplitudy takiej



Rys. 9. Paczka falowa poruszająca się z prędkością v

fali określa prawdopodobieństwo znalezienia się cząstki w danym punkcie przestrzeni. Przykładem paczki falowej w skali makroskopowej jest kulista fala rozchodząca się po powierzchni wody po wrzuceniu do niej kamienia.

Paczka falowa powstaje przez dodanie się do siebie wielu fal o nieco różnych długościach. Jej rozmiary przestrzenne (Δx na rys. 9) zależą od przedziału długości fal tworzących ją — im przedział $\Delta\lambda$ jest większy, tym mniejsze jest rozmycie paczki Δx. Fala o długości λ przedstawia cząstkę o pędzie $p = h/\lambda$ (h — stała Plancka). Jeżeli długość fali λ nie jest ściśle określona (rozmyta), to to samo dotyczy pędu. Ogólnie, ilościowo ujmuje tę zależność zasada nieokreśloności Heisenberga

$$\Delta x \cdot \Delta p \geqslant \hbar.$$

Paczka falowa przesuwająca się z prędkością v równą prędkości poruszania się cząstki mija jakiś punkt w przestrzeni w czasie

$$\Delta t = \Delta x/v \geqslant \hbar/(v\,\Delta p),$$

przy tym

$$v\,\Delta p = \Delta E$$

($E = p^2/2m$, gdyż zakładamy, że cząstka porusza się tak wolno, że można pominąć efekty relatywistyczne; stąd — różniczkując — $\Delta E = p\Delta p/m = v\Delta p$).

Zatem z pędowo-położeniową relacją nieokreśloności wiąże się ściśle relacja czasowo-energetyczna:

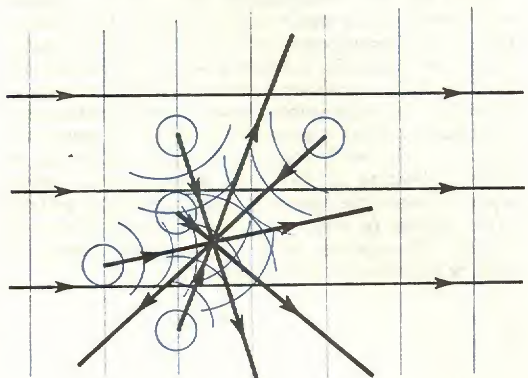
$$\Delta E \cdot \Delta t \geqslant \hbar.$$

cząstka o określonym pędzie

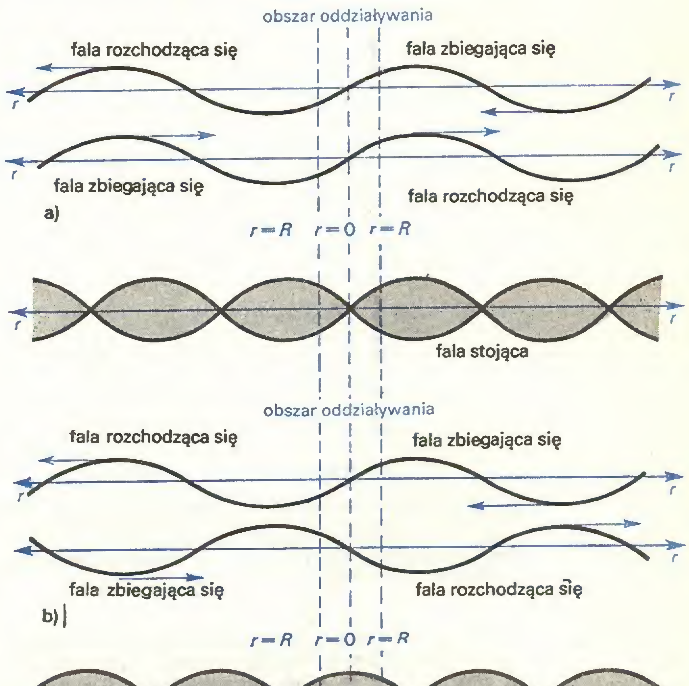
W opisie teoretycznym reakcji jądrowych wygodnie jest rozpatrywać cząstkę o dokładnie określonym wektorze pędu ($\Delta p = 0$), a tym samym — o ściśle określonej energii ($\Delta E = 0$). Jak wynika z zasady Heisenberga cząstka taka jest reprezentowana przez paczkę rozmytą na nieskończenie duży obszar — jest to po prostu fala płaska, biegnąca w kierunku wektora pędu cząstki. Innymi słowy, ponieważ wartość pędu cząstki jest dobrze określona, jej położenie jest zupełnie nieokreślone. Gęstość prawdopodobieństwa znalezienia cząstki w dowolnym punkcie jest stała i nie zależy od współrzędnych punktu. Umownie przyjmuje się, że odpowiada ona jednej cząstce w 1 cm³. Cząstki poruszają się z prędkością v, więc taka fala płaska reprezentuje strumień v cząstek/ /cm²·s. Ze względu na dobrze określoną energię cząstki również czas zajścia reakcji jest zupełnie nieokreślony. Nie ma sensu mówić o sytuacji przed zderzeniem lub po zderzeniu — opis powinien zawierać wszystkie sytuacje naraz. Strumieniowi cząstek padających na centrum rozpraszające stale towarzyszy strumień cząstek rozproszonych. Naturalnie, można wprowadzić rozmycie pędu cząstki i przez to zlokalizować ją w przestrzeni. Wówczas jest możliwe opisanie ewolucji procesu w czasie — zbliżanie się cząstek do jądra, a po oddziaływaniu — oddalanie się cząstki rozproszonej. Opis taki jest jednak bardziej skomplikowany matematycznie i przez to mniej wygodny.

fala płaska jako superpozycja fal kulistych

Można wykazać, że każdą falę płaską da się przedstawić w postaci superpozycji nieskończonego szeregu fal kulistych o odpowiednio zmodulowanych amplitudach zbiegających się do pewnego punktu w przestrzeni (np. może nim być centrum rozpraszające) oraz fal rozchodzących się z tego punktu. Łatwo to wyjaśnić stosując zasadę Huygensa, która mówi, że każdy punkt powierzchni falowej jest źródłem fali kulistej



Rys. 10. Falę płaską można rozłożyć na szereg fal zbiegających się do pewnego punktu i rozbiegających się z tego punktu. Wynika to z zasady Huyghensa



Rys. 11. Superpozycja fal: a) Nie ma przesunięcia w fazie między falą zbiegającą się i falą rozchodzącą się. Powstała w wyniku interferencji między nimi fala stojąca ma węzeł w centrum rozpraszania. b) Fala rozchodząca się jest przesunięta w fazie o $\vartheta = \pm\pi$ względem fali zbiegającej się. Powstała fala stojąca przesunięta w fazie o $\delta = \pm\pi/2$ posiada strzałkę w centrum rozpraszania. Duże jest prawdopodobieństwo przebywania cząstki w obszarze oddziaływania

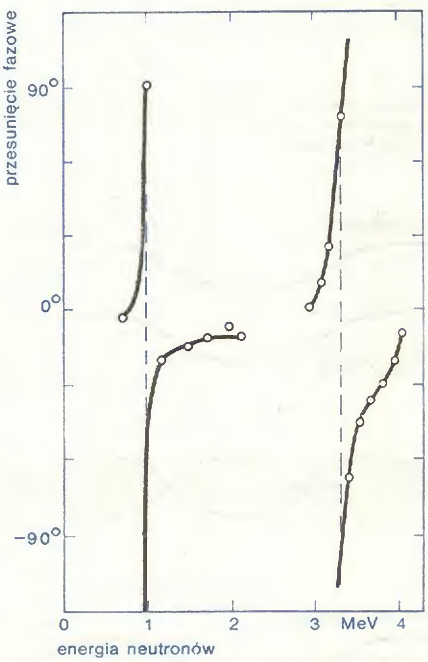
(rys. 10). Rozkład na fale kuliste jest o tyle pomocny, że, zgodnie z tym co podpowiada intuicja, pojawienie się w wyniku oddziaływania strumienia cząstek rozproszonych prowadzi do modyfikacji wyłącznie fal rozchodzących się z centrum rozpraszającego, a nie ma wpływu na fale zbiegające się.

opis za pomocą fal kulistych

Przyjmijmy, że obszar działania sił rozpraszających ma kształt kuli o promieniu R, której środek znajduje się w początku układu współrzędnych. Fala kulista wchodząca do tego obszaru ulega oddziaływaniu, w wyniku czego opóźnia się i rozchodzi się dalej z pewnym przesunięciem w fazie ϑ. Interferując z napotkaną falą zbiegającą się, tworzy falę stojącą, przesłoniętą w fazie o $\delta = \vartheta/2$ w porównaniu z sytuacją bez oddziaływania (rys. 11).

Gdy nie ma rozpraszania ($\delta = 0$), w punkcie $r = 0$ znajduje się węzeł funkcji falowej. Dla oddziaływania, któremu odpowiada strzałka drgań w punkcie $r = 0$ (czyli dla którego $\delta = \pm\pi/2$), funkcja falowa ma dużą amplitudę w obszarze oddziaływania — duże jest prawdopodobieństwo przebywania bombardującej cząstki w obszarze działania sił rozpraszających, a więc wniknięcia jej do jądra. Mówimy, że mamy wówczas do czynienia z rezonansem jądrowym. Przesunięcie fazowe δ zależy od energii padających cząstek (a więc od długości odpowiadających im fal). Okazuje się, że zależność ta jest szczególnie silna w obszarze rezonansu (rys. 12).

rezonans jądrowy



Rys. 12. Zależność od energii przesunięcia fazowego fal odpowiadających rozpraszanym cząstkom o orbitalnym momencie pędu $l = 2$ i całkowitym spinie $J = 3/2$ w reakcji $^{16}O\ (n, n)\ ^{16}O$. Przy energiach ok. 1 MeV i ok. 3,3 MeV występują rezonanse w rozpraszaniu neutronów

Weźmy pod uwagę falę rozchodzącą się od punktu $r = 0$, której faza w wyniku istnienia oddziaływania w obszarze $r \leqslant R$ ulega przesunięciu o $\vartheta = 2\delta$. Tę falę kulistą można zapisać w postaci:

$$\frac{1}{r}\exp[i(kr+2\delta(E)-Et/\hbar)] \quad \text{przy } r \geqslant R,$$

gdzie $E = \hbar^2k^2/2m$ i $k = p/\hbar$ — energia i liczba falowa cząstki o masie m i pędzie p. Z takich fal, o nieco różnych energiach, tworzy się paczka falowa, której maksimum jest w chwili $t = 0$ umiejscowione w punkcie $r = 0$. Maksimum paczki falowej znajdzie się w odległości $r = R$ wtedy, gdy w miejscu tym nastąpi konstruktywna interferencja fal tworzących paczkę. Odpowiada to warunkowi, że różnica faz fal o nieco różnych wektorach falowych k i k' jest w tym miejscu minimalna:

$$(k-k')R+2\delta(E)-2\delta(E')-(1/\hbar)(E-E')t = 0,$$

stąd dla $E' \to E$

$$t = \hbar R\, dk/dE+2\hbar\, d\delta/dE,$$

czyli

$$t = R/v+2\hbar\, d\delta/dE,$$

gdyż

$$dk/dE = m/\hbar^2k = m/\hbar p = 1/\hbar v$$

(v — prędkość cząstki).

Podobnie paczka falowa utworzona z fal zbiegających się do punktu $r = 0$ o postaci

$$\frac{1}{r}\exp[i(-kr-Et/\hbar)] \quad \text{przy } r \geqslant R,$$

reprezentująca cząstki padające, przebiega odcinek od $r = R$ do $r = 0$ w czasie $t = R/v$. Tak więc czas τ potrzebny na przejście cząstki przez całą sferę oddziaływania o średnicy $2R$ wynosi:

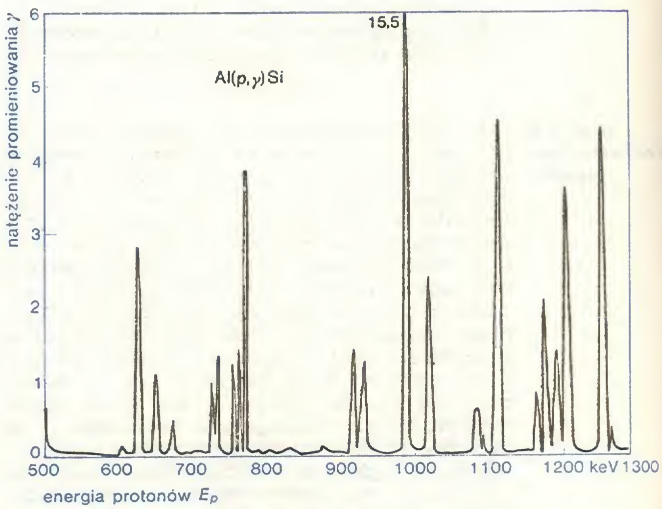
$$\tau = \tau_0+2\hbar\, d\delta/dE,$$

opóźnienie cząstki

gdzie $\tau_0 = 2R/v$ jest czasem przejścia w nieobecności oddziaływania. Oddziaływanie wprowadza opóźnienie w ruchu cząstki. Jest ono największe wtedy, gdy pochodna $d\delta/dE$ ma największą wartość, a więc np. w maksimum rezonansu (por. rys. 12). Zatem przy energiach cząstki odpowiadających rezonansowi w oddziaływaniu z potencjałem rozpraszającym, jej wyjście z obszaru oddziaływania (emisja) następuje z pewnym opóźnieniem i to tym większym, im wartość jej energii jest bliższa maksimum rezonansu. Im raptowniejsze są zmiany przesunięć fazowych przy zmianie energii (rezonans ma mniejszą szerokość energetyczną Γ — jest węższy w skali energii), tym większe jest opóźnienie.

Wartości przesunięcia fazowego wiążą się ściśle z przekrojem czynnym na rozpraszanie elastyczne. Znając je znamy również postać funkcji falowej poza obszarem oddziaływania: $\psi(\delta)$. Przez odjęcie fali płaskiej ψ_0, opisującej strumień cząstek padających, można z niej wyodrębnić falę rozchodzącą się, odpowiadającą cząstkom rozproszonym:

$$\psi_r = \psi(\delta)-\psi_0.$$



Rys. 13. Krzywa wzbudzenia reakcji wychwytu protonów przez jądra ^{27}Al w wyniku czego zostaje wysłane promieniowanie elektromagnetyczne. Występowanie wąskich rezonansów oznacza, że przy pewnych energiach emisja promieniowania jest znacznie opóźniona (tworzą się długożyjące stany wzbudzone jądra)

Znajomość wielkości amplitudy fali rozproszonej pozwala obliczyć strumień cząstek rozproszonych, a więc i przekrój czynny dla elastycznego rozpraszania. Tak więc przesunięcia fazowe, które możemy określić zakładając postać potencjału oddziaływania i rozwiązując odpowiednie równanie falowe (równanie Schrödingera), zdają sprawę z prawdopodobieństwa rozpraszania cząstki.

reakcje szybkie i ze stanem pośrednim

Reakcje jądrowe mogą przebiegać w bardzo różny sposób. Są reakcje szybkie — zachodzące w czasie bliskim czasowi przejścia cząstki przez jądro (10^{-22}–10^{-21} s), w innych reakcjach może się tworzyć stan pośredni, trwający stosunkowo długo (10^{-17}–10^{-16} s), i dopiero w trakcie jego rozpadu emitowana jest cząstka — produkt reakcji.

W świetle tego, co powiedziano poprzednio, cechą charakterystyczną reakcji szybkich, o małych opóźnieniach emisji, jest gładka zależność przekroju czynnego od energii wynikająca ze słabej zależności energetycznej przesunięć fazowych. W przypadku tworzenia się długożyjących stanów pośrednich sytuacja jest odwrotna — w krzywej wzbudzenia muszą występować silne zmiany, wąskie rezonanse, zdające sprawę z dużych opóźnień (rys. 13).

Mechanizmy reakcji jądrowych

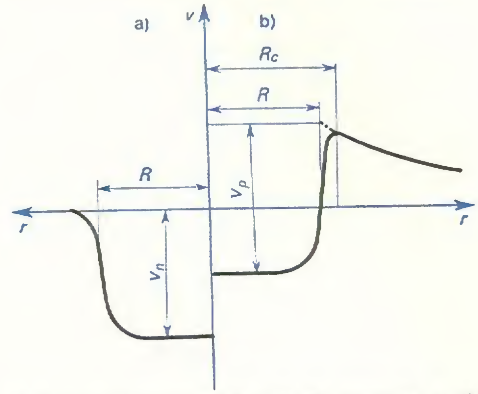
rozpraszanie elastyczne

Przykładem reakcji zachodzącej w krótkim czasie może być rozpraszanie elastyczne. Ruch cząstki padającej zostaje zakłócony przez pole sił działających wokół i wewnątrz jądra, jednakże nie wpływa na stan wewnętrzny jądra. Jądro można zastąpić po prostu pewnym obszarem działania sił. Pole sił opisuje się potencjałem jądrowym oraz, gdy cząstka ma ładunek dodatni, potencjałem odpychania elektrostatycznego. Wystarcza znajomość potencjału, aby w pełni opisać zjawisko przez rozwiązanie równania Schrödingera — różniczkowego równania na funkcję falową układu. Opisane zjawisko nie jest jednak jedynym możliwym sposobem przebiegu rozpraszania elastycznego. Cząstka, po wniknięciu do jądra, może oddziaływać z nukleonami powodując ich przegrupowanie. Jej energia rozdziela się między wiele nukleonów. Taki stan wzbudzenia trwa stosunkowo długo, aż przypadkowo energia skupi się ponownie na nukleonie lub grupie nukleonów tworzących cząstkę umożliwiając jej emisję. Jeśli energia emitowanej cząstki będzie taka sama (w układzie środka masy), jak energia cząstki padającej, będziemy znowu mieć do czynienia z rozpraszaniem elastycznym, z tym że teraz emisja nastąpi ze znacznym opóźnieniem.

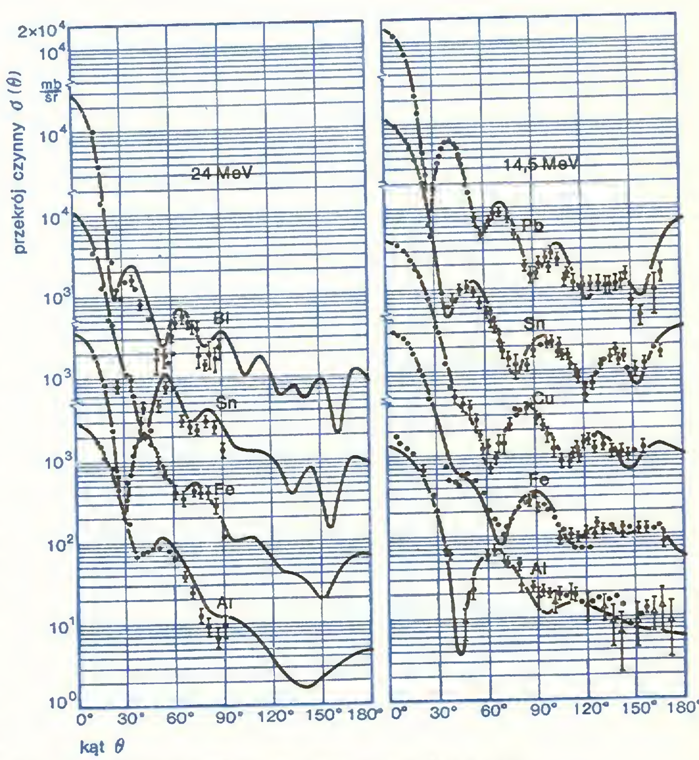
Aby wyodrębnić procesy szybkie, związane z małymi przedziałami czasowymi, należy je dobrze zlokalizować w czasie. Zlokalizowana w czasie paczka falowa odznacza się dużym rozmyciem wektorów falowych, czyli dużym rozmyciem energii. Innymi słowy, aby śledzić zachowanie się wielkości związanych z procesami szybkimi (np. przekrojów czynnych na te procesy), należy dokonać ich uśrednienia po dość znacznym przedziale energii ($\gtrsim 1$ MeV). Fale, odpowiadające emisji z jądra wzbudzonego, są tak bardzo opóźnione, że nie dają już wkładu do paczki falowej odpowiadającej szybko zachodzącemu rozpraszaniu bezpośrednio przez potencjał jądrowy. Tak więc opóźnione cząstki, tracone z punktu widzenia rozpraszania potencjałowego, są pochłaniane przez jądro. Pochłanianie padających cząstek można uwzględnić dołączając do potencjału reprezentującego jądro człon urojony. Rozwiązanie równania Schrödingera z takim potencjałem prowadzi do odpowiedniego zmniejszenia amplitudy funkcji falowych cząstek rozpraszanych, a więc też ich strumienia. Jest to pełna analogia do opisu za pomocą zespolonego współczynnika załamania przechodzenia światła przez mętną, absorbującą i rozpraszającą kulę. Ze względu na tę analogię modelowi rozpraszania elastycznego pod działaniem zespolonego potencjału nadano nazwę modelu optycznego. Wprowadzenie części urojonej potencjału odpowiedzialnej za pochłanianie jest konieczne tym bardziej, że w trakcie bombardowania jądra może zachodzić wiele innych procesów, różnych od rozpraszania elastycznego. W tych procesach cząstki mogą być zarówno wysyłane z jądra wzbudzonego, jak też mogą być wybijane bezpośrednio z jądra przez padającą cząstkę. Prowadzi to do usuwania cząstek z wiązki rozpraszanej elastycznie.

model optyczny

Stosując model optyczny można przewidywać przekroje czynne na rozpraszanie elastyczne oraz na absorpcję cząstek. Model ten jest również przydatny przy obliczaniu wielu innych wielkości stosowanych do opisu rozpraszania i pochłaniania, jak np. współczynników transmisji określających prawdopodobieństwo wniknięcia do wnętrza jądra cząstki o określonej wartości momentu pędu czy też przesunięć fazowych. Warunkiem uzyskania właściwych rezultatów jest zastosowanie w obliczeniach odpowiednio



Rys. 14. Kształt części rzeczywistej potencjału jądrowego: a) dla neutronów, b) dla protonów. R jest bliskie średniej wartości promienia rozkładu nukleonów w jądrze. R_c jest promieniem kulombowskim jądra



Rys. 15. Porównanie przewidywań modelu optycznego z odpowiednio dobranymi potencjałami (linie ciągłe) ze zmierzonymi rozkładami kątowymi neutronów elastycznie rozpraszanych na różnych jądrach (punkty)

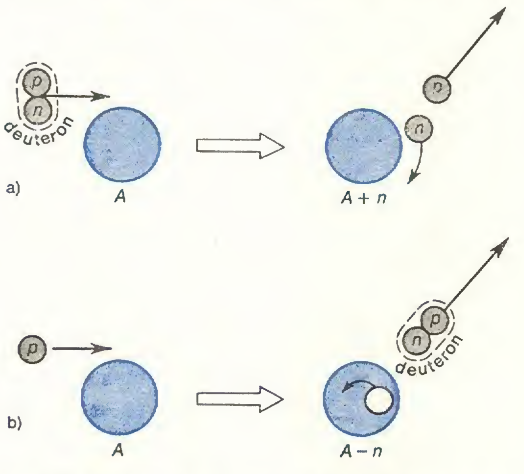
dobrze dobranego kształtu potencjału. Z reguły przyjmuje się potencjał jądrowy w kształcie prostokątnej jamy z zaokrąglonymi brzegami o rozmiarach odpowiadających rozmiarom jądra (rys. 14). Parametry opisujące kształt potencjału dobiera się przez porównanie obliczeń z doświadczalnymi rozkładami kątowymi rozpraszania elastycznego zmierzonymi dla wielu jąder i energii padających cząstek (rys. 15).

Spodziewamy się, że każda reakcja, w której bierze udział niewielka liczba nukleonów, a tym samym niewielka liczba stopni swobody (mogą to być również kolektywne stopnie swobody, np. oscylacje powierzchni jądrowej lub rotacje całego jądra), powinna być procesem szybkim (tzn. niewiele dłuższym od czasu przelotu cząstki przez jądro τ_0). Emisja cząstek następuje bezpośrednio po zderzeniu się cząstki bombardującej z jądrem. Stąd też pochodzi ogólna nazwa nadana takim reakcjom — reakcje bezpośrednie. Cechą charakterystyczną reakcji bezpośrednich jest to, że ze względu na udział niewielu stopni swobody, stan jądra utworzonego w wyniku reakcji ma wiele cech upodabniających go do stanu jądra bombardowanego. Jak się przekonamy później, fakt ten ma istotne znaczenie, gdy wykorzystuje się reakcje jądrowe do badania struktury jądra. Cząstki emitowane zachowują również w pewnej mierze kierunek padania wiązki — wyróżnione są kąty odpowiadające emisji do przodu. Cząstka padająca, oddziałując z niewieloma stopniami swobody jądra, traci na ogół stosunkowo niewiele energii. Stąd też wśród produktów reakcji przeważają cząstki o energii zbliżonej do maksymalnej.

reakcje bezpośrednie

Do najbardziej typowych reakcji bezpośrednich, obok rozpraszania elastycznego i nieelastycznego, można zaliczyć reakcję jedno- i wielonukleonowego zdarcia (reakcja „strippingu"), w której bombardująca cząstka złożona, oddziałując z jądrem, oddaje mu w czasie ruchu jeden lub kilka nukleonów, oraz reakcję jedno- lub wielonukleonowego wychwytu (reakcja „pick-up") polegającą na przechwyceniu przez cząstkę nukleonów z jądra (rys. 16). Do opisu

reakcje zdarcia i wychwytu

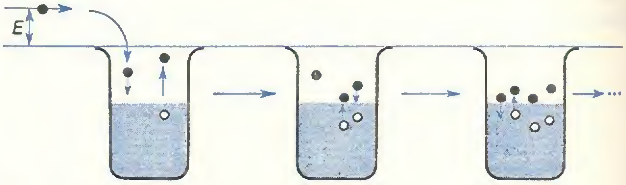


Rys. 16. Przykłady reakcji bezpośrednich: a) Reakcja zdarcia: deuteron składający się z protonu i neutronu oddaje neutron jądru bombardowanemu. Proton kontynuuje lot. b) Reakcja wychwytu: proton wyrywa neutron z jądra bombardowanego, powstały deuteron kontynuuje lot

tych procesów nie wystarcza potencjał optyczny, będący przybliżeniem pomijającym istnienie nukleonów w jądrze czy też jakąkolwiek możliwość wzbudzenia jądra. Oddziaływanie trzeba uściślić przez dodanie do potencjału dodatkowego członu, tzw. oddziaływania resztkowego. W obliczeniach teoretycznych zakłada się na ogół, że potencjał optyczny wpływa tylko na tory i pochłanianie cząstek padających i emitowanych, natomiast za samo przechwycenie nukleonów przez cząstkę lub jądro odpowiedzialne jest oddziaływanie resztkowe. Najczęściej przyjmuje się je w bardzo uproszczonej postaci ułatwiającej obliczenia (np. funkcja δ Diraca we współrzędnych przestrzennych będąca funkcją różną od zera tylko wtedy, gdy współrzędne cząstki padającej i emitowanej są identyczne). Mimo przybliżeń rezultaty obliczeń są w większości wypadków zgodne z doświadczeniem.

oddziaływanie resztkowe

Rozpatrzmy bardziej szczegółowo przebieg reakcji jądrowej (rys. 17). Cząstka, np. nukleon po wniknięciu do jądra może zderzyć się z innym nukleonem



Rys. 17. Schematyczny obraz przebiegu oddziaływania cząstki bombardującej z nukleonami jądra bombardowanego. Cząstka zderzając się z jednym z nukleonów może przenieść go w stan o wyższej energii tracąc swą energię kinetyczną E. W wyniku kolejnych zderzeń coraz więcej jest wzbudzonych nukleonów i coraz mniejsza jest średnia energia wzbudzenia przypadająca na jeden nukleon. Emisja nukleonów jest coraz bardziej utrudniona

przekazując mu część swej energii i następnie może opuścić wzbudzone jądro. Przekazana energia może być na tyle duża, że cząstka nie będzie mogła pokonać sił jądrowych i opuścić jądra — zamiast niej może być wysłany uderzony nukleon, który zyskał sporą porcję energii. Istnieje jednak taka możliwość, że zderzenie nastąpi z jednym z „głębiej" położonych, silniej związanych nukleonów. Wówczas może się zdarzyć, że żaden z obu nukleonów nie będzie miał energii wystarczającej do opuszczenia jądra. Powstaje jak gdyby „związany" stan wzbudzony, nie mogący się rozpaść przez emisję cząstki (choć całkowita energia stanu jest większa niż energia wiązania przez jądro jednego nukleonu). Stan ten charakteryzuje się istnieniem dwóch wzbudzonych nukleonów oraz luki, „dziury", w miejscu poprzednio zajmowanym przez jeden z nukleonów. Dopiero w wyniku dalszego oddziaływania między wzbudzonymi nukleonami jeden z nich może uzyskać energię dostateczną do pokonania przyciągających sił jądrowych. Jednak znacznie bardziej prawdopodobne będzie zderzenie wzbudzonego nukleonu z jednym z wielu pozostałych nukleonów jądra. W ten sposób powstanie stan, w którym istnieją trzy wzbudzone nukleony i dwie dziury. Prawdopodobieństwo emisji nukleonu jest teraz mniejsze, gdyż energia jest rozłożona na większą liczbę cząstek. W wyniku dalszych oddziaływań coraz więcej nukleonów zostaje wzbudzonych i w końcu ustala się pewna równowaga, podobna do równowagi termodynamicznej. Jest to stan jądra złożonego, który może trwać bardzo długo w skali oddziaływań jądrowych. Najbardziej dogodne rozłożenie energii między wzbudzone nukleony zachodzi przy pewnych określonych jej wartościach. Na krzywej wzbudzenia reakcji, w której tworzone jest jądro złożone, występuje szereg wąskich rezonansów odpowiadających długożyciowym stanom jądra złożonego.

stadium „dwie cząstki – jedna dziura"

jądro złożone

Z przytoczonego opisu wynika, że z powodu dwuciałowego charakteru oddziaływania jądrowego (jest to ogólnie przyjęte założenie) tworzenie jądra złożonego musi zawsze przejść przez stadium „dwie cząstki — jedna dziura". Stany takie są jakby początkiem bardziej skomplikowanych konfiguracji, dlatego też noszą nazwę „stanów wejściowych". W niektórych przedziałach energii cząstki wywołującej reakcję tworzenie pewnych stanów wejściowych może być szczególnie prawdopodobne. Spodziewamy się wówczas, że również prawdopodobieństwo tworzenia jądra złożonego będzie odpowiednio większe.

Zasadniczo emisja cząstki może nastąpić w każdym stadium rozwoju kaskady jądrowej prowadzącej do

powstania jądra złożonego. Najbardziej jest prawdopodobna w początkowej fazie rozwoju kaskady, gdy energia rozkłada się jeszcze na niewiele nukleonów. Z tego też powodu widmo cząstek emitowanych przed osiągnięciem przez jądro równowagi termodynamicznej jest widmem stosunkowo „twardym", o dużym udziale cząstek o większej energii. Wkład tego procesu do reakcji może być znaczny, szczególnie w przypadku cząstek naładowanych, gdy ich emisję z jądra złożonego dodatkowo utrudnia bariera kulombowska.

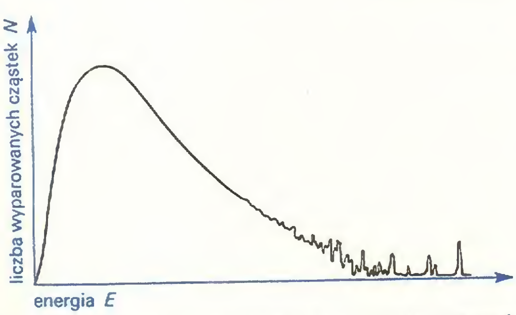
rozpad jądra złożonego

Prawdopodobieństwo emisji z jądra złożonego cząstki o określonej energii zależy praktycznie tylko od łatwości przechodzenia jej przez barierę potencjału w pobliżu powierzchni jądra. Jądro złożone żyje bowiem na tyle długo, że każda możliwa ze względów energetycznych konfiguracja wzbudzonych nukleonów pojawia się wielokrotnie. Informację o prawdopodobieństwie przejścia cząstki przez barierę daje model optyczny. Kierunek emisji cząstki jest w zasadzie dowolny, z tym że musi być zachowany całkowity moment pędu układu. Dlatego rozkład kątowy produktów reakcji niekoniecznie musi być izotropowy, ale zawsze musi być symetryczny względem kąta emisji 90°.

W miarę wzrostu energii wzbudzenia jądra złożonego rośnie liczba możliwych sposobów jego rozpadu. Zwiększa się prawdopodobieństwo rozpadu, maleje czas życia, a zatem zwiększa się energetyczna szerokość stanu jądra złożonego. Z drugiej strony szybko wzrasta liczba sposobów rozdzielenia energii wzbudzenia między nukleony — gwałtownie rośnie liczba poziomów jądra złożonego przypadająca na jednostkowy przedział energii wzbudzenia. W rezultacie poziomy jądra złożonego zaczynają się pokrywać. Pojawia się możliwość trudnej do przewidzenia i opisu interferencji między nimi, istotnie modyfikującej prawdopodobieństwo emisji. Przy dalszym wzroście energii wzbudzenia liczba pokrywających się poziomów staje się na tyle duża, że obserwowane efekty interferencyjne łatwo się uśredniają. Emisję cząstek można teraz opisać w sposób statystyczny (stąd nazwa — model statystyczny) zakładając, że jej prawdopodobieństwo zależy tylko od przenikalności przez barierę potencjału oraz gęstości osiąganych po emisji stanów jądra końcowego. Im większa jest gęstość stanów, czyli im większa jest energia wzbudzenia jądra końcowego, tym większe powinno być prawdopodobieństwo emisji. Z drugiej strony mniejsza jest wówczas energia emitowanej cząstki, co utrudnia przejście tej cząstki przez skok potencjału na powierzchni jądra. W rezultacie w widmie emitowanych cząstek, które w obszarze dużej gęstości stanów jądra końcowego jest widmem ciągłym, pojawia się dla niezbyt dużych energii maksimum odpowiadające optymalnej energii emisji (rys. 18). Można się tu doszukać daleko idącej analogii z parowaniem cząsteczek z powierzchni nagrzanej kropli cieczy. Dlatego też widmo energetyczne emitowanych cząstek często nazywa się widmem parowania.

model statystyczny

widmo parowania



Rys. 18. Widmo parowania cząstek z jądra złożonego; E energia wyparowanych cząstek, N ich liczba

Powodzenie obliczeń prowadzonych na podstawie modelu statystycznego jest w dużej mierze uwarunkowane dobrą znajomością gęstości stanów, funkcji bardzo szybko rosnącej wraz ze wzrostem energii wzbudzenia. Do wyznaczenia tej funkcji stosuje się również metody statystyczne, traktując nukleony jako cząstki poruszające się niezależnie w polu sił jądrowych (model gazu Fermiego) lub też włączając do rozważań pewne możliwości oddziaływania między nukleonami (np. tendencję do łączenia się w pary tych nukleonów, których rozkłady przestrzenne w polu sił jądrowych w dużym stopniu się pokrywają).

Jeśli w obszarze pokrywających się poziomów jądra złożonego dokładnie prześledzić zależność przekroju czynnego od energii dla ściśle określonego kanału wyjściowego reakcji, to można zauważyć fluktuacje będące wynikiem interferencji emisji z różnych poziomów. Fluktuacje te, zw. fluktuacjami eriksonowskimi (od nazwiska T. Ericsona, fizyka, który je zinterpretował), są przypadkowe i szybko nikną przy uśrednieniu po pewnym niewielkim przedziale energii. Jednakże w przedziałach energii rzędu szerokości poziomu fluktuacje nie są zupełnie niezależne, gdyż w proces emisji są zaangażowane w przybliżeniu te same poziomy jądra złożonego interferujące w określony sposób. Można zbudować „funkcję autokorelacji" przekrojów czynnych w postaci:

fluktuacje eriksonowskie

„funkcja autokorelacji" przekrojów czynnych

$$F(\varepsilon) \sim \langle\sigma(E)\sigma(E+\varepsilon)\rangle - \langle\sigma(E)\rangle\langle\sigma(E+\varepsilon)\rangle.$$

Nawiasy $\langle\ \rangle$ oznaczają uśrednienie w odpowiednio dużym przedziale energetycznym ΔE, a ε — pewne małe przesunięcie energii. Zbudowanie tej funkcji na podstawie danych doświadczalnych pozwala oszacować średnią w przedziale ΔE szerokość poziomu Γ, gdyż $F(\varepsilon) \neq 0$ tylko przy $\varepsilon \lessapprox \Gamma$. Jest to możliwość atrakcyjna, gdyż inne metody zawodzą z powodu wzajemnego pokrywania się wielu poziomów.

Statystyczne metody opisu reakcji jądrowych próbuje się również zastosować do emisji cząstek w trakcie tworzenia jądra złożonego, przed osiągnięciem przez nie pełnej równowagi termodynamicznej. Podstawą stosowanych modeli jest założenie, że jest możliwych tak wiele konfiguracji utworzonych z dwóch nukleonów i jednej dziury, trzech nukleonów i dwóch dziur itd., że można w każdym stadium rozwoju kaskady jądrowej traktować sytuację w sposób uśredniony, statystyczny. Prawdopodobieństwo emisji cząstki jest uwarunkowane przenikalnością przez barierę potencjału na powierzchni jądra oraz stosunkiem liczby możliwych konfiguracji wzbudzonych nukleonów i dziur (nadaje im się wspólną nazwę — ekscytony) pozostałych po emisji cząstki do liczby możliwych konfiguracji ekscytonów przed emisją.

Informacje o budowie jądra uzyskiwane z badania reakcji jądrowych

Strukturę jądra w wielu sytuacjach dobrze oddaje model powłokowy, wg którego średni efekt wzajemnych oddziaływań nukleonów w jądrze jest taki, że ich ruch odbywa się wzdłuż pewnych orbit scharakteryzowanych zorientowaniem w przestrzeni, kierunkiem ruchu nukleonu, jego energią, momentem pędu oraz wzajemnym położeniem spinu i momentu pędu. Wzdłuż takiej orbity może się poruszać tylko jeden nukleon (jest to konsekwencja zakazu Pauliego odnoszącego się do cząstek o spinach połówkowych; zakaz ten głosi, że w określonym stanie kwantowym może znajdować się tylko jedna cząstka). Przeniesienie nukleonu z orbity na orbitę wymaga dostarczenia energii (lub energia wydziela się, gdy przejście następuje z orbity położonej wyżej w skali energetycznej na orbitę niższą). Jądro atomowe, będące zespołem silnie wpływających na siebie nukleonów, może ulegać różnego typu wzbudzeniom. Dostarczona do wnętrza jądra energia może być skupiona na jednym nukleonie

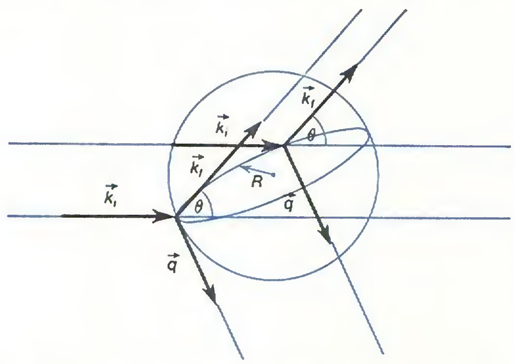
orbity w modelu powłokowym

wzbudzenia jednocząstkowe i kolektywne

(wzbudzenia jednocząstkowe) lub na kilku, może też powodować zmianę stanu wielu nukleonów. We wzbudzeniach o niedużej energii może brać udział tylko niewielka liczba nukleonów. Zdarza się, że wzbudzenie przenosi się w określony sposób z jednej grupy nukleonów na drugą. W rezultacie we wzbudzeniu bierze udział wiele nukleonów. Takie kolektywne wzbudzenia, których przykładem mogą być drgania powierzchni lub obrót jądra, wygodnie jest opisywać wprowadzając stopnie swobody związane z ich kolektywnym ruchem.

badanie reakcji zdarcia

Do badania wzbudzeń jednocząstkowych szczególnie przydatna jest obserwacja reakcji zdarcia z przekazaniem jednego nukleonu. Przekazany nukleon lokuje się na jednej z orbit bombardowanego jądra. Zidentyfikowanie tej orbity daje informację o stanie jądra utworzonego w reakcji. Identyfikacji orbity dokonuje się określając orbitalny moment pędu przechwyconego nukleonu. W tym celu wystarczy zmierzyć rozkład kątowy cząstek emitowanych w reakcji. Kątowi emisji odpowiada pewna wartość pędu przekazana wraz z nukleonem do jądra bombardowanego (rys. 19). Okazuje się, że największy wkład do prze-

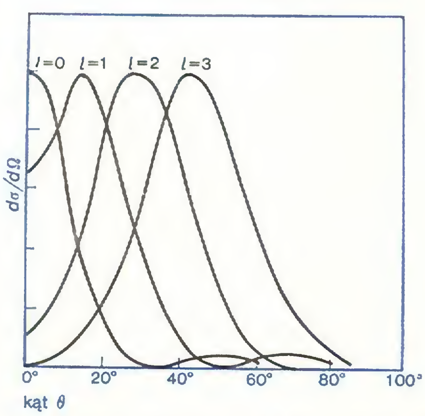


Rys. 19. Schemat obrazujący relacje między pędami cząstek biorących udział w reakcji zdarcia. Wraz z nukleonem zostaje do jądra przekazany pęd $\vec{q}(\theta) = \vec{k}_i - \vec{k}_f$. Przekrój czynny jest maksymalny, gdy reakcja zachodzi na powierzchni jądra w pobliżu płaszczyzny równikowej prostopadłej do $\vec{q}$. Wówczas moment pędu wniesiony do jądra wynosi qR

kroju czynnego dają przyczynki pochodzące od reakcji zachodzącej na powierzchni jądra w pobliżu „równika" wyznaczonego przez płaszczyznę prostopadłą do kierunku przekazanego pędu. Fale reprezentujące emitowaną cząstkę, wysyłane z różnych punktów „równika", interferują wówczas konstruktywnie dając wzmocnienie efektu. Zatem najbardziej korzystnym kątem emisji jest kąt θ, przy którym jest spełniony warunek:

rozkłady kątowe

$$l\hbar = q(\theta) \cdot R$$



Rys. 20. Typowe kształty rozkładów kątowych dla różnych momentów pędu wniesionych przez nukleon do jądra (znormalizowane w maksimum)

($l\hbar$ — moment pędu przekazanego nukleonu, $\hbar = h/2\pi$, h — stała Plancka, q — przekazany pęd, R — promień jądra). Każdej wartości orbitalnego momentu pędu odpowiada pewien charakterystyczny kąt, przy którym przekrój czynny powinien być maksymalny (rys. 20).

badanie rozkładów kątowych

Mierzone rozkłady kątowe można stosunkowo dokładnie odtwarzać w obliczeniach teoretycznych. Zakłada się, że stan wytworzony w wyniku reakcji jest czystym stanem jednocząstkowym — jądro bombardowane + nukleon. Przeważnie jednak każdy stan jest mieszaniną różnych konfiguracji wzbudzeń nukleonów. Konfiguracja: „jądro bombardowane + nukleon" jest tylko jedną z kilku składowych, z których skonstruowany jest dany stan. Można powiedzieć, że jądro spędza tylko pewną część czasu w konfiguracji jednocząstkowej. W związku z tym przekrój czynny reakcji jest odpowiednio mniejszy. Porównanie zmierzonego przekroju czynnego z obliczeniem pozwala wyznaczyć tzw. czynnik spektroskopowy, mówiący o wielkości wkładu konfiguracji jednocząstkowej do całkowitej funkcji falowej danego stanu. Z pomiaru kątowego różniczkowego przekroju czynnego na reakcję zdarcia dowiadujemy się więc o rodzaju i wkładzie konfiguracji jednocząstkowej „jądro bombardowane + nukleon" do budowy badanego stanu jądra utworzonego w reakcji. Te same wnioski tyczą się innych reakcji zdarcia i wychwytu nukleonów: wychwyt nukleonu daje informację o konfiguracji „jednodziurowej", przekazanie dwóch nukleonów — informację o konfiguracjach par cząstek czy dziur i tak dalej.

czynnik spektroskopowy

badanie reakcji bezpośrednich

Reakcje bezpośrednie dostarczają również innych cennych informacji. Na przykład z wielkości przekroju czynnego na tworzenie stanów wzbudzonych w reakcjach rozpraszania nieelastycznego można otrzymać informację o obrocie jąder zdeformowanych i wielkości ich deformacji. Przekroje czynne reakcji wzbudzenia kulombowskiego (wzbudzenie jądra przez pole elektromagnetyczne przebiegającej w pobliżu jądra naładowanej cząstki) pozwalają wysnuć wnioski na temat rozkładu ładunku jądra wzbudzonego. Tego typu przykłady można mnożyć. Wspólną cechą reakcji bezpośrednich jest dostarczanie informacji o wzbudzeniach związanych z niewielką liczbą stopni swobody.

Reakcje, w których tworzone jest jądro złożone, mogą być źródłem informacji o wysoko wzbudzonych stanach jądra (np. o ich szerokości) oraz o własnościach statystycznych jąder (np. o gęstości poziomów jądrowych). Pomiary parametrów oddzielnych rezonansów jądra złożonego w różnych kanałach wejściowych i wyjściowych umożliwiają wyciągnięcie wniosków o najbardziej istotnych w nich konfiguracjach nukleonów.

Reakcje wywołane przez ciężkie jony

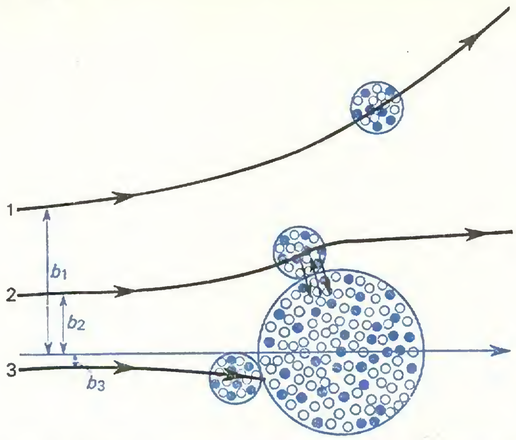
Na specjalną uwagę zasługują reakcje wywołane przez jądra atomowe, tzw. ciężkie jony, przyspieszone do znacznej energii (od kilku do kilkunastu MeV/nukleon). Cechą charakterystyczną tych reakcji jest to, że dzięki dużej masie cząstek długość odpowiadających im fal jest na tyle mała, że ruch cząstek można traktować niemal klasycznie.

Zjawiska towarzyszące bombardowaniu jąder przez ciężkie jony można bardzo ogólnie podzielić na trzy grupy (rys. 21).

rozpraszanie elastyczne lub wzbudzanie kulombowskie

1. Przy małych energiach lub przy dużych parametrach zderzenia, gdy cząstka nie wchodzi w bezpośredni kontakt z powierzchnią jądra, głównym procesem, poza rozpraszaniem elastycznym, jest wzbudzenie kulombowskie zarówno jądra bombardowanego, jak i ciężkiego jonu. Ze względu na duże ładunki oddziałujących obiektów prawdopodobieństwo tego procesu jest znaczne. Możliwy, choć mało prawdopodobny jest również proces przekazania jednego lub

kilku nukleonów, zachodzący mimo braku kontaktu obu jąder dzięki istnieniu kwantowego efektu tunelowego.



Rys. 21. Reakcje wywołane przez ciężkie jony: 1 rozpraszanie elastyczne, 2 otarcie się obu jąder z wymianą nukleonów, 3 zderzenie prawie centralne prowadzące do połączenia się obu jąder. $b_1 > b_2 > b_3$ — parametry zderzenia

2. Przy energiach przekraczających barierę kulombowską i przy odpowiednich parametrach zderzenia jon „ociera się" o powierzchnię jądra. Towarzyszy temu wymiana pewnej, czasami dość znacznej liczby nukleonów. Można rozróżnić dwa mechanizmy: zderzenia „kwazielastyczne", gdy energia rozproszonego jonu zmienia się bardzo mało w porównaniu z energią początkową, oraz zderzenia „silnie tłumione", którym towarzyszy przekazanie jądru bombardowanemu dużej energii. Przypuszcza się, że w drugim wypadku dzięki siłom jądrowym jądro bombardujące wiąże się na pewien czas z jądrem bombardowanym w układ, który szybko ulega rozerwaniu. Przekazanie dużej energii można interpretować jako efekt podobny do klasycznego efektu działania sił tarcia między dwoma ciałami.

zderzenia „kwazielastyczne"

zderzenia „silnie tłumione"

3. Przy energiach przekraczających barierę kulombowską i przy małych parametrach zderzenia następuje połączenie się obu jąder (fuzja). Dzięki swej dużej masie ciężki jon może wnieść do utworzonego jądra złożonego bardzo duży moment pędu (ok. $100\hbar$) nie spotykany w innego typu reakcjach jądrowych. W ten sposób można wytwarzać i badać jądra w bardzo nietypowych stanach. Moment pędu wnoszony do jądra jest ograniczony z góry przez warunek, aby siła odśrodkowa działająca na środki mas obu cząstek w układzie „jon stykający się z jądrem" była co najmniej skompensowana przez siły napięcia powierzchniowego stykających się kropli materii jądrowej. Z tego też względu, ze wzrostem masy bombardującego jonu maleje prawdopodobieństwo połączenia się obu jąder.

fuzja jąder

Najbardziej atrakcyjną stroną reakcji wywoływanych przez ciężkie jony jest możliwość uzyskiwania jąder położonych daleko od ścieżki trwałości oraz jąder w ekstremalnych stanach, o dużych wartościach spinów. Od reakcji tego typu oczekuje się również odpowiedzi na temat ewentualnego istnienia tzw. jąder superciężkich.

W.E. Burcham *Nuclear Physics*, London 1967; J.B.A. England *Metody doświadczalne fizyki jądrowej*, Warszawa 1981; Z. Wilhelmi *Fizyka reakcji jądrowych*, Warszawa 1976.

Jądra atomowe w stanach ekstremalnych

Zdzisław Szymański

Substancja, z której zbudowane jest jądro atomu, nie jest podobna do stanów skupienia materii znanych nam z codziennego doświadczenia. Małe rozmiary jądra oraz duża gęstość i spoistość uniemożliwiają wniknięcie w jego strukturę za pomocą prostych i bezpośrednich eksperymentów. Wiedza nasza o siłach działających pomiędzy nukleonami — podstawowymi składnikami jądra — jest bardzo niepełna. Brak też teorii obejmującej całokształt zjawisk jądrowych. Kwantowa natura praw fizyki opisujących jądro atomu obca jest naszej intuicji, będącej sumą doświadczeń z makroskopowego świata fizyki klasycznej. Wszystko to sprawia, że własności jądra atomowego są trudne do badania i nie w pełni jeszcze znane.

Otaczająca nas na Ziemi materia zawiera jądra w stosunkowo bardzo dobrze określonych stanach fizycznych. Jądra te na ogół znajdują się w stanach podstawowych. Badanie ich było pierwszym etapem poznania budowy jądra. Jednakże, obok doświadczeń zaburzających tylko w minimalnym stopniu stan podstawowy, niezwykle cennym źródłem informacji o strukturze jądra i jego własnościach jest wprowadzenie (lub próba sztucznej produkcji) jądra w stany ekstremalne, różniące się w istotny sposób od tych, które umownie potraktowaliśmy jako stany normalne.

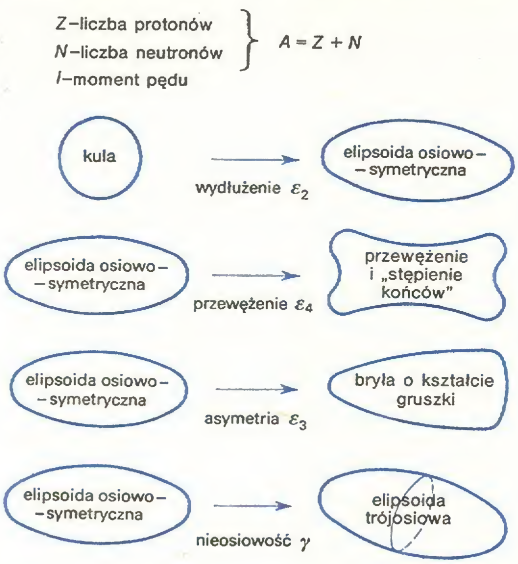
ekstremalne a normalne stany jąder

Chcąc mówić o stanach ekstremalnych jąder atomowych należy najpierw poznać własności jądra w stanie normalnym (→ Jądra atomowe i ich wzbudzenia). Możemy przyjąć, że jądro atomowe to układ kwantowy wielu nukleonów oddziałujących silnie (→ Oddziaływania silne). Istnieje pogląd, że punktem wyjściowym teorii jądra atomowego powinno być rozwiązanie równania Schrödingera dla wielu nukleonów. Obecnie podejście takie napotyka duże trudności, związane z rozwiązywaniem równania Schrödingera oraz niedostateczną znajomością oddziaływań silnych. Poza tym zgromadzony został bardzo bogaty materiał doświadczalny dotyczący jąder atomowych. Wobec tego jest rzeczą naturalną ograniczenie się do bardziej fenomenologicznego podejścia, poprzez opis jądra za pomocą wielu parametrów charakteryzujących jego własności. Tak więc obok liczby protonów Z, liczby neutronów N oraz momentu pędu I jądra (w jednostkach $\hbar$) istotną rolę odgrywają parametry charakteryzujące zmianę kształtu jądra w stosunku do kształtu kulistego, takie jak: wydłużenie ε_2 (osiowo symetryczne), parametr ε_4 — opisujący przewężenie tworzące się w powierzchni jądra rozszczepiającego się, asymetria ε_3 względem płaszczyzny prostopadłej do osi symetrii jądra, odstępstwo od kształtu osiowo symetrycznego γ i inne parametry kształtu, a ponadto parametry opisujące radialny rozkład materii w jądrze, wreszcie parametry charakteryzujące stopień struktury nadciekłej powstającej w jądrze (w wielu jądrach atomowych wytwarzają się korelacje podobne do tych, które wywołują nadciekłość helu w niskich temperaturach). Rysunek 1 charakteryzuje schematycznie parametry odgrywające istotną rolę w opisie jądra.

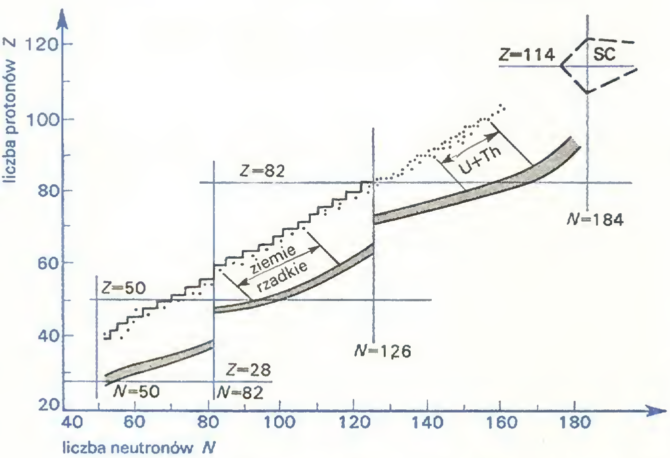
parametry charakteryzujące jądro

Jądro atomowe w stanie nieekstremalnym odznacza się pewnym stosunkiem liczby neutronów N do liczby protonów Z odpowiadającym tzw. ścieżce jąder trwałych (ścieżka trwałości → Jądra atomowe i ich wzbudzenia, rys. 2), niewielkim momentem pędu (powolny obrót niezaburzający struktury wewnętrznej jądra, np. $I \lesssim 10$), niewysokimi energiami wzbudzenia, kulistym kształtem ($\varepsilon_2 = \varepsilon_3 = \varepsilon_4 = \gamma = 0$) lub rozkładem masy odpowiadającym niewielkiej deformacji kuli, polegającym głównie na wydłużeniu, oraz występowaniem słabych korelacji typu nadciekłego (u znacznej większości jąder).

Stany ekstremalne jąder atomowych odznaczają się znacznymi odstępstwami od prawidłowości omówionych powyżej. Do jąder w stanach ekstremalnych



Rys. 1. Najważniejsze parametry charakteryzujące jądro



Rys. 2. Ścieżka jąder trwałych. Linia łamana oraz kropki określają położenie jąder trwałych. Obszary zacieniowane opisują przebieg procesu r (proces szybkiego wychwytu neutronów). Obszar zaznaczony napisem SC odpowiada hipotetycznym nuklidom superciężkim

występowanie jąder w stanach ekstremalnych

można zaliczyć jądra o nadmiarze neutronów, które mogą powstawać np. jako produkty rozszczepienia jądrowego, bądź też w niezwykle silnych strumieniach neutronów, np. w gwiazdach supernowych, jądra o znacznym niedomiarze neutronów, powstałe w wyniku reakcji syntezy ciężkich jonów, oraz jądra o szczególnie dużych liczbach protonów i neutronów, większych od znanych obecnie na Ziemi (pierwiastki transuranowe lub hipotetyczne pierwiastki superciężkie). Najcięższymi znanymi w przyrodzie pierwiastkami są tor (Th) i uran (U) o liczbach atomowych odpowiednio $Z = 90$ oraz $Z = 92$; występują one na Ziemi. Przed paru laty wykryto również na Ziemi śladowe ilości cięższego pierwiastka — plutonu (izotop ^{244}Pu). Następny pierwiastek o liczbie atomowej $Z = 93$, nazwany neptunem (Np), był wytworzony sztucznie. Izotop ^{239}Np uzyskano w procesie wychwytu neutronów przez jądro uranu ^{238}U. Powstaje wówczas cięższy izotop uranu, ^{239}U, który ulega rozpadowi β (zamiana jednego z neutronów na proton) z okresem połowicznego zaniku $T_{1/2} = 23$ min, jak również nuklid ^{239}Np, który też nie jest trwały. Kolejny rozpad β z okresem połowicznego zaniku $T_{1/2} = 2{,}3$ dnia prowadzi do powstania nowego pierwiastka, plutonu (izotop ^{239}Pu). Opisane powyżej procesy można przedstawić za pomocą następującego schematu:

$$^{238}\mathrm{U} + n \rightarrow {}^{239}\mathrm{U} \xrightarrow{\beta} {}^{239}\mathrm{Np} \xrightarrow{\beta} {}^{239}\mathrm{Pu}.$$

wytwarzanie transuranowców

Próby tworzenia nowych, coraz cięższych pierwiastków, tzw. transuranowców, są prowadzone nadal. Wytworzono i zbadano do chwili obecnej własności wielu izotopów nowych pierwiastków od neptunu (Np, $Z = 93$) aż do pierwiastka $Z = 102$ (nobel, No) oraz $Z = 103$ (lorens, Lr). W ostatnich latach badane są reakcje jądrowe (na ogół polegające na bombardowaniu ciężkich jąder innymi ciężkimi jądrami przyspieszanymi w akceleratorach ciężkich jonów) prowadzące do pierwiastków o liczbach atomowych $Z = 104$, 105, 106 oraz 107. Nuklidem o największej znanej liczbie neutronów ($N = 159$) jest obecnie ^{259}Fm (ferm). W chwili obecnej udaje się jednak wytwarzać zaledwie nieznaczne ilości tych pierwiastków, co utrudnia badanie ich własności fizycznych i chemicznych.

Inną grupę jąder ekstremalnych tworzą jądra o dużym momencie pędu. Oddziaływania typu siły Coriolisa, występujące w bardzo szybko obracającym się jądrze (np. $I \gg 10$) prowadzą do zniszczenia w nim struktury nadciekłej analogicznie do zjawiska Meissnera (→ Nadprzewodnictwo) występującego w nadprzewodzącym metalu. Przy jeszcze szybszym obrocie (np. $I \approx 50$ lub nawet $I \approx 100$) następuje znaczna modyfikacja lub całkowite przeobrażenie potencjału jądra i ruchu nukleonów w tym potencjale. Mogą również wystąpić silne deformacje jądra, polegające np. na znacznym odstępstwie od symetrii osiowej.

Jeszcze inny typ jądrowych stanów ekstremalnych związany jest z gwałtownymi zmianami kształtu jądra. Na przykład procesowi rozszczepienia jądra, towarzyszy bardzo znaczne wydłużenie jądra ε_2 związane z wyraźnym występowaniem przewężenia ($\varepsilon_4 \neq 0$), w wielu wypadkach także asymetrii ($\varepsilon_3 \neq 0$), a być może również odstępstw od osiowej symetrii ($\gamma \neq 0$). Innym przykładem, związanym zresztą najczęściej z procesem rozszczepienia, są tzw. izomery kształtu, będące krótkożyciowymi stanami jądra o znacznej deformacji prowadzącej do rozpadu przez rozszczepienie.

powierzchnia energetyczna

Szukając możliwości występowania stanów ekstremalnych skorzystamy ze znanego powszechnie faktu, że jądro atomowe, podobnie jak każdy zamknięty układ fizyczny, dąży do przyjęcia konfiguracji odpowiadającej możliwie najniższej energii. Przedstawiając zależność energii jądra od parametrów opisujących jego stan jako powierzchnię energetyczną (na ogół wielowymiarową) możemy przypuszczać, że stanowi podstawowemu jądra odpowiada absolutne minimum na powierzchni energetycznej, podczas gdy innym obserwowanym stanom, a więc odznaczającym się dostatecznie długim czasem życia, odpowiadają lokalne minima tej powierzchni. Stąd badanie własności powierzchni energetycznej, jej właściwa parametryzacja, poszukiwanie minimów itp. jest podstawową metodą teorii jądra atomowego.

Pojęcie o przebiegu powierzchni energetycznej można sobie wyrobić na podstawie znajomości własności materii jądrowej rozpatrywanej jako całość bez wnikania w szczegóły, które zależą np. od konkretnych konfiguracji nukleonów. W tym wypadku do opisu powierzchni jądrowej wystarczają na ogół prawa fizyki klasycznej (niekwantowej) oraz użycie takich wielkości charakteryzujących środowisko jądrowe, jak: energia wiązania przypadająca na jeden nukleon, energia napięcia powierzchniowego na powierzchni ograniczającej „kroplę" cieczy jądrowej, energia symetrii, która decyduje o ustaleniu się odpowiedniego stosunku ilości neutronów i protonów, energia elektrostatycznego odpychania protonów itp. W wyniku takiego opisu otrzymujemy ogólny prze-

bieg powierzchni energetycznej w postaci wolnozmiennej i na ogół gładkiej funkcji parametrów (np. dla modelu kroplowego jądra → Modele jądrowe).

Na tę gładką i regularną powierzchnię nakładają się fluktuacje wynikające z kwantowej struktury jądra atomowego — układu wielu nukleonów. Fluktuacje te nazwane ogólnie poprawkami powłokowymi „żłobią" w gładkiej powierzchni energetycznej, a zwłaszcza w tych jej okolicach, gdzie zależność energii od parametrów jest stosunkowo najmniejsza, wiele lokalnych minimów i maksimów, decydując ostatecznie o możliwości przybierania przez jądro atomowe określonych stanów. Poprawki powłokowe, jakkolwiek znacznie mniejsze od energii wyznaczonej przez ogólne własności materii jądrowej, mogą więc w sposób istotny wpływać na własności jąder. Znak poprawki powłokowej (jest ona umownie uznana za ujemną wówczas, gdy „żłobi" minimum na powierzchni energetycznej) oraz jej wielkość związane są ściśle ze strukturą widma kwantowych poziomów energetycznych układu. Jeśli poziomy energetyczne rozłożone są mniej więcej w jednakowych odległościach (widmo jednorodne), to poprawka powłokowa jest bardzo bliska zeru i o losach układu decyduje średni przebieg powierzchni energetycznej, wynikający np. z modelu kroplowego jądra. Natomiast w stanach, w których widmo poziomów energetycznych jest niejednorodne, a więc poziomy grupują się w wyraźne powłoki, poprawka powłokowa jest duża co do wartości bezwzględnej i ujemna, o ile liczba nukleonów odpowiada wypełnionym powłokom, oraz duża i dodatnia dla jąder o liczbie nukleonów odpowiadającej powłoce wypełnionej do połowy. Tak więc jądro atomowe, które dąży do przyjęcia stanu o najniższej energii, faworyzuje takie wartości parametrów (liczby neutronów N, protonów Z, momentu pędu, parametrów kształtu itp.), którym odpowiadają silnie niejednorodne widma energetyczne o zamkniętych powłokach.

poprawka powłokowa

Dlatego też poznanie kwantowych widm energetycznych jądra atomowego ma zasadnicze znaczenie w poszukiwaniu stanów ekstremalnych. Widmo poziomów energetycznych jądra można na ogół z dobrym przybliżeniem traktować jako widmo poziomów energetycznych niezależnych nukleonów poruszających się w średnim potencjale jądra. Fizycy duńscy, A. Bohr i B.R. Mottelson, zaproponowali badanie widm jąder, a więc i ocenę poprawki powłokowej przez uwzględnienie zależności energii ε niezależnego nukleonu od liczb kwantowych charakteryzujących stan nukleonu w jądrze. Okazuje się, że struktura powłokowa w widmie energetycznym może występować wówczas, gdy pochodne energii ε względem liczb kwantowych $n_1, n_2, \ldots$ mają się do siebie tak, jak niewielkie liczby całkowite:

warunek współmierności częstości

$$\frac{\partial\varepsilon}{\partial n_1} : \frac{\partial\varepsilon}{\partial n_2} : \ldots = a:b:\ldots \qquad (1)$$

Pochodna energii ε względem liczby kwantowej n_i ($i = 1, 2 \ldots$) ma prosty sens fizyczny — jest wielkością proporcjonalną do częstości podstawowej ω_i odpowiadającej i-temu stopniowi swobody nukleonu. Relację (1) można więc traktować jako warunek współmierności częstości ω_i ($i = 1, 2, \ldots$), który w języku fizyki klasycznej oznacza warunek tworzenia się orbit zamkniętych. Na przykład gdy ruch składa się z dwóch drgań harmonicznych w dwóch wzajemnie prostopadłych kierunkach „1" i „2", otrzymujemy ruch po orbicie zamkniętej tylko wówczas, gdy częstości ω_1 oraz ω_2 są współmierne; orbity mają najprostszy charakter przy stosunku $\omega_1:\omega_2$ równym np. 1:1, 2:1, 3:2, 3:1 itp. Podobna sytuacja znana jest w teorii muzyki, gdzie interwały konsonansowe odznaczają się prostymi stosunkami częstości: 1:1 (pryma), 2:1(oktawa), 3:2 (kwinta), 3:1 (duodecima) itp.

Każdemu układowi liczb całkowitych a, b, ... w (1), a więc i każdemu stosunkowi częstości odpowiada inna struktura kwantowego widma poziomów energetycznych, a więc także i orbit jednocząstkowych. Liczby cząstek odpowiadające zamkniętym powłokom (zwyrodniałym lub prawie zwyrodniałym poziomom kwantowym) określają układy o stosunkowo większej trwałości. Nazywa się je liczbami magicznymi. W fizyce atomu odpowiadają one konfiguracjom gazów szlachetnych. Stopień zwyrodnienia powłoki (tj. ilość blisko leżących poziomów energetycznych) określa więc liczby magiczne. Jeśli liczba neutronów lub protonów w jądrze w danym stanie kwantowym jest równa liczbie magicznej, to stan taki będzie faworyzowany przez jądro, a w wypadku szczególnie dużej (ujemnej) poprawki powłokowej może to prowadzić do stanu trwałego lub dostatecznie długożyciowego. Poniżej podane są przykłady ilustrujące znane bądź też hipotetyczne stany ekstremalne jąder. Między innymi uwzględniono związek tych stanów z występowaniem odpowiednich liczb magicznych, charakterystycznych dla danej konfiguracji.

liczby magiczne

Rysunek 2 przedstawia — jak już wspomnieliśmy — ścieżkę jąder trwałych, określającą typowe stosunki liczb neutronów i protonów. Nuklidy leżące z dala od tej ścieżki odznaczają się bardzo krótkimi czasami życia względem rozpadu beta (β^-, β^+ lub wychwytu elektronu) i wobec tego są na ogół bardzo trudne do obserwacji. Jądra o znacznym niedomiarze neutronów można wytwarzać doprowadzając do reakcji syntezy jąder lżejszych, np. przyspieszając ciężkie jony. Tak wytworzone jądra mogą się ponadto odznaczać bardzo dużym, przekazanym w reakcji, momentem pędu wynoszącym np. $I = 50$ lub więcej jednostek $\hbar$. Szczegółowe badanie takich nuklidów możliwe będzie tylko wówczas, jeżeli okaże się, że przy pewnych wartościach Z, N oraz I jądro może trwać dostatecznie długo, tzn. wtedy, gdy struktura widma energetycznego ma charakter wyraźnych powłok. Liczby magiczne Z, N oraz moment pędu I dla takich hipotetycznych stanów jądrowych, określanych angielskim terminem „superdizzy" (wirujące z zawrotną szybkością), nie są jeszcze znane, można jednak przypuszczać, że w najbliższych latach będą przeprowadzane poszukiwania eksperymentalne i teoretyczne tych stanów.

stan „superdizzy"

Omówimy teraz jądra charakteryzujące się nadmiarem neutronów, a więc takie, które leżą poniżej ścieżki trwałości na rys. 2. Nuklidy tego typu mogą powstawać w wyniku rozszczepienia ciężkich jąder. Ciekawym przykładem jest podwójnie magiczny izotop cyny ^{132}Sn o magicznych liczbach protonów i neutronów $Z = 50$, $N = 82$, badany obecnie w niektórych laboratoriach (magiczne liczby protonów i neutronów w jądrach o kształcie kulistym są Z, $N = 2$, 8, 20, 28, 50, 82 a także $N = 126$). Każdy inny nuklid o niemagicznych liczbach Z, N położony w tak dużej odległości od ścieżki trwałości byłby nieobserwowalny z powodu zbyt krótkiego czasu życia ze względu na rozpad β. Innym przykładem nuklidów o dużym nadmiarze neutronów są jądra występujące w tzw. procesie r, zachodzącym prawdopodobnie podczas gwałtownych ewolucji gwiazd (supernowych) produkujących niezwykle intensywne strumienie neutronów (→ Reakcje jądrowe w gwiazdach). Powstają wówczas nowe izotopy w wyniku wychwytu licznych neutronów. Czas życia ze względu na rozpad β dla tych izotopów jest porównywalny z odstępem czasu upływającym między dwoma kolejnymi przyłączeniami neutronów. Proces ten, którego przebieg jest zaznaczony na rys. 2, stanowi jedyne wytłumaczenie istnienia na Ziemi pierwiastków cięższych od ołowiu, takich jak uran (U) i tor (Th).

proces *r*

W ciągu ostatnich kilku lat bardzo żywo dyskutowana jest przez fizyków możliwość wytworzenia nowych pierwiastków o liczbach protonów Z i neutronów N, większych od znanych obecnie, tzw. pierwiastków superciężkich. Prosta ekstrapolacja znanych liczb magicznych prowadzi do wniosku, że układ złożony z $N = 126$ protonów mógłby tworzyć jądro o dużej (ujemnej) poprawce powłokowej. Jed-

pierwiastki superciężkie

114 — nowa liczba magiczna

nakże bardziej prawdopodobny wydaje się obecnie wybór pierwiastka o liczbie $Z = 114$. Aby zrozumieć, skąd się bierze nowa liczba magiczna 114, rozważmy, jak mechanika kwantowa tłumaczy przytoczone powyżej liczby magiczne jąder kulistych. Energia pojedynczego nukleonu poruszającego się w potencjale trójwymiarowego oscylatora harmonicznego dana jest wzorem:

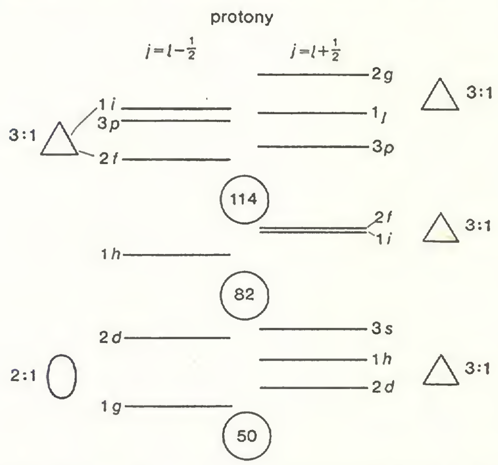
$$\varepsilon(n, l) = \hbar\omega[2(n-1)+l+3/2] \quad (2)$$

(n, l — odpowiednio radialna liczba kwantowa oraz orbitalny moment pędu). Zgodnie ze wzorem (1) otrzymujemy stosunek

$$\frac{\partial\varepsilon}{\partial n} : \frac{\partial\varepsilon}{\partial l} = 2:1 \quad (3)$$

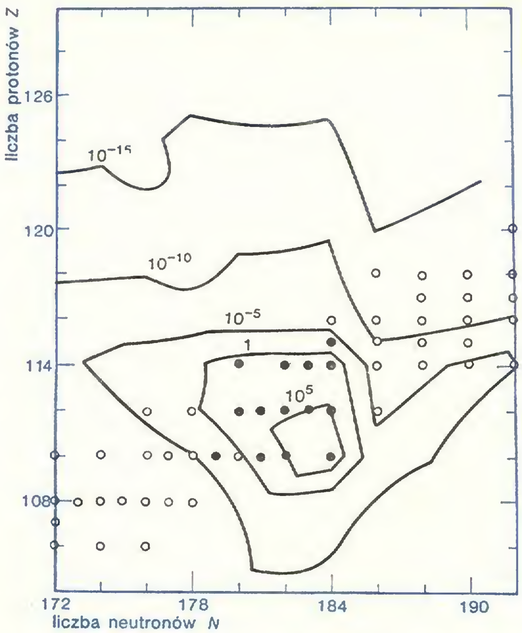
odpowiadający oktawie w terminologii muzycznej. Uwzględniając $2(2l+1)$-krotne zwyrodnienie poziomów jądrowych w jądrze kulistym związane z kwantowaniem przestrzennym i istnieniem spinu, otrzymujemy (przez sumowanie tej krotności dla każdej pary liczb n, l, prowadzących do tej samej energii $\varepsilon(n, l)$ danej wzorem 2) liczby magiczne dla kulistego jądra o potencjale oscylatora harmonicznego w postaci ciągu: 2, 8, 20, 40, 70, 112, ... Silne sprzężenie typu spin-orbita w jądrze zmienia ten ciąg prowadząc do liczb magicznych 2, 8, 20, 28, 50, 82, 126, 184, obserwowanych w rzeczywistości z wyjątkiem ostatniej. Jednakże rzeczywisty potencjał działający na nukleony w jądrze różni się od potencjału oscylatora harmonicznego, zwłaszcza dla jąder cięższych. Odstępstwa te są bardziej wyraźne dla protonów, co jest zapewne częściowo związane z istnieniem elektrostatycznego odpychania. Okazuje się, że w jądrach ciężkich stosunek $(\partial\varepsilon/\partial n):(\partial\varepsilon/\partial l)$ jest bliski wielkości 3:1, co odpowiada bardziej interwałowi duodecimy niż oktawie (2:1). W tej sytuacji grupy stanów protonowych typu ($1j$) ($2g$); ($1i$) ($2f$) oraz ($1h$) ($2d$) są bliskie zwyrodnienia (wielkość w nawiasie oznacza wartość liczby kwantowej n, litera zaś — wielkość liczby kwantowej l zgodnie ze standardową notacją spektroskopii; $l = 0, 1, 2, 3, 4, 5, 6, 7, \ldots$ oznaczane jest odpowiednio przez $s, p, d, f, g, h, i, j, \ldots$). Odnosi się to szczególnie do tej części widma, która odpowiada całkowitemu momentowi pędu nukleonu j, równemu $l+{}^1/_2$ (zgodna równoległość momentu orbitalnego i spinu protonu). Wynikający stąd układ poziomów energetycznych protonu przedstawiony jest na rys. 3.

poziomy energetyczne protonu w jądrze superciężkim



Rys. 3. Schemat poziomów protonowych jąder superciężkich Symbolami △ i ○ oznaczono te poziomy, które byłyby zwyrodniałe w przypadku spełnienia wzoru (1) dla stosunku odpowiednio 3:1, 2:1. W kółkach podane są liczby magiczne

Widać tu wyraźnie zamkniętą powłokę dla $Z = 114$. Hipotetyczne pierwiastki superciężkie występowałyby więc zgodnie z przywidywaniami teorii w okolicy $Z = 114$ oraz $N = 184$ (rys. 2). Trwałość tych jąder ograniczona jest przez możliwość rozszczepienia lub przez rozpad α i β. Rysunek 4 podaje teoretyczne oszacowania czasów życia hipotetycznych jąder su-



Rys. 4. Teoretyczne oszacowania czasów życia nuklidów superciężkich. Kropki i kółka oznaczają jądra trwałe względem rozpadu beta, zaś liczby przy konturach — czas życia w latach względem rozszczepienia i rozpadu alfa

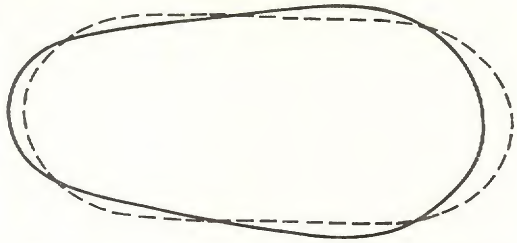
perciężkich. Jak widać, czasy te są dostatecznie długie na to, by raz wytworzone (lub odkryte w naturze) nuklidy superciężkie mogły być poddane różnym badaniom. W ciągu ostatnich kilku lat wiele szczegółowych badań poświęcono próbom znalezienia w naturze tych interesujących pierwiastków lub też próbom sztucznego ich wytworzenia. Jak na razie próby te nie powiodły się. Jest to zrozumiałe, jeśli uwzględni się fakt, że np. na Ziemi, której wiek szacuje się na około 10^9 lat, nie mogłyby pozostać nawet niewielkie ilości pierwiastków o czasach życia wyraźnie krótszych od jej wieku. Poza tym wytworzenie nuklidu o $Z = 114$ oraz $N = 184$, w drodze syntezy dwóch jąder lżejszych, jest obecnie niemal niewykonalne ze względu na fakt, że złożenie fragmentu o $Z = 114$ z dwóch ciężkich jonów dostępnych obecnie w laboratoriach prowadzi z reguły do nuklidu o N znacznie mniejszej od 184. Być może przyśpieszenie cięższych jonów, takich jak ksenon lub uran pozwoli na rozwiązanie tej trudności. Wytworzenie pierwiastków superciężkich byłoby donioslym wydarzeniem nie tylko dla fizyki jądra atomowego. Mogłoby się ono przyczynić do lepszego zrozumienia lub nawet rozwiązania wielu problemów astrofizyki, związanych ze wzajemnym przeobrażaniem się pierwiastków w gwiazdach. Byłoby krokiem milowym w fizyce atomowej i chemii, gdzie otworzyłyby się możliwości badania nowych powłok atomowych i ich wpływu na własności chemiczne ciał. Stworzyłoby także fascynujące możliwości sprawdzenia praw elektrodynamiki w warunkach ekstremalnych w związku z występowaniem bardzo dużych ładunków elektrycznych w stosunkowo niewielkim obszarze.

problem wytworzenia jąder superciężkich

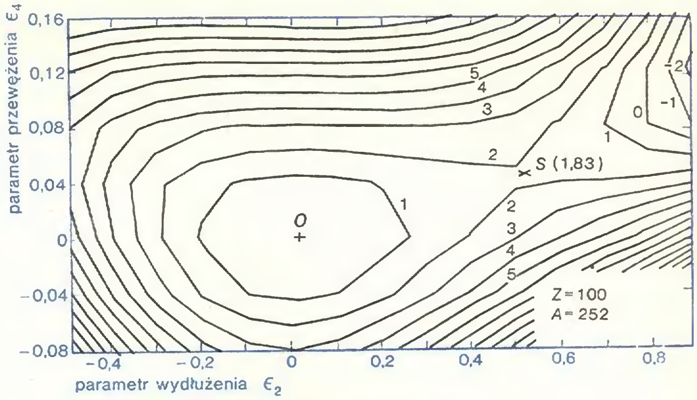
proces rozszczepienia

Innym przykładem ekstremalnego stanu jądra jest stan jądra występujący w procesie rozszczepienia. W stanie, w którym zaczyna się ten proces (punkt siodłowy na powierzchni energetycznej) jądro atomowe charakteryzuje się znacznym wydłużeniem, odpowiadającym stosunkowi osi równym w przybliżeniu 2:1. Ponadto istotną rolę odgrywa parametr ε_4,

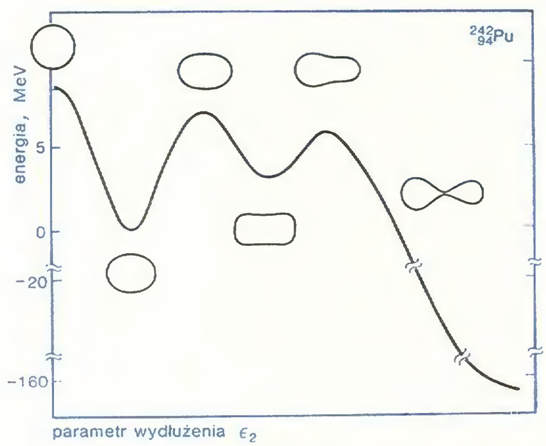
określający tworzenie się przewężenia prowadzącego do podziału jądra na dwa fragmenty. Wreszcie fakt, że wiele jąder ulega podziałowi na dwie nierówne części, wiąże się z przypuszczeniem naruszenia symetrii względem płaszczyzny prostopadłej do osi symetrii



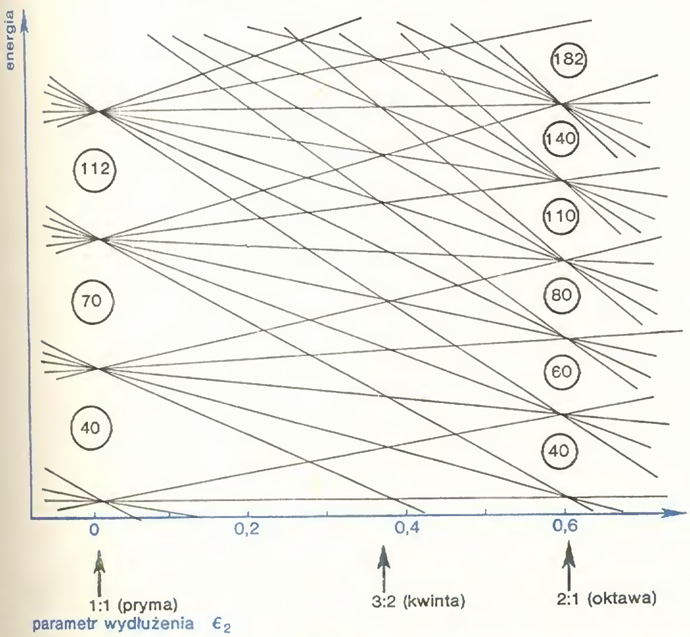
Rys. 5. Kształt jądra rozszczepiającego się asymetrycznie ($\varepsilon_3 \neq 0$) w konfiguracji siodłowej. Linia przerywana określa kształt jądra symetrycznego ($\varepsilon_3 = 0$)



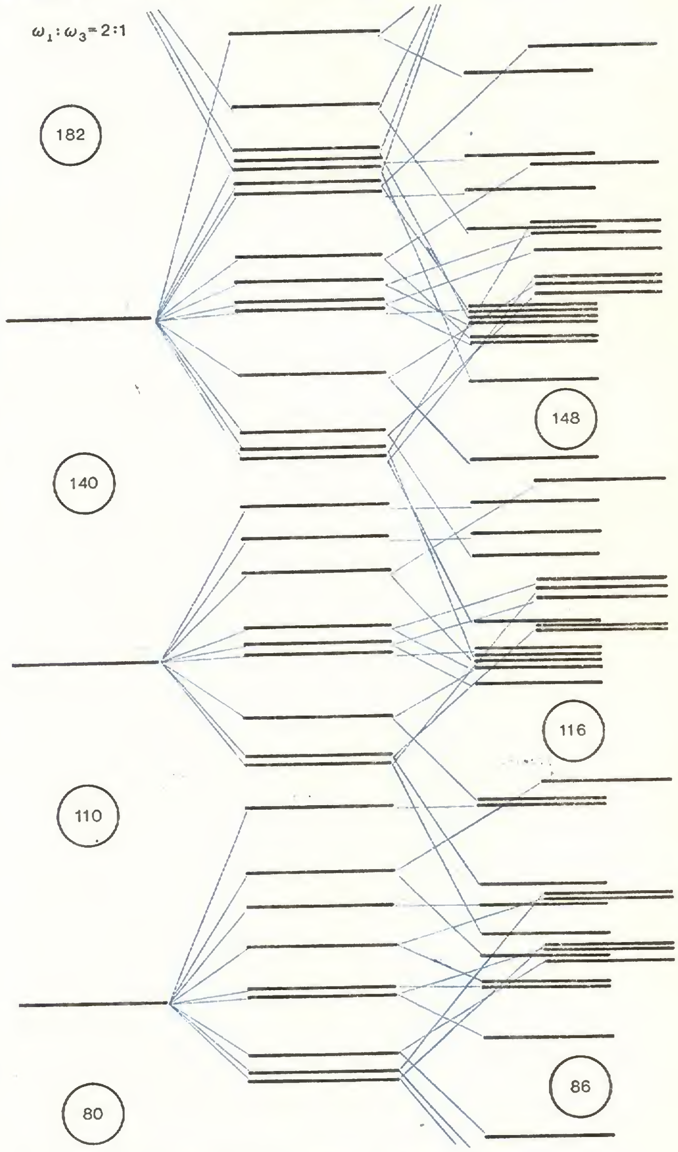
Rys. 6. Rzut powierzchni energetycznej na płaszczyznę parametrów ε_2 oraz ε_4. Kontury (warstwice) odpowiadają określonym wartościom energii. Obliczenia odnoszą się do jądra ^{252}Fm bez uwzględnienia poprawki powłokowej. Literami *O* i *S* oznaczono odpowiednie położenia konfiguracji stanu podstawowego oraz punktu siodłowego



Rys. 7. Kształt dwugarbnej bariery potencjału (energia w MeV) w rozszczepiających się jądrach, które mają izomery kształtu. Nad krzywą oraz pod zobrazowane są schematyczne kształty odpowiadające kolejno: jądru sferycznemu, stanowi równowagi, konfiguracji siodłowej pierwszego garbu, stanowi izomerycznemu, konfiguracji siodłowej drugiego garbu i wreszcie konfiguracji rozerwania jądra



Rys. 8. Poziomy energetyczne niezależnego nukleonu poruszającego się w potencjale oscylatora harmonicznego w funkcji wydłużenia ε_2. Podane są liczby magiczne odpowiadające stosunkom częstości 1:1 (pryma) oraz 2:1 (oktawa)



Rys. 9. Poziomy energetyczne niezależnego nukleonu w potencjale jądrowym dla stosunku częstości 2:1. Z lewej strony wykreślone są poziomy w potencjale oscylatora harmonicznego, z prawej zaś — dla bardziej realistycznego potencjału jądrowego. W kółkach podane są odpowiednie liczby magiczne określające zamknięte powłoki w zdeformowanym potencjale

jądra ($\varepsilon_3 \neq 0$). Tak rozszczepiające się jądro przybiera w punkcie siodłowym kształt zbliżony do gruszki (rys. 5). Konfiguracja odpowiadająca punktowi siodłowemu odznacza się wyższą energią potencjalną w porównaniu z konfiguracją stanu podstawowego. Proces rozszczepienia przebiega więc albo przez kwantowy efekt tunelowy albo wskutek dostarczenia do jądra energii w ilości wystarczającej do pokonania bariery potencjału na rozszczepienie (→ Rozpady jąder atomowych), tj. różnicy energii w obu wspomnianych konfiguracjach. Powierzchnia energetyczna, jako funkcja parametru wydłużenia ε_2 oraz parametru przewężenia ε_4, zilustrowana jest na przykładzie podanym na rys. 6, który jest mapą powierzchni energetycznej (linie ciągłe są warstwicami, tj. odpowiadają stałej energii). Na rys. 6 zaznaczone są również położenia konfiguracji stanu podstawowego i punktu siodłowego (oznaczone odpowiednio literami *O*, *S*).

Okazuje się jednak, że wiele spośród jąder ciężkich ulegających rozszczepieniu wykazuje znacznie bardziej skomplikowaną strukturę powierzchni energetycznej w okolicy punktu siodłowego. Obok minimum w stanie podstawowym występuje w tych jądrach drugie minimum, odpowiadające bardziej dużemu wydłużeniu o stosunku osi rzędu 2:1. Jądra w takiej konfiguracji żyją bardzo krótko i rozpadają się przez rozszczepienie z czasem życia o kilka lub kilkanaście rzędów wielkości krótszym niż w wypadku rozszczepienia zaczynającego się w stanie podstawowym. Takie stany ekstremalne zwane są izomerami kształtu. Były one obserwowane u wielu jąder grupy aktynowców. Istnienie tego typu stanów możliwe jest jedynie wówczas, gdy w barierze na rozszczepienie „wyżłobione" jest dodatkowe minimum, tak jak to przedstawia rys. 7.

izomery kształtu

Wyjaśnienie występowania izomerów kształtu na gruncie badania struktury widm energetycznych jest ciekawym przykładem zastosowania warunku (1) Bohra i Mottelsona. Energia pojedynczego nukleonu poruszającego się w potencjale anizotropowego oscylatora harmonicznego dana jest wzorem

$$\varepsilon = \hbar\omega_\perp(n_\perp+1)+\hbar\omega_3(n_3+{}^1/_2), \qquad (4)$$

gdzie $\omega_\perp$ oraz ω_3 oznaczają dwie różne częstości charakteryzujące kształt potencjału, zaś $n_\perp$ i n_3 — odpowiednie liczby kwantowe. Rysunek 8 przedstawia wykres poziomów energetycznych w funkcji wydłużenia ε_2. Widać, że oprócz struktury powłokowej, odpowiadającej kształtowi kulistemu (1:1), istnieje wyraźna struktura powłokowa dla $\omega_\perp : \omega_3 = 2:1$ (oktawa), czyli stosunku osi równego 1:2. Liczby magiczne 2, 4, 10, 16, 28, 40, 60, 80, 110, 140, 182... odpowiadają w tym wypadku potencjałowi czystego oscylatora harmonicznego. Istnienie odstępstw od potencjału oscylatora harmonicznego prowadzi do modyfikacji tego prostego schematu (rys. 9). Wskutek tego w rzeczywistych jądrach należy oczekiwać istnienia liczb magicznych... 86, 116, 148... Właśnie w pobliżu liczby neutronów $N = 148$ wykryto szereg izomerów kształtu.

Omówione powyżej przykłady hipotetycznych lub obserwowanych jądrowych stanów ekstremalnych nie wyczerpują wszystkich możliwości. Można oczekiwać, że w przyszłości wysunięte zostaną nowe hipotezy i odkryte będą nowe ekstremalne stany jąder atomowych. Będą one niewątpliwie cennym źródłem dalszych informacji o strukturze i własnościach jąder atomowych.

A. Bohr, B.R Mottelson *Nuclear Structure*, Vol 2, Massachusetts (London–Amsterdam–Don Mills, Ontario–Sydney, Tokyo) 1975; E.K Hyde i in. *The Nuclear Properties of The Heavy Elements*, Englewood Cliffs, New Jersey 1967; S.G. Nilsson *Shapes and Shells* w *Scuola Internationale „Enrico Fermi"* 1974, Amsterdam 1956.

Fizyka ciężkich jonów

Adam Sobiczewski

ciężkie jony

Ciężkimi jonami nazywamy, dosyć umownie, jony (tzn. zjonizowane atomy) pierwiastków cięższych od helu, tj. pierwiastków o liczbie atomowej $Z > 2$. Przyspieszane w akceleratorach do różnych energii, znajdują one zastosowanie w wielu dziedzinach nauki (głównie fizyka jądrowa, ale także atomowa, ciało stałe, medycyna i in.) oraz techniki. Jądra ciężkich jonów rozpędzane do energii o bardzo szerokim zakresie — od niskich do bardzo wysokich — mogą wywoływać w jądrach tarczy różnorodne procesy, trudne lub niemożliwe do osiągnięcia za pomocą cząstek lżejszych, jak elektrony, miony czy mezony, a nawet protony, deuterony czy jądra helu. Badanie tych procesów dostarcza jedynej w swoim rodzaju wiedzy o jądrach, ich strukturze i przemianach.

Konieczność przyspieszania jonów sprawia, że rozwój fizyki ciężkich jonów jest ściśle związany z rozwojem akceleratorów. Pierwszą wiązkę jonów ciężkich (węgla ^{12}C) — o natężeniu jeszcze bardzo małym, ale wystarczającym już do badania niektórych reakcji jądrowych — otrzymano w 1950 r. w USA (Berkeley). Był to początek zrazu powolnego, a ostatnio bardzo burzliwego rozwoju fizyki ciężkich jonów. Obecnie potrafimy już przyspieszać jony wszystkich pierwiastków, a budowa coraz większych akceleratorów umożliwia przyspieszanie ich do coraz wyższych energii i z coraz większym natężeniem wiązki (→ Akceleratory).

Szybki i bardzo wszechstronny rozwój fizyki ciężkich jonów sprawia, że z pomocniczej uprzednio techniki zaczyna ona wyrastać na samodzielny i ważny dział fizyki jądrowej. Wydaje się, że w najbliższych latach będzie to jeden z działów najbardziej dynamicznych i bogatych w wyniki.

Do tych kilku uwag dodajmy następującą uwagę terminologiczną. Ponieważ w reakcjach jądrowych powłoka elektronowa jonu nie odgrywa prawie żadnej roli, przeto gdy w fizyce jądrowej mówimy o ciężkim jonie czy reakcji z nim, mamy z reguły na myśli samo jego jądro czy odpowiednio reakcję z tym jądrem. Skrótu tego nie używa się w innych działach fizyki, gdzie powłoka elektronowa jonu jest istotna.

specyfika reakcji z ciężkimi jonami

Reakcje wywołane przez ciężkie jony mają kilka cech szczególnych, wyróżniających je spośród innych reakcji jądrowych. Cechy te decydują w dużej mierze o znaczeniu reakcji. Wymienimy główne z nich:

1) Duży ładunek elektryczny jądra jonu. Daje to możliwość oddziaływania na jądro tarczy silnym polem elektrycznym (wzbudzenie kulombowskie — patrz niżej). Pozwala także na utworzenie, choćby na krótko (mianowicie na okres czasu, w którym jądra ciężkiego jonu i tarczy przebywają w bezpośrednim sąsiedztwie), układu o bardzo dużym ładunku. Dostarcza to szansy zbadania powłoki elektronowej odpowiadającej jądru o ładunku, którego nie można otrzymać w inny sposób (atomy superciężkie).

2) Duża masa jonu. Zderzenia jąder o dużej masie umożliwiają przekazanie z jednego jądra do drugiego wielu nukleonów, lub wymianę ich między jądrami. Oznacza to możność wytwarzania nuklidów znacznie różniących się od naturalnych lub wytwarzanych w reakcjach z lekkimi cząstkami (nuklidy o dużym niedomiarze lub nadmiarze neutronów, nuklidy bardzo ciężkie lub nawet superciężkie).

3) Duży pęd jonu. Wskutek tego duży jest także pęd przekazany produktom reakcji, co z kolei sprawia, że produkty te „wybijane" są z tarczy i mogą być niemal natychmiast poddane badaniom. Jest to

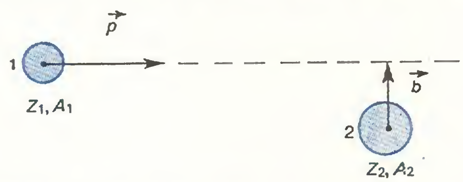
szczególnie ważne przy badaniu produktów krótkożyciowych. Duża prędkość produktów umożliwia także stosowanie wielu specjalnych technik detekcyjnych i identyfikacyjnych (np. wykorzystujących analizatory magnetyczne z selektorem prędkości, które pozwalają na określenie zarówno masy, jak i ładunku produktu, czy też detektorów śladowych).

4) Duży moment pędu jonu względem jądra tarczy. Pozwala on na wytwarzanie jąder złożonych w stanach o dużym momencie pędu (spinie). Badanie jądra w takim stanie umożliwia poznanie wpływu szybkości obrotu jądra na jego strukturę i własności.

Duże pędy i momenty pędu występujące w reakcjach z ciężkimi jonami pozwalają przy tym na korzystanie w dużej mierze z klasycznego i półklasycznego ich opisu.

Rodzaje reakcji z ciężkimi jonami i ich mechanizm

O przebiegu procesu zderzenia ciężkiego jonu *1* z jądrem tarczy *2* decyduje kilka wielkości. Są to (rys. 1): liczby atomowe Z_1, Z_2 obu jąder (a więc i ładunki



Rys. 1. Wielkości charakteryzujące zderzenie ciężkiego jonu *1* z jądrem tarczy *2*

Z_1e i Z_2e, gdzie e — ładunek protonu), ich liczby masowe A_1, A_2, pęd p lub energia E padającego jonu (jądro tarczy na początku procesu zderzenia spoczywa) oraz parametr zderzenia b. Parametr ten jest odległością środka jądra tarczy od prostej będącej przedłużeniem wektora pędu czy prędkości początkowej padającego jonu. Na przykład $b = 0$ oznacza, że jon pada wprost na jądro tarczy (tzw. zderzenie czołowe). Łącznie z wektorem $\vec{p}$ parametr zderzenia określa moment pędu jonu względem jądra tarczy, $\vec{l} = \vec{b} \times \vec{p}$, który jest zachowany w czasie całego procesu zderzenia.

bariera kulombowska

Liczby Z_1 i Z_2 razem z liczbami masowymi A_1 i A_2, które określają wymiary jąder (→ Jądra atomowe i ich wzbudzenia), wyznaczają wartość bariery potencjału między jonem i jądrem tarczy. Pochodzenie bariery wyjaśnia schematycznie rys. 2. Z rys. 2a widać, że potencjał składa się z dwu części: długozasięgowej, dodatniej $U_{\text{kul}}(r)$, opisującej odpychanie elektrostatyczne (kulombowskie) między jądrami, oraz krótkozasięgowej, ujemnej $U_j(r)$, opisującej przyciąganie czysto jądrowe (→ Siły jądrowe). Wobec krótkiego zasięgu sił jądrowych przyciąganie występuje dopiero przy bardzo małej odległości między jądrami, niemal wtedy, gdy „stykają się" one swoimi powierzchniami. Jest jednak silniejsze od odpychania kulombowskiego, toteż nałożenie obu oddziaływań daje barierę potencjału (barierę kulombowską) przedstawioną na rys. 2b. Jest to ta sama bariera, którą musi pokonać cząstka naładowana w procesie odwrotnym do opisywanego tutaj (tzn. w procesie wydostania się z jądra), np. cząstka α przy rozpadzie α (→ Rozpady jąder atomowych). Jeśli przynajmniej jedna ze zderzających się cząstek nie jest naładowana elektrycznie (np. neutron), bariera taka nie istnieje.

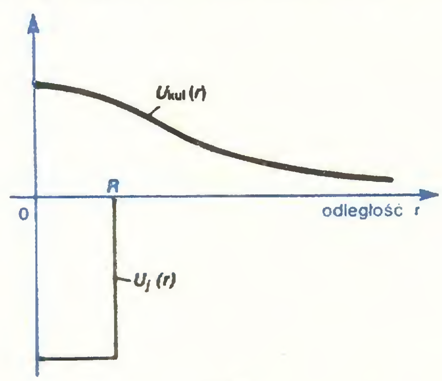
Wysokość bariery kulombowskiej można łatwo obliczyć. Zgodnie z rys. 2

$$E_{\text{kul}} = U(R) = \frac{(Z_1e)(Z_2e)}{R}, \tag{1}$$

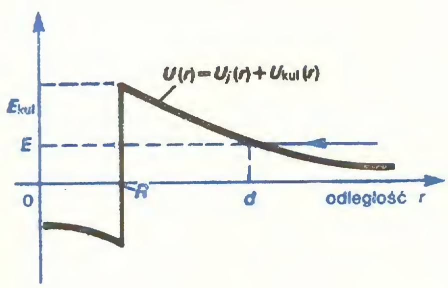
gdzie R jest (w przybliżeniu) odległością między środkami jąder przy ich „zetknięciu", a więc sumą ich promieni

$$R \approx R_1 + R_2 = r_0(A_1^{1/3} + A_2^{1/3}). \tag{2}$$

We wzorze (2) $r_0 = 1{,}3$–$1{,}5$ fm. Ponieważ energię

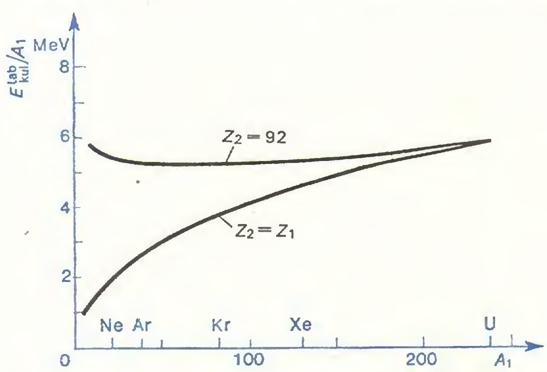


a)



b)

Rys. 2. Objaśnienie pochodzenia bariery kulombowskiej (a) oraz wykres bariery (b)



Rys. 3. Energia, jaką musi mieć ciężki jon, aby pokonać barierę kulombowską jąder tarczy; przeliczona na jeden nukleon jonu i wykreślona w funkcji jego liczby masowej A_1

energia na pokonanie bariery

padającego jonu podaje się zwykle w przeliczeniu na jeden jego nukleon, tzn. E/A_1, wygodnie jest barierę kulombowską podawać również w takim przeliczeniu, tzn. E_{kul}/A_1, a zamiast podawać samą barierę kulombowską E_{kul}/A_1 lepiej podać energię $E_{\text{kul}}^{\text{lab}}/A_1$, jaką ciężki jon musi mieć w układzie laboratoryjnym, by mógł pokonać barierę kulombowską jąder tarczy. Energia $E_{\text{kul}}^{\text{lab}}$ jest barierą kulombowską E_{kul} powiększoną o energię ruchu środka masy. Na rys. 3 podane są dwie krzywe przedstawiające $E_{\text{kul}}^{\text{lab}}/A_1$ w funkcji A_1: gdy $Z_2 = Z_1$ oraz gdy $Z_2 = 92$. W drugim wypadku tarcza wykonana jest z najcięższego pierwiastka występującego w naturze, tzn. z uranu. Wartości $E_{\text{kul}}^{\text{lab}}/A_1$ zawarte między obiema krzywymi wyczer-

pują zatem niemal wszystkie kombinacje Z_1 i Z_2, jakie mogą występować (ze względu na symetrię wzoru (1) przy zamianie Z_1 i Z_2 można rozważać tylko wypadek, gdy $Z_2 > Z_1$). Z rys. 3 wynika np., że przy zderzeniu C+C (węgiel z węglem) wystarczy energia ok. 1 MeV/nukleon, by jony węgla mogły pokonać barierę kulombowską, podczas gdy w reakcji C+U (węgiel z uranem) trzeba już ponad 5 MeV/nukleon. Energia zaś $E/A_1 = 6–7$ MeV/nukleon wystarcza w zasadzie do pokonania każdej bariery kulombowskiej.

Obecność bariery kulombowskiej w zderzeniach ciężkiego jonu z jądrami narzuca podział tych zderzeń na dwie zasadnicze klasy: zderzeń podbarierowych, gdy $E < E_{kul}$, i nadbarierowych, gdy $E \geqslant E_{kul}$. W pierwszym wypadku jon o energii E może się zbliżyć do jądra najwyżej (przy zderzeniu czołowym) na odległość $d > R$ (rys. 2b), może więc spowodować tylko proces, w którym występuje oddziaływanie kulombowskie (elastyczne rozpraszanie kulombowskie i wzbudzenie kulombowskie), a nie jądrowe. W wypadku drugim jon może pokonać barierę kulombowską, dostaje się wtedy do obszaru oddziaływania jądrowego i powoduje różnorodne reakcje jądrowe.

zderzenia podbarierowe

zderzenia nadbarierowe

Ze względu na dużą rozpiętość energii ciężkich jonów dogodnie jest wydzielić przynajmniej trzy zakresy: energii niskich, pośrednich i wysokich. Odpowiadające tym zakresom procesy są bardzo różne.

Niskie, to energie do ok. 10 MeV/nukleon, a więc energie zarówno podbarierowe, jak i nadbarierowe, ale niezbyt odległe od wartości bariery kulombowskiej. Ten zakres energii jest najlepiej zbadany, ponieważ większość obecnych akceleratorów przyspiesza jony do takich właśnie energii. Wysokie, to energie od kilkuset do kilku tysięcy MeV/nukleon. W tym zakresie jest na razie stosunkowo mało wyników doświadczalnych. Wreszcie w zakresie energii pośrednich brak, jak dotąd, wyników doświadczalnych. Jedynie pod względem teoretycznym jest on w pewnym stopniu opracowany.

Zgodnie z tym stanem rzeczy najbardziej szczegółowo omówimy zakres energii niskich.

Energie niskie

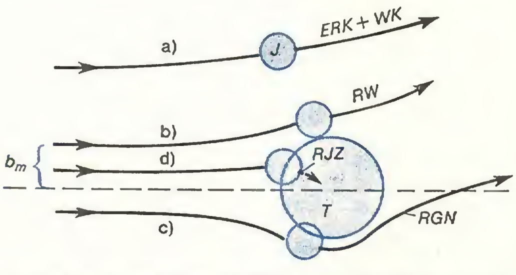
Przy energii nadbarierowej rodzaj procesu wywołanego ciężkim jonem zależy w istotny sposób od wartości parametru zderzenia b. Przy niskiej energii mogą zajść następujące procesy (rys. 4):

a) Przy dużych wartościach b, mianowicie gdy $b > b_m$ (zderzenia odległe), jon nie dostaje się do strefy oddziaływania jądrowego i może oddziaływać z jądrem tylko kulombowsko. Przez to oddziaływanie może on doznać jedynie rozproszenia elastycznego (w którym ani jądro tarczy, ani jon nie zostają wzbudzone) lub nieelastycznego (w którym bądź jądro tarczy, bądź jon, bądź też jedno i drugie zostaje wzbudzone — wzbudzenie kulombowskie).

zderzenia odległe

b) Przy wartości $b \approx b_m$ jon dostaje się zaledwie na granicę oddziaływania jądrowego; zachodzi wówczas oddziaływanie tylko między kilkoma nukleonami na powierzchni obu jąder. Takie zderzenie „muśnięcia" prowadzi do tzw. reakcji wprost, lub inaczej — reakcji bezpośrednich.

„muśnięcie"



Rys. 4. Różne rodzaje zderzeń ciężkiego jonu J z jądrem tarczy T, w zależności od wartości parametru zderzenia b: a) elastyczne rozpraszanie kulombowskie (ERK) i wzbudzenie kulombowskie (WK), b) reakcja wprost (RW), c) reakcja głęboko nieelastyczna (RGN), d) reakcja przez jądro złożone (RJZ)

c) Przy mniejszych b zderzenie ma charakter głębszy, więcej nukleonów jest w nim zaangażowanych. Następuje przekazanie znacznej ilości energii i momentu pędu od ruchu względnego do wewnętrznych stopni swobody obu jąder, a także wymiana pewnej liczby nukleonów między jądrami. Takie zderzenie nazywa się głęboko nieelastycznym. Jądra po zderzeniu (jądra końcowe) różnią się od jąder przed zderzeniem (jądra początkowe) bardziej niż w reakcjach wprost.

zderzenia głęboko nieelastyczne

d) Dopiero przy dostatecznie małych b energia i moment pędu ruchu względnego mogą w pełni zostać przekazane wewnętrznym stopniom swobody. Powstaje wówczas układ silnie wzbudzony, w którym energia rozłożona jest między wszystkie nukleony. Mówimy, że zachodzi synteza (lub fuzja) zderzających się jąder lub że zostaje utworzone jądro złożone. Jądro to „zapomina" niejako o swojej historii, sposób jego późniejszego rozpadu nie zależy od sposobu jego powstania.

powstanie jądra złożonego

Omówimy główne cechy i mechanizm poszczególnych procesów.

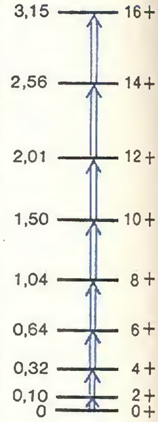
Elastyczne rozpraszanie kulombowskie, zwane także rozpraszaniem rutherfordowskim, jest najprostszym z wymienionych procesów. Nie następuje tu żadna zmiana energii ruchu względnego na energię wewnętrzną jąder. Pod wpływem odpychania kulombowskiego jon zostaje odchylony o kąt, który jest prostą funkcją energii padania E i parametru zderzenia b. Wzór na różniczkowy przekrój czynny rozpraszania otrzymać można zarówno na gruncie mechaniki klasycznej, jak i kwantowej. Jest to jeden z nielicznych wzorów, które w obu wypadkach są identyczne. Proces rozpraszania rutherfordowskiego nie daje jednak prawie żadnej (oprócz ładunku) informacji o jądrach tarczy.

elastyczne rozpraszanie kulombowskie

Wzbudzenie kulombowskie jest nadal stosunkowo jeszcze prostym i dobrze zrozumianym procesem; nie występuje w nim bowiem oddziaływanie jądrowe. Jest przy tym stosunkowo łatwe do zrealizowania w praktyce — stąd szerokie jego stosowanie. Przelot ciężkiego jonu w pobliżu jądra tarczy poddaje jądro silnemu zmiennemu polu elektrycznemu (siła oddziaływania dochodzi do kilkudziesięciu kG!); pole powoduje dynamiczną deformację jądra lub (jeśli ma ono kształt trwale zdeformowany, jak np. jądra obszaru ziem rzadkich lub aktynowców) jego obrót. Prowadzi to do wzbudzeń kolektywnych typu wibracyjnego (drgania) lub rotacyjnego (obrót; → Modele jądrowe). Najłatwiej zachodzą wzbudzenia kolektywne o niskiej energii i małej multipolowości (tzn. o małym spinie), głównie kwadrupolowe (spin 2). Łatwość i jednocześnie „czystość" wzbudzeń była powodem, że z zastosowaniem ciężkich jonów zostały zbadane stany kolektywne niemal wszystkich dostępnych jąder. Wyznaczono energie stanów, prawdopodobieństwa ich wzbudzenia (elektromagnetycznego), spiny i parzystości, a więc podstawowe wielkości charakteryzujące stan wzbudzony i jego relację ze stanem podstawowym.

wzbudzenia kulombowskie

Prawdopodobieństwo wzbudzenia kulombowskiego rośnie ze wzrostem energii padającej cząstki, rośnie też szybko ze wzrostem jej masy. Jeśli np. przy jakiejś energii prawdopodobieństwo wzbudzenia za pomocą elektronów jest bardzo małe, to za pomocą protonów o tej samej energii jest znaczne, a za pomocą ciężkiego jonu — duże. Prawdopodobieństwo wzbudzenia stanu kolektywnego przeciętnego jądra z obszaru ziem rzadkich jonem ^{16}O jest kilka tys. razy większe niż protonem — przy takiej samej energii jonu na jeden jego nukleon jak energia protonu. Fakt ten zdecydował o powszechności stosowania ciężkich jonów do wzbudzania kulombowskiego. Pozwolił on także na realizowanie za pomocą ciężkich jonów



Rys. 5. Ośmiokrotne wzbudzenie kulombowskie jądra ^{170}Hf. Z lewej strony podane są energie stanów w MeV, z prawej — spin i parzystość

wielokrotnego (lub wielostopniowego) wzbudzenia kulombowskiego. W procesie tym w czasie przelotu jonu w pobliżu jądra tarczy jądro doznaje kilku kolejnych wzbudzeń. Rys. 5 podaje przykład, w którym jądro ^{170}Hf doznaje ośmiu wzbudzeń, a każde z nich podwyższa kolejno jego spin o 2 i o pewną wartość energii. Stosując odpowiednio ciężkie jony można drogą wielokrotnego wzbudzenia kulombowskiego osiągać bardzo wysokie spiny. Należy przy tym zwrócić uwagę, że prawdopodobieństwo osiągnięcia stanu 4+ bezpośrednio ze stanu 0+ (wzbudzenie jednokrotne) jest znacznie mniejsze niż pośrednio, przez stan 2+ (wzbudzenie dwukrotne). Odnosi się to tym bardziej do stanów o wyższym spinie: można je osiągnąć niemal jedynie drogą wielokrotnego („drabinkowego") wzbudzenia kulombowskiego.

reakcje wprost

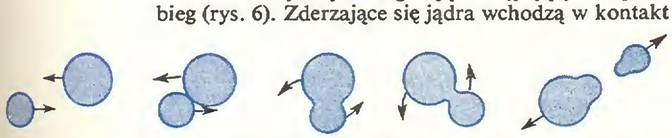
Reakcje wprost są procesami szybkimi. Czas ich trwania jest rzędu czasu swobodnego przelotu jonu przez obszar silnego (jądrowego) oddziaływania z jądrem tarczy, a więc ok. 10^{-22} s. Oddziaływanie następuje między kilkoma tylko nukleonami na powierzchni jąder; reszta nukleonów nie bierze udziału w zderzeniu, pozostają one tylko jego „obserwatorami". W konsekwencji przekaz energii i momentu pędu jest nieduży. Nieduża jest także liczba nukleonów, które zostają przekazane z jednego jądra do drugiego.

reakcje głęboko nieelastyczne

Reakcje głęboko nieelastyczne obejmują cały zakres między reakcjami wprost a reakcjami przez jądro złożone. Przekrój czynny, czyli prawdopodobieństwo takich reakcji, rośnie ze wzrostem energii zderzenia oraz ze wzrostem mas zderzających się jąder. Dość duże wartości przekroju czynnego otrzymuje się, gdy energia jonu przewyższa barierę kulombowską przynajmniej o 1-2 MeV/nukleon, a masy jąder są znaczne.

Czas trwania reakcji głęboko nieelastycznej mieści się w szerokim przedziale 10^{-22}–10^{-18} s, średnio jest więc rzędu 10^{-20} s. W tym czasie zderzające się jądra zdążą przejąć dużą część energii kinetycznej i momentu pędu ruchu względnego (dyssypacja energii i momentu pędu), co spowoduje silne ich wzbudzenie wewnętrzne.

Wyniki doświadczalnych badań nad reakcjami głęboko nieelastycznymi sugerują następujący ich przebieg (rys. 6). Zderzające się jądra wchodzą w kontakt



Rys. 6. Przebieg reakcji głęboko nieelastycznej (w układzie środka masy): od momentu nalatywania na siebie jąder, przez zlepienie się i wirowanie aż do rozerwania

ze sobą, „zlepiają się" na pewien czas (rzędu 10^{-20} s), tworząc wspólny układ. Połączenie to ma jednak charakter tylko dynamiczny. Czas jego trwania wystarcza z reguły do pełnej dyssypacji energii kinetycznej radialnego ruchu względnego, nie wystarcza jednak do pełnej dyssypacji względnego momentu pędu. W rezulatacie — po wykonaniu obrotu o pewien kąt siła odśrodkowa pochodząca od tej części względnego momentu pędu, która nie uległa dyssypacji, rozrywa zlepione jądro na dwie części. Podczas połączenia między zlepionymi jądrami może nastąpić wymiana znacznej liczby nukleonów.

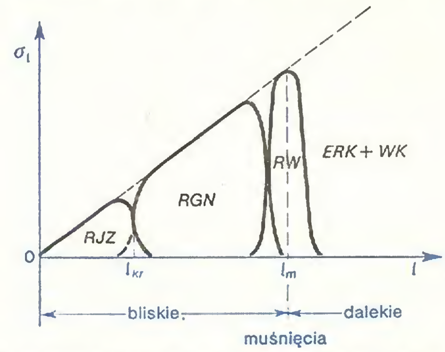
Z przedstawionego mechanizmu reakcji głęboko nieelastycznych widać, że badanie ich może być ważnym źródłem wiedzy o wielkości i naturze efektów dyssypacyjnych w jądrach.

reakcje przez jądro złożone

Przy odpowiednio małym parametrze zderzenia b wartość względnego momentu pędu l jest dostatecznie mała, by siła odpychająca (siła odśrodkowa plus siła kulombowska), działająca na jądra w czasie ich zetknięcia, była mniejsza od siły jądrowej przyciągającej je do siebie. Powstają wtedy warunki do utworzenia przez zderzające się jądra układu względnie trwałego. Przy utworzeniu jego zarówno energia kinetyczna jak i moment pędu ruchu względnego ulegają całkowitej dyssypacji i zostają rozłożone między wszystkie nukleony. Mówi się, że układ jest w równowadze (termodynamicznej) i nazywa się go jądrem złożonym. W jądrze tym zatarte zostają ślady sposobu, w jaki ono powstało. Ogólne charakteryzujące je wielkości, jak liczba masowa A, atomowa Z, energia wzbudzenia i spin nie są jednoznacznie związane z tym sposobem, można je otrzymać przy zderzeniu wielu różnych par jąder. Tylko te wielkości, a nie „zapomniany" sposób powstania, mają pewien, statystyczny w swoim charakterze wpływ na sposób rozpadu jądra złożonego. Może to być rozpad przez emisję lekkich cząstek, rozpad γ lub rozszczepienie. Czas tworzenia się jądra złożonego jest rzędu 10^{-17} s. W skali czasu procesów jądrowych jest to bardzo dużo, ok. stu tysięcy razy dłużej, niż trwa reakcja wprost.

Badanie warunków, w jakich może powstać jądro złożone, ważne jest m.in. dla poznania możliwości syntezy jąder bardzo ciężkich.

Granica między omówionymi rodzajami reakcji niskich energii nie jest ostra. Przy płynnej zmianie takich parametrów reakcji jak parametr zderzenia b czy względny moment pędu l, jeden rodzaj przechodzi pyłnnie w drugi. Tak np. na rys. 7 podany jest wy-



Rys. 7. Wykres prawdopodobieństwa σ_l zajścia procesu o danym względnym momencie pędu l, w zależności od l (por. rys. 4).

zależność σ od momentu pędu

kres przekroju czynnego, czyli wykres prawdopodobieństwa zajścia różnego rodzaju procesów w zależności od l. Przy dużym l zachodzi elastyczne rozpraszanie kulombowskie (*ERK*) i wzbudzenie kulombowskie (*WK*), przy mniejszym — reakcje wprost (*RW*), jeszcze mniejszym — reakcje głęboko nieelastyczne (*RGN*), a dopiero przy wartościach l mniejszych od pewnej wartości krytycznej l_{kr} tworzy się jądro złożone (*RJZ*). Przy przejściu od jednego obszaru do drugiego prawdopodobieństwo odpowiedniego procesu zmienia się stopniowo, płynnie (na rysunku linie ciągłe). Rysunek 7 przedstawia w inny trochę sposób to samo w zasadzie co poglądowy rys. 4. Moment pędu $l_m = pb_m$ (gdzie p — pęd padającego jonu) odpowiada zderzeniom, w których jądra zaledwie się muskają powierzchniami. Rozdziela on wszystkie możliwe zderzenia na dwie klasy: zderzeń bliskich i zderzeń odległych.

Energie pośrednie

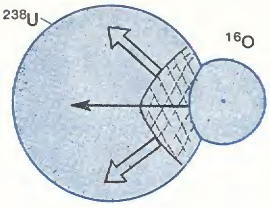
proces kruszenia

Przy wzroście energii jonu rośnie rola procesu bezpośredniego wybijania ze zderzających się jąder małych lub większych ich fragmentów (proces kruszenia lub fragmentacji).

Oczekuje się, że w pewnym zakresie energii, właściwym już dla obszaru procesów kruszenia, wystąpi bardzo charakterystyczny proces, inny niż omówione dotychczas. Mianowicie, przy energii większej od ok. 20 MeV/nukleon ciężki jon może wywołać w jądrze tarczy falę uderzeniową. Byłaby to fala zgęszczenia

fala uderzeniowa w materii jądrowej

materii jądrowej, analogiczna do fali uderzeniowej w cieczy lub gazie. Jeden z możliwych obrazów rozchodzenia się takiej fali w jądrze ^{238}U uderzonym



Rys. 8. Rozchodzenie się fali uderzeniowej w jądrze ^{238}U uderzonym jonem ^{16}O

jonem ^{16}O przedstawia rys. 8. Czoło fali, biegnące z prędkością większą od szybkości „dźwięku" w materii jądrowej jądra ^{238}U (określonej przez jej ściśliwość), oddziela „chłodną" część jądra od silnie „nagrzanej" (wzbudzonej). Z geometrii procesu widać, że emitowane z „gorącego" obszaru cząstki (nukleony, deuterony, ^{3}He, ^{4}He i cięższe) wylatywałyby głównie na boki, co by świadczyło o obecności fali.

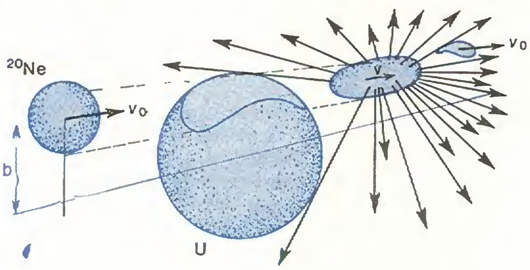
jądra supergęste

Zagadnienie występowania fali uderzeniowej w jądrach związane jest z problemem jąder supergęstych, przewidywanych przez teorię. Oczekuje się mianowicie, że oprócz znanych jąder o normalnej gęstości mogą występować jądra o gęstości znacznie większej (ok. dwóch razy). Wytworzenie fal uderzeniowych w materii jądrowej mogłoby być pomocne w otrzymaniu takich jąder. Dotychczas jednak nie stwierdzono istnienia tych fal i nie wiadomo, czy można je wytworzyć w jądrze ani też, czy można by za ich pomocą otrzymać jądra supergęste.

Energie wysokie (relatywistyczne)

Obecnie wytwarza się już wiązki ciężkich jonów (^{20}Ne) o energii w zakresie ok. 250–2000 MeV/nukleon. Taka energia kinetyczna jest zbliżona lub nawet znacznie większa od energii spoczynkowej jonu (wynoszącej ok. 940 MeV/nukleon), przeto do opisu ruchu należy stosować mechanikę relatywistyczną; stąd nazwa „energie relatywistyczne".

Przeprowadzone dotychczas badania reakcji jonu ^{20}Ne z jądrami tarczy uranowej sugerują następujący przebieg procesu (rys. 9). Padający jon wybija w jąd-



Rys. 9. Przebieg zderzenia jonu ^{20}Ne o relatywistycznej energii z jądrem uranu. Padający jon wybija w jądrze tarczy rów. Części jąder z obszaru zderzenia poruszają się po zderzeniu jako silnie wzbudzony, „ognisty" kawałek materii jądrowej, którego fragmenty rozlatują się na wszystkie strony. Pozostałe części jąder są „chłodnymi" obserwatorami zderzenia

wybijanie „ognistych" kawałków

rze uranu „rów" lub „tunel". Ta część jądra, która nie leżała na drodze jonu, pozostaje „chłodnym" obserwatorem zderzenia. Podobnie część jonu, która nie napotkała jądra tarczy, pozostaje „chłodna", mało wzbudzona i kontynuuje swój ruch z nie zmienioną prędkością początkową. Tylko ta część materii, która się znalazła dokładnie w obszarze zderzenia, biegnie z odpowiednio mniejszą prędkością jako „gorący", bardzo silnie wzbudzony, „ognisty" — jak się często mówi — kawałek pokruszonej materii jądrowej. Rozlatują się z niego na wszystkie strony nie powiązane już ze sobą jego fragmenty — głównie nukleony, lecz także deuterony, jądra ^{3}He, ^{4}He oraz jądra cięższe.

Z obrazu tego, zgodnego z wynikami doświadczenia, widać, jak bardzo różny jest ten proces od procesu wzbudzenia kulombowskiego, od reakcji wprost czy nawet reakcji głęboko nieelastycznych, które zachodzą przy energiach niskich i które są reakcjami dwuciałowymi albo dokładnie (wzbudzenie kulombowskie), albo w dobrym przybliżeniu (reakcje wprost i reakcje głęboko nieelastyczne). Dwa silnie wzbudzone w tych reakcjach jądra końcowe mogą się rozpadać dopiero po pewnym czasie od momentu zderzenia emitując na ogół niewiele lekkich cząstek lub kwantów γ, lub rozszczepiając się. W omówionej zaś reakcji wysokiej energii sam pierwotny proces jest od razu wielociałowy.

Zastosowanie w fizyce jądrowej

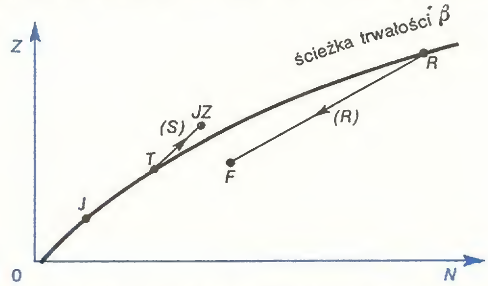
Jedno z pierwszych ważnych rodzajów zastosowania ciężkich jonów w fizyce jądrowej — wzbudzenie kulombowskie — omówione zostało w poprzednim paragrafie. Tutaj opiszemy kilka dalszych, które wraz z nim zdecydowały i nadal decydują o rozwoju fizyki ciężkich jonów.

Otrzymywanie i badanie jąder dalekich od ścieżki trwałości β

Reakcje z ciężkimi jonami są bardzo skuteczną metodą otrzymywania jąder dalekich od ścieżki trwałości β (→ Jądra atomowe i ich wzbudzenia). Jądra te otrzymywane są zarówno w reakcjach przez jądro złożone, jak i w procesie przekazu (transferu) nukleonów w reakcji wprost czy reakcji głęboko nieelastycznej. W obu tych procesach lekkie jony nie pozwalają odejść daleko od ścieżki trwałości i dopiero jony ciężkie dają taką możliwość. Na przykład w reakcjach przez jądro złożone z cząstkami α, $T(\alpha, xn)$, liczba emitowanych neutronów jest mała i możemy się oddalić od ścieżki tylko o parę jednostek masy. Zapis $T(\alpha, xn)$ oznacza tutaj, że jądro złożone utworzone jest przez naświetlanie tarczy T cząstkami α i, silnie wzbudzone, emituje następnie x neutronów n.

jądro złożone

Rozpatrzymy najpierw proces przez jądro złożone. W wytwarzaniu jąder dalekich od ścieżki trwałości wykorzystujemy fakt, że im cięższe jest jądro na ścieżce, tym większy ma stosunek liczby neutronów N do protonów Z. Tak np. w jądrach $^{16}_{8}$O, $^{44}_{20}$Ca i $^{208}_{82}$Pb stosunek ten wynosi odpowiednio $8/8 = 1{,}00$, $24/20 = 1{,}20$ i $126/82 \approx 1{,}54$. Oznacza to, że na mapie nuklidów, tzn. we współrzędnych (N, Z), ścieżka trwałości odchyla się od przekątnej coraz silniej w kierunku osi N. Fakt ten ilustruje (nieco przesadnie) rys. 10, z którego widać, że synteza jonu J z jądrem tarczy T zawsze prowadzi do jąder neutronodeficytowych, tzn. mających za mało neutronów w stosunku do odpowiadającego im izotopu trwałego. Deficyt neutronów jest tym większy, im cięższy jest



Rys. 10. Ilustracja faktu, że synteza (S) jonu J z jądrem tarczy T prowadzi zawsze do jąder neutronodeficytowych JZ, a rozszczepienie (R) jądra R — do jąder neutrononadmiarowych F

jon J i jądro tarczy T. Pogłębia go jeszcze fakt, że silnie wzbudzone jądro złożone JZ emituje kilka neutronów (a nie protonów, którym utrudnia taką emisję bariera kulombowska) przed przejściem do stanu podstawowego. W rezultacie, w procesie przez jądro złożone można osiągnąć nuklidy o deficycie kilkunastu (do ok. 20) neutronów, zbliżając się w ten sposób do linii zerowej energii wiązania protonu (→ Jądra atomowe w stanach ekstremalnych).

Z rys. 10 widać także, że w procesie rozszczepienia, który jest pod wieloma względami odwrotny do syntezy, rzecz przedstawia się odwrotnie. Dwa np. równe produkty (fragmenty) F rozszczepienia jądra R są neutrononadmiarowe. Nadmiar neutronów może wynosić kilkanaście, do ok. 20, podobnie jak niedomiar przy syntezie. Samorzutnie rozszczepiają się jedynie jądra ciężkie o $A \approx 240$, dając produkty głównie w obszarze mas $A \approx 100$ i $A \approx 140$. Pobudzone jednak ciężkimi jonami, mogą się także rozszczepiać jądra znacznie lżejsze, dając odpowiednio znacznie lżejsze produkty rozszczepienia F. Dzięki ciężkim jonom możemy więc wytwarzać jądra neutrononadmiarowe F niemal wzdłuż (poniżej) całej ścieżki trwałości β.

proces transferu

Za pomocą procesu transferu nukleonów osiągnięto również wiele nuklidów dalekich od ścieżki trwałości; przede wszystkim wśród lekkich jąder neutrononadmiarowych. Przykładem są reakcje:

$$^{232}_{90}\mathrm{Th}+^{15}_{7}\mathrm{N} \rightarrow {}^{236}_{94}\mathrm{Pu}+^{11}_{3}\mathrm{Li},$$

$$^{232}_{90}\mathrm{Th}+^{22}_{10}\mathrm{Ne} \rightarrow {}^{228}_{90}\mathrm{Th}+^{26}_{10}\mathrm{Ne},$$

$$^{232}_{90}\mathrm{Th}+^{22}_{10}\mathrm{Ne} \rightarrow {}^{230}_{92}\mathrm{U}+^{24}_{8}\mathrm{O}.$$

Pierwsza z nich — to reakcja zdarcia 4 protonów z jądra $^{15}_{7}\mathrm{N}$, druga — przechwytu 4 neutronów przez jądro $^{22}_{10}\mathrm{Ne}$, trzecia — zdarcia 2 protonów i przechwytu 4 neutronów przez to jądro.

W taki właśnie sposób osiągnięto granicę trwałości jąder względem emisji neutronu (czyli linię zerowej energii wiązania neutronu) dla kilku najlżejszych pierwiastków: H, He, Li, Be i B. Przyczyniło się to do poprawienia wzorów masowych dla lekkich jąder. Uprzednio wzory te przewidywały, że jądra $^{11}_{3}\mathrm{Li}$, $^{14}_{4}\mathrm{Be}$, $^{17}_{5}\mathrm{B}$ i $^{20}_{6}\mathrm{C}$ są nietrwałe ze względu na emisję neutronu. Po doświadczalnym otrzymaniu ich jednak w reakcjach transferu przekonano się, że tak nie jest.

Niełatwym technicznie problemem jest badanie własności jąder położonych z dala od ścieżki trwałości β. Jądra te mają bowiem bardzo krótkie czasy życia i pomiary należy wykonywać możliwie jednocześnie z ich wytwarzaniem. Rozwiązuje się to przez ustawienie seperatora izotopów oraz detektorów bezpośrednio lub prawie bezpośrednio „na wiązce" cząstek wywołujących reakcję.

Synteza ciężkich pierwiastków

Wszystkie pierwiastki transuranowe, tzn. pierwiastki leżące w tablicy Mendelejewa za uranem ($Z > 92$), otrzymano sztucznie. W naturze one nie występują, gdyż czas ich życia (czas połowicznego rozpadu) jest znacznie mniejszy od wieku Ziemi. Większość z nich otrzymano po raz pierwszy w reakcjach z jonami. Mianowicie pluton ($Z = 94$) otrzymano w reakcjach z deuteronami, a kiur ($Z = 96$), berkel ($Z = 97$), kaliforn ($Z = 98$) i mendelew ($Z = 101$) — w reakcjach z cząstkami α (tzn. jądrami $^{4}\mathrm{He}$). Dalszych pierwiastków nie można było już otrzymać za pomocą lekkich jonów, gdyż nie dysponowano i nadal się nie dysponuje dostatecznymi ilościami pierwiastków cięższych niż einstein do sporządzenia tarczy. Konieczne zatem było i jest stosowanie ciężkich jonów. Tak np. nobel otrzymano w reakcjach

synteza transuranowców

$$^{246}_{96}\mathrm{Cm}+^{12}_{6}\mathrm{C} \rightarrow {}^{254}_{102}\mathrm{No}+4\,^{1}_{0}n,$$

$$^{238}_{92}\mathrm{U}+^{22}_{10}\mathrm{Ne} \rightarrow {}^{254}_{102}\mathrm{No}+6\,^{1}_{0}n,$$

a pierwiastek o liczbie atomowej $Z = 106$, ostatni z otrzymanych dotychczas — w reakcjach

$$^{249}_{98}\mathrm{Cf}+^{18}_{8}\mathrm{O} \rightarrow {}^{263}_{106}\mathrm{X}+4\,^{1}_{0}n,$$

$$^{208}_{82}\mathrm{Pb}+^{54}_{24}\mathrm{Cr} \rightarrow {}^{259}_{106}\mathrm{X}+3\,^{1}_{0}n$$

(symbolem X oznaczyliśmy pierwiastki nie mające jeszcze uzgodnionej nazwy lub w ogóle jeszcze nie znane).

synteza pierwiastków superciężkich

Obecnie trwają próby nad syntezą pierwiastków 107 i 108 oraz tzw. pierwiastków superciężkich. Według przewidywań teoretycznych pierwiastki superciężkie powinny tworzyć wyspę nuklidów o podwyższonej trwałości wokół nuklidu $^{298}_{114}\mathrm{X}$, miałyby więc o ok. 40 nukleonów więcej niż najcięższe znane dotychczas jądro (tj. jądro $^{263}_{106}\mathrm{X}$; o czasie życia 0,9 s). A zatem i własnościami prawdopodobnie bardzo by się różniły od znanych jąder, wobec czego stanowią niezwykle interesujący obiekt badań.

Obecnie próbuje się otrzymać te jądra za pomocą wiązki jonów $^{48}\mathrm{Ca}$, np. w reakcji

$$^{248}_{96}\mathrm{Cm}+^{48}_{20}\mathrm{Ca} \rightarrow {}^{294}_{116}\mathrm{X}+2\,^{1}_{0}n.$$

Duża użyteczność jonów $^{48}\mathrm{Ca}$ polega głównie na dwóch ich własnościach. Pierwsza — to że są one bogate w neutrony, co pozwala dostać się za ich pomocą stosunkowo blisko środka wyspy jąder superciężkich, a więc do jąder najtrwalszych. Druga — to że są one szczególnie silnie związane (są to bowiem jądra podwójnie magiczne), co sprzyja stosunkowo niskiej energii wzbudzenia wytworzonego za ich pomocą jądra superciężkiego. Daje to większe szanse przejścia tego jądra do stanu podstawowego, w którym jest najtrwalsze, i uniknięcia tym samym szybkiego rozpadu ze stanu wzbudzonego.

Wzbudzanie stanów o wysokim spinie

Duża masa, pęd, a także spore rozmiary ciężkiego jonu sprawiają, że może on tworzyć z jądrem tarczy układ o bardzo dużym momencie pędu, nieosiągalnym w reakcjach z lżejszymi cząstkami. Na przykład jon $^{20}\mathrm{Ne}$ o energii 10 MeV/nukleon może z jądrem tarczy $^{197}\mathrm{Au}$ utworzyć układ o momencie pędu do ok. 110 jednostek $\hbar$. Dla porównania — jon deuteronu o tej samej energii na nukleon, z takim samym jądrem $^{197}\mathrm{Au}$ może utworzyć układ o spinie tylko do ok. $8\hbar$.

dyssypacja momentu pędu

Interesujące są tu dwa zagadnienia. Jedno — to wyjaśnienie, co się dzieje z dużym momentem pędu ruchu względnego, jaka jego część jest przekazywana ruchowi wewnętrznemu zderzających się jąder i jak szybko. Jest to więc zagadnienie dyssypacji momentu pędu ruchu względnego — zagadnienie mechanizmu przekazywania go wewnętrznym stopniom swobody jąder.

niszczenie korelacji par

Z omówionego wyżej faktu istnienia spinu krytycznego wynika, że przy bardzo dużym momencie pędu część momentu pędu, która nie zdąży ulec dyssypacji w czasie reakcji, może (wskutek sił odśrodkowych, które wywołuje) nie dopuścić do utworzenia się jądra złożonego. Drugie zagadnienie dotyczy własności samego jądra złożonego. Interesujący jest mianowicie wpływ wysokiego spinu na te własności i na strukturę jądra. Stwierdzono np., że już przy stosunkowo niskich spinach (12–$16\hbar$) zachodzi gwałtowna zmiana momentu bezwładności jąder zdeformowanych. Zjawisko to wiąże się z rozrywaniem par nukleonów przez szybki obrót jądra. Niszczenie korelacji par zwiększa znacznie moment bezwładności i spowalnia przejścia elektromagnetyczne między odpowiednimi stanami rotacyjnymi. Duży spin jądra ma także istotny wpływ na sposób jego rozpadu. Podwyższa np. prawdopodobieństwo emisji cząstki α w stosunku do prawdopodobieństwa emisji pojedynczego nukleonu, zwiększa także prawdopodobieństwo rozszczepienia jądra.

Inne zastosowania

Omówimy tu tylko kilka spośród wielu zastosowań ciężkich jonów poza fizyką jądrową. Będą to zastosowania ich w dziedzinie ciała stałego, techniki, biologii i medycyny. We wszystkich ważne jest oddziaływanie powłoki elektronowej jonu z materią.

Implantacja ciężkich jonów

Naświetlanie dowolnego materiału wiązką ciężkich jonów jest bardzo dogodnym sposobem wprowadzania („wbijania", implantacji) atomów dowolnego pierwiastka do materiału. Takie wprowadzanie innych, domieszkowych atomów pozwala zmienić w bardzo szerokim zakresie własności fizyczne materiału, szczególnie elektryczne i mechaniczne.

prostota i uniwersalność metody

Dogodność implantacji polega na jej prostocie i uniwersalności. W procesie tym nie odgrywają roli relacje chemiczne między pierwiastkami materiału naświetlanego i implantowanego. Możliwość bardzo dobrego ogniskowania wiązki ciężkich jonów (bardzo małe jej rozmiary poprzeczne, rzędu 1 mm, a niekiedy nawet 1 μm) oraz dokładnego doboru energii jonów (a więc dokładnego doboru głębokości ich implantacji) pozwala na precyzyjne sterowanie geometrią wprowadzanych domieszek. Dzięki temu można w naświetlanej próbce — nawet przy miniaturowych jej rozmiarach — wytwarzać dosyć złożone i jednocześnie precyzyjne struktury geometryczne o kompleksowych własnościach elektrofizycznych. Ponadto, implantację można przeprowadzać przy stosunkowo niskich temperaturach, a niepożądane uszkodzenia struktury naświetlanej próbki można usunąć przez podgrzanie do niezbyt wysokiej temperatury. Pozwala to, dzięki uniknięciu silnych podgrzewań, wpływać na własności próbki w sposób dobrze kontrolowany.

zastosowanie na skalę przemysłową

Jeśli chodzi o zastosowanie implantacji na skalę przemysłową, istotny jest fakt, że proces ten można całkowicie zautomatyzować i osiągać prawie dokładną powtarzalność własności wytwarzanych materiałów. Stosując implantację uzyskuje się istotne polepszenie własności tranzystorów i innych układów półprzewodnikowych. Np. naświetlanie jonami boru, fosforu, tantalu i in. pierwiastków znacznie poprawia własności krzemowych i germanowych detektorów półprzewodnikowych, stosowanych w fizyce jądrowej.

wytwarzanie nowych materiałów

Implantacja pozwala także na wytworzenie stopów o niezwykłych własnościach mechanicznych i termicznych, trudnych lub niemożliwych do wytworzenia innymi metodami. Wymagane własności stopu otrzymuje się najczęściej już przy bardzo nieznacznych dodatkach odpowiednio dobranych domieszek. Zachęcające są wstępne wyniki stosowania implantacji w badaniach nad zagadnieniem wytwarzania nadprzewodników wysokotemperaturowych. Interesująca jest też możliwość zastosowania tej metody do wytwarzania światłowodów w układach optyki zintegrowanej. Światłowód może być wykorzystywany do szybkiego przekazywania dużej ilości informacji.

Modelowanie uszkodzeń radiacyjnych w materiałach reaktorowych

Szybkie neutrony emitowane przy rozszczepieniu jądrowym w reaktorze powodują uszkodzenia radiacyjne w materiałach konstrukcyjnych reaktora. Wybijają one mianowicie jądra z zajmowanych położeń, wprowadzając zmiany w strukturze materiału. Zmiany takie należy uwzględnić przy projektowaniu reaktorów.

Ponieważ ciężki jon oddziałuje z całym atomem, a nie tylko — jak neutron — z jądrem, jest on setki tysięcy lub nawet miliony razy efektywniejszy w wywoływaniu uszkodzeń radiacyjnych. Uszkodzenia powstające w materiałach reaktora dużej mocy w ciągu kilku lat jego pracy można odtworzyć w ciągu kilku godzin przez naświetlanie próbki materiału wiązką ciężkich jonów. Próbka staje się przy tym mniej promieniotwórcza, co upraszcza badanie. Także łatwiejsza i dokładniejsza jest kontrola dawki promieniowania, temperatury i innych parametrów.

Filtry jądrowe

Ciężkie jony mogą być wykorzystane jako „mikroigły przekłuwające" cienką folię z materiału plastycznego, szkła czy miki. „Przekłucie" jest po prostu uszkodzeniem radiacyjnym, które po odpowiedniej obróbce chemicznej staje się wąskim kanalikiem, a cała naświetlana folia — subtelnym i precyzyjnym filtrem (tzw. filtr jądrowy). Średnica kanaliku zależy od rodzaju i energii jonu, materiału folii oraz od warunków obróbki chemicznej. Obecnie uzyskuje się filtry jądrowe o średnicy porów od 4 nm do kilkudziesięciu μm (tj. rzędu 10^4 nm). Rysunek 11 pozwala porównać filtr chemiczny wysokiej jakości z filtrem jądrowym. Widać, że pierwszy ma otwory o różnych średnicach i nieregularnych kształtach, wskutek czego rozmiary przepuszczanych cząstek mają duży rozrzut. Filtr jądrowy natomiast ma otwory o jednakowej niemal średnicy i o regularnym, kołowym kształcie. Ponadto istnieje możliwość prawie ciągłej zmiany średnicy w bardzo szerokich, jak podano wyżej, granicach.

filtr chemiczny

filtr jądrowy



Rys. 11. Zdjęcie zwykłego filtru chemicznego o średniej efektywnej średnicy pora 0,45 μm (rys. a) oraz filtru jądrowego o średnicy pora 0,40 μm (rys. b), uzyskane za pomocą mikroskopu elektronowego

Filtry jądrowe mogą być wykorzystywane np. do zimnej sterylizacji. Ponieważ bakterie mają rozmiary powyżej 0,5 μm, można je łatwo zatrzymać już za pomocą filtrów o średnicy kanalików 0,45 μm.

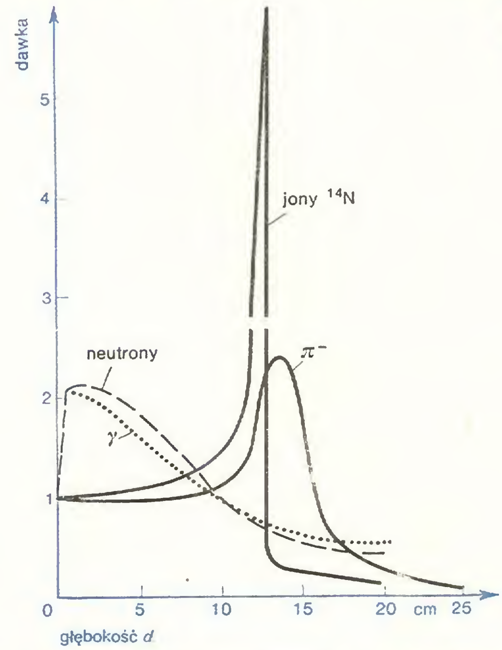
Zastosowanie w medycynie

radioterapia

Podstawowym zastosowaniem wiązki ciężkich jonów w medycynie może być radioterapia, przede wszystkim niszczenie komórek nowotworowych. Wielką zaletą tej wiązki jest to, że pozwala ona na bardzo dobrą lokalizację naświetlanego obszaru zarówno w kierunku poprzecznym (możliwość bardzo małych rozmiarów wiązki), jak i podłużnym (możliwość naświetlania tkanki niemal wyłącznie na żądanej głębokości). Można więc uniknąć niszczenia sąsiedniej, zdrowej tkanki w czasie naświetlania. Ani stosowane najczęściej w praktyce promieniowanie X i γ, ani strumienie elektronów i neutronów nie dają możliwości dobrej lokalizacji podłużnej. Ilustruje to rys. 12, na którym porównane są rozkłady dawki (tj. wydzielonej energii) promieniowania w wodzie w funkcji głębokości d przy różnych rodzajach naświetlania: promieniami γ z ^{60}Co, neutronami o energii 14,6 MeV, mezonami π^- o energii 65 MeV i jonami ^{14}N o energii 278 MeV/ /nukleon. Widać, że promieniowanie γ i neutronowe nie dają w ogóle możliwości lokalizacji podłużnej, wiązka mezonów π^- (i podobnie protonów) daje już niezłą lokalizację, ale dopiero wiązka ciężkich jonów pozwala na lokalizację bardzo dobrą.

J. Bartke *Eksperymenty z jądrami cięższymi od wodoru przyspieszonymi do bardzo wysokich energii*, Post. Fiz. 24, 423 (1973); R. Broda, J. Wilczyński *Perspektywy zastosowania ciężkich jonów w fizyce jądrowej*, Raport no. 922/PL Instytutu Fizyki Jądrowej, Kraków 1976; *Cyklotron ciężkich jonów U-200 w Instytucie Fizyki Doświadczalnej Uniwersytetu Warszawskiego* (sprawozdanie z sympozjum), Jabłonna 1972; G.N. Flerow, W.S. Baraszenkow *Zastosowanie wiązek ciężkich jonów*, Post. Fiz. 27, 53 (1976); J. Grabowski *Aktualny stan teorii reakcji jądrowych z ciężkimi jonami*, Post. Fiz. 19, 257 (1968); P. Marmier and E. Sheldon *Physics of nuclei and particles*, v. 2, New York 1970; W. Rosiński *Implantacja jonów*, Warszawa 1975; E. Ruchowska *Synteza pierwiastka o Z = 106*, Post. Fiz., 26, 355 (1975); A. Sobiczewski, J. Żylicz *Oddziaływanie ciężkich jonów z jądrami i synteza nowych pierwiastków*, Post. Fiz., 27, 171 (1976); A. Sobiczewski, Z. Sujkowski *Europejska konferencja fizyki jądrowej uprawianej z pomocą ciężkich jonów w Caen*, Post. Fiz., 28, 189 (1977); J. Wilczyński *Makroskopowe aspekty mechanizmu zderzeń ciężkich jonów z jądrami atomowymi*, Raport no. 900/PL, Inst. Fizyki Jądrowej, Kraków 1975.

rozkład dawki w wodzie



Rys. 12. Zależność dawki promieniowania w wodzie od głębokości d, na której jest ona pochłonięta, przy różnych rodzajach promieniowania

Spektroskopia jądrowa

Andrzej Hrynkiewicz

Wszystkie dane eksperymentalne dotyczące własności poziomów jądrowych oraz charakteru oddziaływań jądrowych są podstawą, na której opierają się koncepcje struktury jąder atomowych. Dane te można podzielić na dwie kategorie: 1) parametry statyczne stanów jądrowych, takie jak energie wzbudzenia, spiny i izospiny, parzystości i momenty elektromagnetyczne oraz masy i rozmiary jąder, 2) wielkości charakteryzujące dynamikę procesów jądrowych, jak czasy życia i typy rozpadu oraz przekroje czynne na oddziaływanie jąder z cząstkami naładowanymi, neutronami i promieniowaniem elektromagnetycznym.

dane statyczne i dynamiczne

Każda metoda umożliwiająca wyznaczanie wielkości charakteryzujących jądra atomowe jest więc metodą badania ich struktury. Tutaj omówimy tylko metody spektroskopii jądrowej, tj. metody, które pozwalają wyznaczyć własności jąder promieniotwórczych na podstawie emitowanego promieniowania.

W ciągu wielu lat przedmiotem badań spektroskopii jądrowej były długożyjące nuklidy promieniotwórcze naturalne lub wytwarzane sztucznie. Długi czas życia nuklidów pozwalał transportować te źródła promieniotwórcze z miejsca ich wytworzenia do odległych nieraz laboratoriów wyposażonych w urządzenia spektroskopowe.

Obecnie większość badań spektroskopowych wykonuje się przy użyciu aparatury pomiarowej ustawionej przy akceleratorach lub reaktorach jądrowych (→ Akceleratory cząstek naładowanych, Energia jądrowa), co pozwala dokonywać pomiary bezpośrednio „na wiązce", tj. badać krótkożyjące produkty reakcji jądrowych. Obiektami badanymi są w tym wypadku jądra, których rozpad następuje przeważnie w czasie krótszym niż 1 ns (10^{-9} s).

pomiary „na wiązce"

Spektroskopia jądrowa „na wiązce" rozszerzyła również w sensie jakościowym zakres badanych stanów jądrowych. Można badać stany o bardzo wysokiej energii wzbudzenia (niedostępne w rozpadach promieniotwórczych) oraz jądra mające znaczny nadmiar lub niedobór neutronów (w stosunku do nuklidów występujących w przyrodzie). W pewnych eksperymentach wykorzystywane są prędkości odrzutu produktów reakcji oraz orientacja ich spinów względem kierunku lotu bombardujących cząstek.

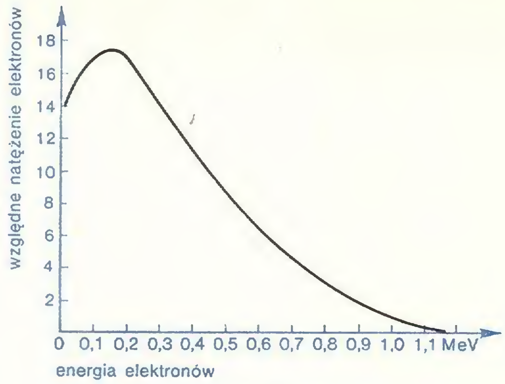
W pracach spektroskopowych „na wiązce" szczególne znaczenie ma również detekcja i spektrometria cząstek naładowanych i neutronów emitowanych w procesie reakcji jądrowej prowadzącej do wytworzenia interesującego nas nuklidu.

Spektroskopia elektronów

Jądra atomowe emitują elektrony ujemne lub dodatnie w rozpadach β^- i β^+ (→ Rozpady jąder atomowych). Niezależnie od tego elektrony są emitowane z powłoki elektronowej atomu albo w wyniku bezpośredniego przekazania im energii wzbudzenia jądra, czyli w procesie konwersji wewnętrznej (elektrony konwersji) lub też jako tzw. elektrony Augera, których emisja towarzyszy procesom wytwarzającym dziury w powłoce elektronowej.

widma energetyczne elektronów

Podstawowym zadaniem spektroskopii elektronowej jest wyznaczanie widm energetycznych emitowanych elektronów, tj. liczby elektronów w funkcji ich

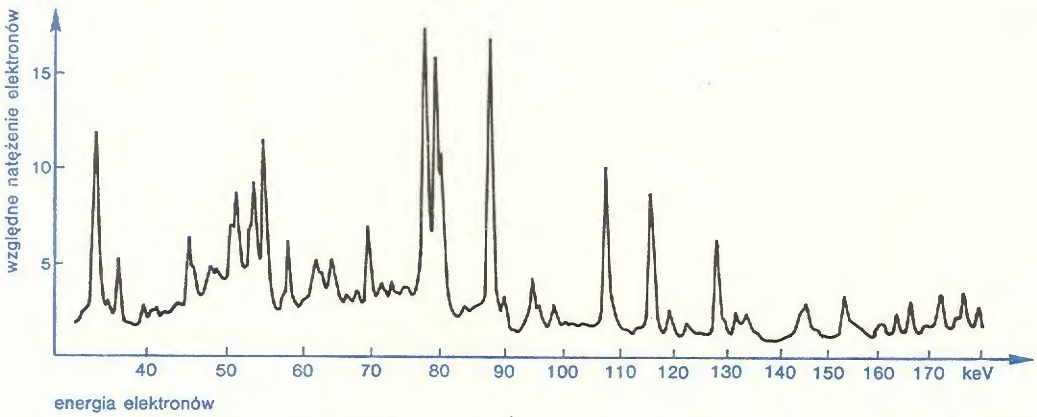


Na rys. 2 pokazany jest przekrój nowoczesnego spektrometru β — urządzenia pozwalającego rejestrować widma energetyczne elektronów. Odpowiednie pole magnetyczne ogniskuje na detektorze elektrony wylatujące ze źródła (którym może być tarcza bombardowana przez wiązkę cząstek z akceleratora). Zależnie od natężenia pola magnetycznego do detektora trafiają elektrony o różnej energii. Zmieniając natężenie pola magnetycznego (przez zmianę natężenia prądu zasilającego elektromagnes) możemy wyznaczyć liczbę elektronów o różnych określonych wartościach energii czyli, jak mówimy, „zdjąć" ich widmo energetyczne

spektrometr β

Typy spektrometrów β różnią się między sobą transmisją i energetyczną zdolnością rozdzielczą. Transmisją nazywamy stosunek liczby elektronów o danej energii docierających do detektora do liczby elektronów o tej energii emitowanych przez źródło. W przypadku spektrometrów toroidalnych transmisja może przekroczyć nawet 20%. Zdolność rozdzielcza dana jest przez szerokość (w połowie wysokości)

transmisja i zdolność rozdzielcza



Rys. 1. Przykłady widm energetycznych elektronów. a) Widmo ciągłe elektronów rozpadu β^- RaE, b) Widmo liniowe elektronów konwersji wewnętrznej izotopów złota wytworzonych w reakcji Pt(p, xn)Au

widmo rozpadu β

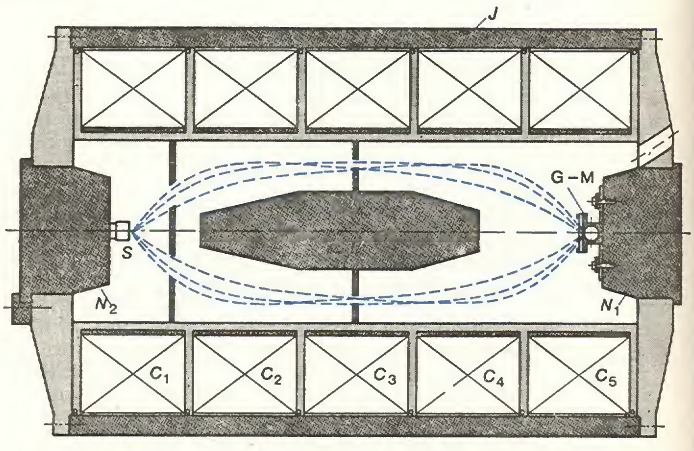
energii. Na rys. 1 pokazane są przykłady widma ciągłego elektronów rozpadu β i widma liniowego elektronów konwersji wewnętrznej. W pierwszym przypadku widmo elektronów jest ciągłe, gdyż w rozpadzie β emisji elektronu towarzyszy emisja neutrina, które nie jest rejestrowane, a które unosi część energii rozpadu. Kraniec widma elektronów β określa energię rozpadu, a z kształtu widma można wyciągnąć wnioski dotyczące typu rozpadu, np. określić tzw. stopień wzbronienia rozpadu β i uzyskać informację o zmianie spinu i parzystości jądra ulegającego rozpadowi.

widmo elektronów konwersji

Widmo elektronów konwersji (rys. 1b) wygląda zupełnie inaczej — jest to widmo liniowe. Poszczególne „piki" w widmie odpowiadają dyskretnym wartościom energii emitowanych elektronów. Energia elektronu konwersji jest dana przez wzór

$$E = E_0 - E_w,$$

gdzie E_0 — energia przejścia jądra z wyższego na niższy poziom energetyczny, E_w — energia wiązania elektronu na danej powłoce w atomie. Pomiar energii elektronów konwersji pozwala znaleźć energię przejścia jądrowego E_0, a z natężenia linii konwersji (z liczby elektronów w danym piku) można wnioskować o konkurencji między konwersją wewnętrzną a emisją fotonu promieniowania γ, czyli wyznaczyć tzw. współczynnik konwersji wewnętrznej. Jego wartość jak również stosunki natężeń elektronów konwersji pochodzących z różnych powłok elektronowych atomu są źródłem informacji o charakterze danego przejścia jądrowego E_0, co z kolei także dostarcza danych o spinach i parzystości stanów jądrowych.



Rys. 2. Schemat spektrometru β z podłużnym polem magnetycznym. Oznaczenia: C_1–C_5 cewki zasilające elektromagnes, J, N_1 i N_2 żelazne jarzmo i nabiegunniki, L osłona ołowiana, S źródło promieniotwórcze emitujące elektrony, G–M licznik elektronów Geigera–Müllera. Liniami przerywanymi niebieskimi oznaczone są tory elektronów poruszających się po spiralach wokół osi spektrometru

piku obserwowanego w widmie, odpowiadającego monoenergetycznym elektronom emitowanym przez źródło. Im lepsza jest zdolność rozdzielcza spektrometru (tzn. im węższe są piki w widmie), tym łatwiej możemy rozróżnić dwie linie leżące blisko siebie. Zdolność rozdzielcza najbardziej precyzyjnych spektrometrów osiąga 0,01% mierzonej energii elektronów. Na ogół spektrometry o dużej energetycznej zdol-

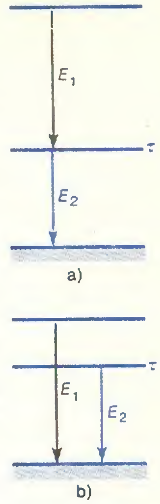
ności rozdzielczej mają małą transmisję i odwrotnie — dużej transmisji towarzyszy niska zdolność rozdzielcza, toteż typ przyrządu pomiarowego powinien być dobierany odpowiednio do problemu, który należy rozwiązać. Obecnie w wielu eksperymentach spektrometry magnetyczne są zastępowane półprzewodnikowymi detektorami elektronów, które szczególnie w obszarze niskich energii elektronów mają zdolność rozdzielczą porównywalną ze zdolnością rozdzielczą spektrometrów magnetycznych.

Spektroskopia promieniowania γ

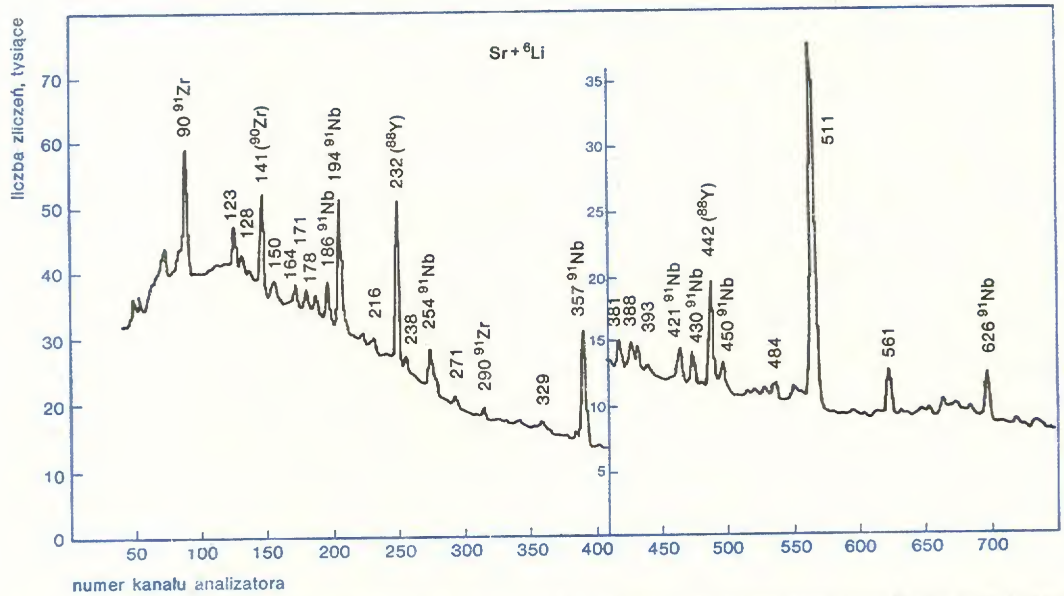
Widmo promieniowania γ emitowanego przez jądra atomowe jest widmem liniowym (rys. 3), gdyż przejściu jądra ze stanu wzbudzonego na niższy odpowiada emisja fotonów o określonej energii. Liniowe jest również widmo promieniowania rentgenowskiego, którego emisja towarzyszy takim procesom jądrowym, jak wychwyt elektronu lub konwersja wewnętrzna.

Metoda koincydencji w spektroskopii jądrowej

Przy rozwiązywaniu wielu problemów spektroskopii jądrowej istotne znaczenie ma informacja o czasowych zależnościach emisji przez jądro różnego rodzaju promieniowania. Dobrym tego przykładem może być problem wyboru między dwoma wariantami rozpadu jądra przedstawionymi schematycznie na rys. 4. W obu wypadkach w widmie energetycznym zaobserwujemy piki odpowiadające fotonom γ o energiach E_1 i E_2 i dopiero stwierdzenie, czy emisja fotonów E_1 i E_2 jest skorelowana w czasie (schemat 4a), czy też nie (schemat 4b), umożliwia rozstrzygnięcie, według którego schematu przebiegał rozpad. Aby stwierdzić, czy emisji fotonu E_1 towarzyszy (w granicach krótkiego przedziału czasowego) emisja fotonu E_2, należy wykonać pomiar koincydencyjny. Stosujemy w tym celu dwa detektory promieniowania γ, z których jeden rejestruje fotony o energii E_1, a drugi — fotony o energii E_2, następnie impulsy z obu detektorów kierujemy do



Rys. 4. Ilustracja zastosowania metody koincydencji; dwa warianty schematu rozpadu: a) emisja dwóch fotonów γ w kaskadzie, b) emisja dwóch nieskorelowanych fotonów γ



Rys. 3. Przykład widma energetycznego promieniowania γ emitowanego z tarczy strontu bombardowanej jonami ^{6}Li. Liczby nad pikami podają energię przejść γ w keV

Jednak pod pikami w widmie obserwuje się ciągłe tło, którego głównym źródłem jest rozproszenie promieniowania na elektronach (przede wszystkim w zjawisku Comptona) zachodzące w detektorze lub w otaczających go materiałach.

W spektrometrach γ stosowane są detektory rejestrujące poszczególne fotony i wytwarzające impulsy elektryczne, których amplituda jest proporcjonalna do energii pochłoniętego fotonu. Widmo energetyczne promieniowania γ, takie jak pokazane przykładowo na rys. 3, otrzymuje się za pomocą analizatora wielokanałowego, który sortuje impulsy zależnie od ich amplitudy.

detektory promieniowania γ

Najpowszechniej stosowanymi detektorami promieniowania γ są detektory półprzewodnikowe, przede wszystkim germanowe, które w porównaniu ze stosowanymi poprzednio licznikami scyntylacyjnymi mają znacznie lepszą (ponad 20-krotnie) energetyczną zdolność rozdzielczą. Szerokość piku uzyskiwanego za pomocą dobrego, dużego detektora germanowego wynosi 2–3 keV, co daje w obszarze mierzonych energii kilku MeV względną zdolność rozdzielczą rzędu 0,1%. Wydajność dużych detektorów germanowo-litowych (zależna od rozmiarów kryształu germanu) jest porównywalna z wydajnością liczników scyntylacyjnych.

układu elektronicznego, zwanego układem koincydencyjnym. Zadaniem tego układu jest wysłanie impulsu wyjściowego tylko wówczas, gdy impulsy z detektorów pojawiają się w momentach różniących się mniej w czasie niż tzw. czas rozdzielczy układu koincydencyjnego (może on wynosić np. 1 μs). W omawianym przykładzie pojawienie się wyjściowych impulsów koincydencji świadczy, że mamy do czynienia ze schematem rozpadu przedstawionym na rys. 4a.

układ koincydencyjny

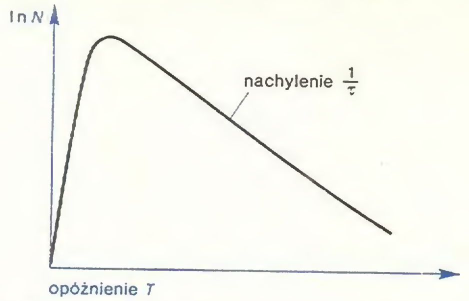
Stosując układ koincydencyjny o dostatecznie krótkim czasie rozdzielczym możemy wykonać pomiar średniego czasu życia τ stanu pośredniego w rozpadzie 4a. W tym celu, między detektorem rejestrującym fotony E_2 i układem koincydencyjnym umieszczamy linię opóźniającą, która pozwala opóźniać dojście impulsu o określony odstęp czasu. Mierząc szybkość zliczeń koincydencji N w zależności od wprowadzonego opóźnienia T otrzymamy krzywą pokazaną na rys. 5, której nachylenie w skali logarytmicznej jest związane ze średnim czasem życia stanu jądrowego τ wzorem

metoda opóźnionych koincydencji

$$\frac{d(\ln N)}{dT} = -\frac{1}{\tau}.$$

Sporządzając więc wykres liczby zliczeń w skali logarytmicznej możemy bezpośrednio wyznaczyć czas

życia τ wzbudzonego stanu jądrowego. Przedstawiona metoda opóźnionych koincydencji, przy zastosowaniu odpowiednio szybko reagujących detektorów i no-



Rys. 5. Przykład krzywej otrzymywanej w pomiarze średniego czasu życia stanu jądrowego metodą opóźnionych koincydencji

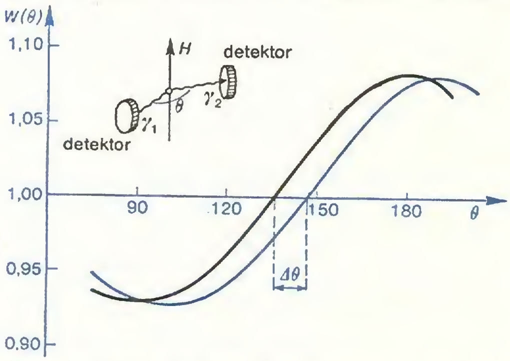
pomiary czasów życia

woczesnych układów koincydencyjnych, pozwala dokonywać pomiarów czasów życia krótszych nawet niż 10^{-9} s. W pomiarach, które wymagają dużej czasowej zdolności rozdzielczej, stosowane są nadal liczniki scyntylacyjne, gdyż wytwarzane w nich impulsy są znacznie krótsze od impulsów z detektorów półprzewodnikowych.

Metoda korelacji kierunkowych promieniowania γ

wyznaczanie korelacji kierunkowych

Pomiar koincydencji między fotonami E_1 i E_2 (rys. 4a) może być rozszerzony na pomiar kierunkowej zależności emisji obu fotonów kaskady. Możemy ją wyznaczyć zmieniając kąt, jaki tworzą kierunki od źródła promieniotwórczego do detektorów, tzn. obracając jeden z detektorów wokół źródła i wyzna-



Rys. 6. Korelacja kierunkowa promieniowania γ emitowanego w kaskadzie

czając dla każdego kąta szybkość zliczeń koincydencji (E_1, E_2) (rys. 6). Prawdopodobieństwo tego, że dwa fotony kaskady wylecą w kierunkach tworzących kąt θ, czyli funkcja korelacji kierunkowej $W(\theta)$ może być zapisana w postaci:

$$W(\theta) = 1+b_2\cos 2\theta+b_4\cos 4\theta,$$

przy czym współczynniki b_2 i b_4 zależą od spinów poziomów, między którymi zachodzi przejście, oraz od rodzaju (elektryczne czy magnetyczne) i multipolowości przejścia. Pomiar współczynników b_2 i b_4 pozwala więc uzyskać informacje o tych parametrach jądrowych. Typowy wykres funkcji korelacji kierunkowej dla dwóch przejść elektrycznych kwadrupolowych między poziomami o spinach kolejno 4–2–0 jest przedstawiony na rys. 6 (linia czarna).

Metoda korelacji kierunkowych może być wykorzystana do wyznaczania momentów elektromagnetycznych wzbudzonych stanów jądrowych. Przypuśćmy, że źródło promieniotwórcze znajduje się w polu magnetycznym skierowanym prostopadle do płaszczyzny, w której umieszczone są detektory, a jądro we wzbudzonym stanie pośrednim o czasie życia τ ma moment magnetyczny μ. Wskutek oddziaływania momentu magnetycznego jądra z polem magnetycznym wystąpi precesja spinu jądra wokół kierunku pola, co spowoduje obrót jądra w czasie życia τ o kąt wynoszący średnio $\Delta\theta$. Wywoła to przesunięcie wykresu funkcji korelacji kierunkowej (linia niebieska na rys. 6). Wyznaczenie tego przesunięcia pozwala obliczyć wartość oddziaływania jądra z polem magnetycznym, a stąd moment magnetyczny stanu jądrowego. W wypadku, gdy w funkcji korelacji można pominąć wyraz zawierający b_4, krzywa przesunięta jest opisana wzorem

wyznaczanie momentu magnetycznego

$$W(\theta, H) = 1+\frac{b_2}{\sqrt{1+(2\omega\tau)^2}}\cos 2(\theta-\Delta\theta),$$

gdzie $\Delta\theta = {}^1/_2 \operatorname{arctg}(2\omega\tau)$, a ω jest częstością precesji momentu magnetycznego jądra μ w polu magnetycznym o natężeniu H:

$$\omega = \mu H/I\hbar.$$

Jak widać pomiar przesunięcia kątowego $\Delta\theta$ pozwala wyznaczyć moment magnetyczny μ pod warunkiem, że znane jest natężenie pola magnetycznego H, średni czas życia τ danego stanu jądrowego i jego spin I.

Metoda pomiaru czasu życia stanów jądrowych oparta na efekcie Dopplera

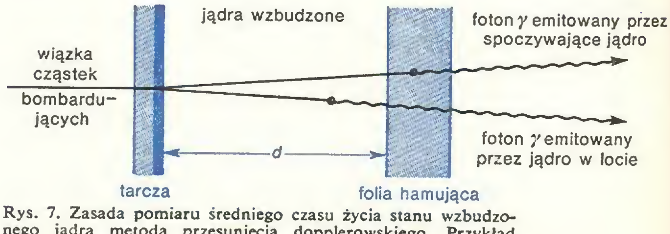
Metodą opóźnionych koincydencji, o której była mowa wyżej, nie można zmierzyć czasów życia stanów jądrowych żyjących krócej niż 10^{-9}–10^{-10} s. W obszarze krótszych czasów życia może być wykorzystana metoda oparta na zjawisku Dopplera, która jest z pewnością dobrze znana Czytelnikowi z dziedziny akustyki. Również w optyce częstość odbieranej fali elektromagnetycznej zależy od prędkości ruchu źródła względem obserwatora. Zmiana rejestrowanej przez obserwatora częstości $\Delta\nu$ dana jest wyrażeniem

$$\Delta\nu = \pm\nu_0 v/c,$$

gdzie ν_0 — częstość fali emitowanej przez źródło spoczywające względem obserwatora, v — prędkość ruchu względnego, c — prędkość światła. Zmiana częstości jest dodatnia, gdy źródło i obserwator zbliżają się do siebie, a ujemna — gdy się oddalają. Ponieważ energia fotonu jest związana z częstością drgań: $E = h\nu$, to zmianie częstości odpowiada zmiana energii rejestrowanego fotonu

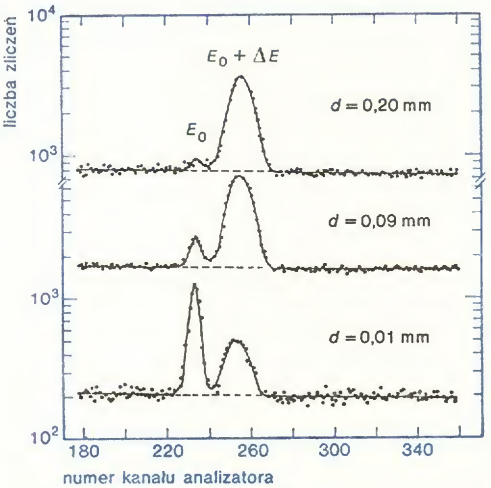
$$\Delta E = \pm E_0 v/c.$$

Opisany efekt może być wykorzystany do pomiaru czasu życia jądra, które w wyniku reakcji jądrowej doznało odrzutu i porusza się z prędkością wynoszącą na ogół kilka procent prędkości światła. Rysunek 7 przedstawia zasadę takiego pomiaru. Jądra wytworzone w cienkiej tarczy wylatują z niej do próżni i po przebyciu drogi d są zatrzymane w folii hamującej. Detektor rejestruje fotony γ emitowane



Rys. 7. Zasada pomiaru średniego czasu życia stanu wzbudzonego jądra metodą przesunięcia dopplerowskiego. Przykład przejścia 0,891 MeV w ^{22}Na

przez jądra wytworzone w stanie wzbudzonym. Jeżeli czas przelotu odcinka d jest znacznie dłuższy od średniego czasu życia danego stanu jądrowego, to niemal wszystkie fotony są wysyłane przez poruszające się jądra i mają wobec tego energię wyższą od E_0 o ΔE. Jeżeli natomiast czas przelotu jest znacznie krótszy od czasu życia τ, to większość fotonów wysyłana jest już przez jądra spoczywające, zahamowane w folii, a wobec tego ma energię E_0. Zmieniając odstęp d między tarczą a folią hamującą obserwujemy w widmie energetycznym γ zmianę natężenia dwóch pików przesuniętych względem siebie o ΔE. Pierwszy z nich odpowiada fotonom γ emitowanym przez jądra zaha-



Rys. 8. Przykład widm promieniowania γ dla różnych odległości d między tarczą a folią hamującą

mowane w folii (natężenie I_0), drugi — fotonom emitowanym przez jądra w locie (natężenie I_v) (rys. 8). Pomiar stosunku natężeń obu pików I_v/I_0 w zależności od odległości d pozwala obliczyć średni czas życia stanu jądrowego τ ze wzoru

wyznaczanie czasu życia

$$\frac{I_v}{I_0} = e^{d/v\tau} - 1.$$

Metoda ta umożliwia pomiary czasu życia rzędu 10^{-11}, a nawet 10^{-12} s.

Inne metody spektroskopii jądrowej

W tak krótkim opisie nie sposób przedstawić całego bogactwa metod spektroskopii jądrowej. Oprócz pomiarów energii, natężeń i rozkładów kątowych promieniowania jądrowego oraz czasów życia wzbudzonych stanów jądrowych można wymienić pomiary liniowej i kołowej polaryzacji promieniowania γ, a także pomiary podłużnej polaryzacji elektronów rozpadu β, tj. stopnia orientacji ich spinów w stosunku do kierunku lotu. Na wzmiankę zasługuje również wykorzystanie zalet detekcji promieniowania jądrowego w pomiarach prowadzonych innymi metodami, jak np. metodą magnetycznego rezonansu jądrowego lub metodą rezonansu w strumieniu atomowym. Zastosowanie w tych metodach detekcji promieniowania jądrowego umożliwia rejestrację pojedynczych jąder promieniotwórczych, a więc pracę z bardzo małymi ilościami badanych substancji. Korzystne okazało się również połączenie metod spektroskopii jądrowej z metodą orientacji jąder w bardzo niskich temperaturach.

Należy także zwrócić uwagę, że metody spektroskopii jądrowej znajdują dziś szerokie zastosowanie w badaniach własności materii skondensowanej (tzn. ciał stałych i cieczy), przy czym jądro atomowe w kryształach lub w cząsteczkach związków chemicznych odgrywa rolę mikroskopowego próbnika czułego na własności bezpośredniego otoczenia.

K. Siegbahn *α-, β-, γ-Ray Spectroscopy*, Amsterdam 1966; A. Strzałkowski *Wstęp do fizyki jądra atomowego*, Warszawa 1978; Sz. Szczeniowski *Fizyka doświadczalna*, cz. 6, Warszawa 1974.

Fizyka jądrowa wielkich energii

Przemysław Zieliński

Fizyka jądrowa wielkich energii jest dziedziną z pogranicza fizyki cząstek elementarnych i fizyki jądrowej. Zakres wielkich energii nie jest sprecyzowany, ale orientacyjnie przyjmuje się, że chodzi o oddziaływania z jądrami cząstek mających energię kinetyczną powyżej 1 GeV (1 gigaelektronowolt = $1 \cdot 10^9$ elektronowoltów). Jest to energia, która przekazana np. jądru o liczbie masowej A ok. 100, spowodowałaby, że wszystkie nukleony oddaliłyby się od siebie (odpowiada to całkowitej energii wiązania tego jądra). Pierwotnie w zakres fizyki jądrowej wchodziła również fizyka cząstek elementarnych, która w ostatnich dziesięcioleciach rozwinęła się w dyscyplinę samodzielną.

Rodzaje oddziaływań

Wymieniony powyżej przykład oddziaływania jest bardzo rzadki. Na ogół oddziaływanie powoduje rozerwanie tylko części wiązań nukleonów w jądrze. Występują rozmaite rodzaje oddziaływań. Najbardziej charakterystyczne z nich ilustruje schematycznie rys. 1.

rozpraszanie elastyczne

Rozpraszanie elastyczne (rys. 1a) jest procesem wyjątkowo prostym. Obie cząstki, zarówno cząstka padająca, jak i jądro-tarcza, pozostają w stanie podstawowym. Proces ten bardzo dobrze opisuje teoria wielokrotnego rozpraszania Glaubera, którą rozwinęli m.in. fizycy polscy, a szczególnie prof. Wiesław Czyż. Jeżeli cząstka padająca nieznacznie zmieni swój pęd (przypadek najczęstszy), to przeważa oddziaływanie z pojedynczym nukleonem, przy większych zmianach pędu należy uwzględnić kolejne rozpraszanie na dwóch, trzech itd. nukleonach (stąd nazwa teorii). Rozkład prawdopodobieństwa rozproszenia cząstki pod różnymi kątami przypomina obraz dyfrakcyjny promieni świetlnych na kuli i, w przypadku jąder ciężkich, teoria wielokrotnego rozpraszania wyjaśnia przewidywania modelu optycznego stosowanego do opisu reakcji jądrowych dla małych i pośrednich energii (→ Reakcje jądrowe).

spójna produkcja cząstek

Spójna produkcja cząstek (rys. 1b) jest charakterystycznym procesem występującym wyłącznie w obszarze wielkich energii. Cząstka padająca „dysocjuje" w polu jądra (np. proton „dysocjuje" na proton i dwa mezony π), przy czym jądro bombardowane pozostaje w stanie podstawowym. Ten ciekawy proces, mało jeszcze zbadany, jest obecnie przedmiotem wielu prac eksperymentalnych i teoretycznych.

procesy nieelastyczne

Procesy nieelastyczne są bardzo rozmaite. Na rys. 1c przedstawiono schematycznie prosty proces, w wyniku którego z jądra bombardowanego emitowany jest jeden fragment jądrowy (np. proton lub deuteron). Reakcje tego typu badane są intensywnie szczególnie przy energiach rzędu 100 MeV, łatwiej dostępnych ze względu na dużą liczbę akceleratorów

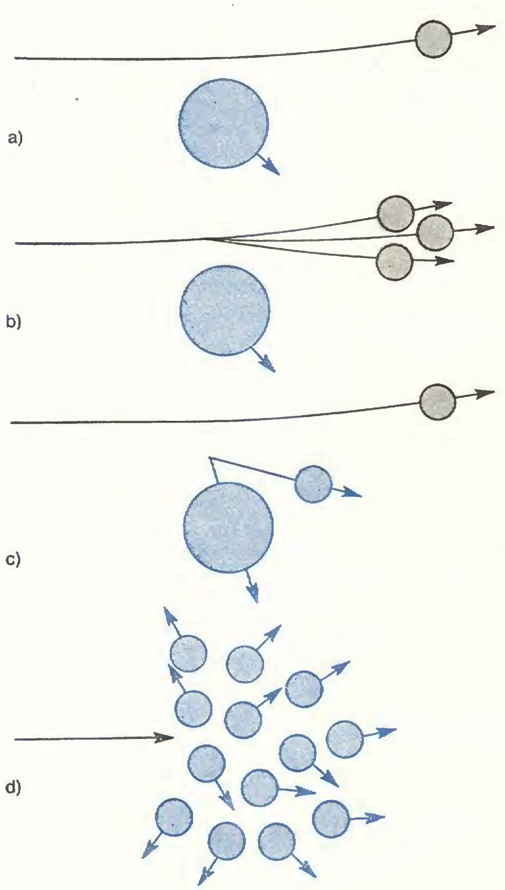
cząstek pośrednich energii. Badanie tych reakcji przy wielkich energiach dostarcza nowych informacji o strukturze jądra, ilustruje to rys. 2. Cząstka o wielkiej energii związana jest ze znacznie krótszą falą de Broglie'a zgodnie z relacją $p = h/\lambda$, gdzie p jest pędem cząstki, h — stałą Plancka, a λ — długością fali i „widzi" — w przeciwieństwie do cząstki o małej energii — tylko małą część jądra bombardowanego (podobnie krótkie fale rentgenowskie służą do badania szczegółów struktury sieci krystalicznej, które trudno zbadać za pomocą światła widzialnego, mającego znacznie większą długość fali).

fragmentacja jądra

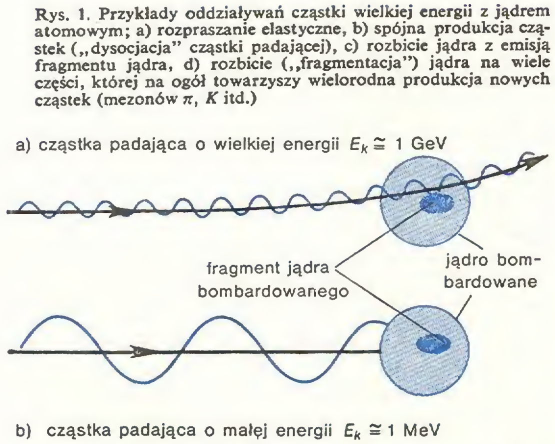
Oddziaływania nieelastyczne mogą mieć jeszcze bardziej skomplikowany charakter. Jeżeli np. jądro bombardowane w wyniku zderzenia z cząstką padającą ulegnie rozbiciu na wiele części, to proces taki nazywamy fragmentacją jądra (rys. 1d). Procesy fragmentacji jądra, mimo swojej złożoności, wykazują wiele regularności empirycznych (np. zbliżony do maxwellowskiego rozkład energii fragmentów), które jeszcze nie są wytłumaczone.

wielorodna produkcja cząstek

Innym, bardzo złożonym procesem nieelastycznym, który często występuje w zderzeniach cząstki o wielkiej energii z jądrem, jest wielorodna produkcja cząstek. W wyniku tego procesu nie tylko zmienia się stan jądra bombardowanego, z którego wyemitowane mogą być zwykłe (składające się z nukleonów) fragmenty jądrowe, ale wyprodukowane są nowe cząstki (np. mezony π). Ogólny opis tego procesu jest bardzo złożony, ale i tu stwierdzono proste empiryczne regularności. Na przykład stwierdzono, że krotność cząstek wyprodukowanych w zderzeniu z jądrem, szczególnie kiedy ograniczymy się do małych kątów emisji, nieoczekiwanie mało różni się od krotności cząstek wyprodukowanych w zderzeniu proton-proton. Analiza tych regularności przyczyni się do głębszego teoretycznego opisu zjawiska, badanego ostatnio niezwykle intensywnie.



Rys. 1. Przykłady oddziaływań cząstki wielkiej energii z jądrem atomowym; a) rozpraszanie elastyczne, b) spójna produkcja cząstek („dysocjacja" cząstki padającej), c) rozbicie jądra z emisją fragmentu jądra, d) rozbicie („fragmentacja") jądra na wiele części, której na ogół towarzyszy wielorodna produkcja nowych cząstek (mezonów π, K itd.)



Rys. 2. Schemat ilustrujący różnicę pomiędzy sposobem oddziaływania z jądrem cząstki o: a) małej, b) wielkiej energii

Znaczenie poznawcze

cząstki elementarne i jądra

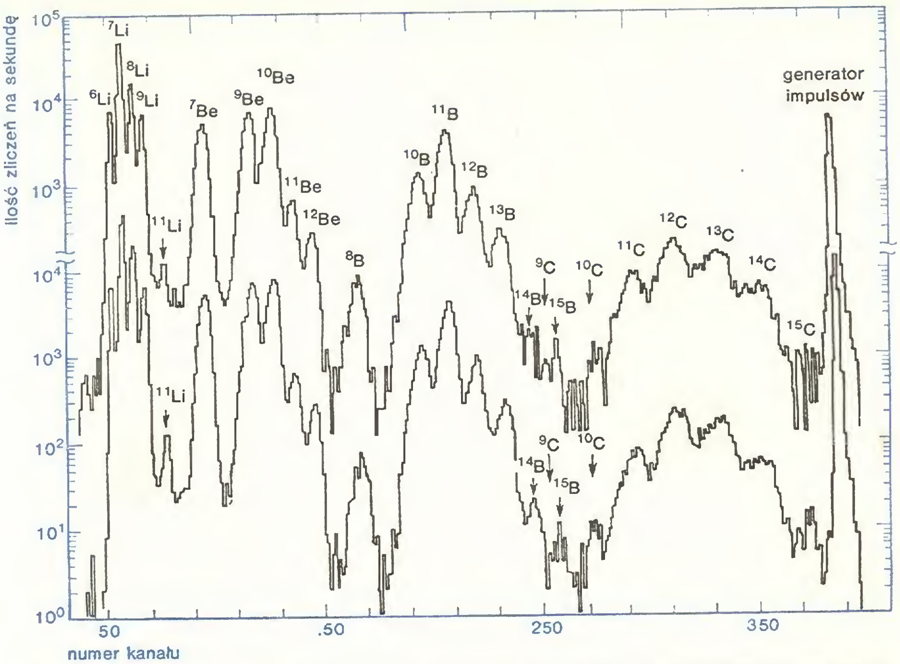
Badanie typów oddziaływań cząstek wielkiej energii z jądrami jest źródłem informacji o cząstkach elementarnych oraz jądrach atomowych. Wiele ważnych odkryć w fizyce cząstek elementarnych dokonano badając oddziaływania cząstek wielkich energii z jądrami (np. odkrycie mezonów π i K, hiperonów, hiperjąder i in.). Własności samych cząstek elementarnych poznajemy jednak głównie badając wzajemne ich oddziaływania, chociaż wiele tych własności można zbadać również w oddziaływaniach jądrowych. Na przykład w ostatnich latach wielkie zainteresowanie fizyków wzbudziły obserwacje wielorodnej produkcji mezonów w oddziaływaniach proton-jądro. Wspomniana już mała krotność produkcji mezonów interpretowana jest jako przejaw długiego okresu trwania oddziaływania, prowadzącego do powstania mezonów (cząstki te „nie zdążą się" wytworzyć wewnątrz jądra). Obecność innych nukleonów w jądrze może stać się w tym wypadku swoistym analizatorem niezwykle krótkich czasów (rzędu 10^{-23} s) oddziaływań, które nie są dostępne w badaniach bezpośrednich. W ten sposób można np. dowiedzieć się o pewnych własnościach krótkotrwałych stanów rezonansowych cząstek.

Powyżej wspomniano o badaniu struktury i własności jąder za pomocą cząstek o wielkich energiach. Jest to dodatkowe źródło informacji w stosunku do tej, której dostarcza nam badanie reakcji jądrowych o małych energiach. Niektóre jednak własności materii jądrowej mogą być badane w warunkach ziemskich jedynie w oddziaływaniach o wielkich energiach. Na przykład w zjawisku fragmentacji jąder cząstka padająca przekazuje jądru energię porównywalną z całkowitą energią wiązania. W ten sposób materia jądrowa w czasie zderzenia zostaje doprowadzona do stanu, w którym w warunkach ziemskich nie występuje (z wyjątkiem bardzo rzadkich oddziaływań cząstek promieniowania kosmicznego z jądrami). Tego rodzaju ekstremalne stany materii jądrowej mogą istnieć jedynie we wnętrzu niektórych osobliwych ciał niebieskich (np. w gwiazdach neutronowych). A zatem w laboratorium na Ziemi za pomocą akceleratorów cząstek wielkich energii możemy wytworzyć i zbadać nowe, dotychczas niedostępne stany i własności materii jądrowej. Ostatnio niektórzy wybitni teoretycy, m.in. A. Migdał i T.D. Lee, wysunęli hipotezę, że w pewnych ekstremalnych warunkach wskutek zgęszczenia powstałego w wyniku zderzenia z cząstką o wielkiej energii materia jądrowa może przejść w zupełnie inny jakościowo stan (sięgając do analogii można posłużyć się przykładem lodu, który w odpowiednim zakresie temperatury i ciśnienia może wystąpić w odmianach allotropowych).

stany materii nie występujące normalnie

Badania oddziaływań jądrowych wielkich energii są pośrednim źródłem dalszych informacji o własnościach jąder. Mianowicie w zderzeniach jądrowych często jądra żelaza, a nawet jądra znacznie cięższe). Ich badanie umożliwia poznanie struktury ciał niebieskich i własności przestrzeni kosmicznej.

informacje z dziedziny kosmologii



Rys. 3. Częstość występowania fragmentów jądrowych, emitowanych w oddziaływaniach protonów o energii 5 GeV z jądrami uranu. Na osi odciętej podany jest numer kanału analizatora mierzącego straty energii w detektorach półprzewodnikowych, użytych dla identyfikacji cząstek (wg A.M. Poskanzer i in. Physical Review Letters 17, 1271, 1966)

stany ekstremalne jąder

o wielkiej energii powstają anomalne jądra ekstremalne. Kiedy cząstka o małej energii pada na jądro, to w wyniku reakcji jądrowej powstać może niewielka ilość różnych produktów końcowych. Jeżeli zwiększamy energię cząstki padającej, to mogą zostać wyprodukowane rzadkie lub zupełnie nowe jądra, które są dalekie od „ścieżki trwałości" (określonej stosunkiem ilości protonów do neutronów dla ustalonej liczby masowej jąder). Sytuację tę ilustruje rys. 3, na którym podane są wyniki identyfikacji niektórych produktów reakcji oddziaływania protonów o energii 5 GeV z jądrami uranu. W doświadczeniu tym odkryto nowe jądra ^{11}Li, ^{14}B, ^{15}B. Wyprodukowane i zarejestrowane w tym doświadczeniu jądro ^{11}Li składa się z trzech protonów i aż ośmiu neutronów, podczas gdy trwałe izotopy litu mają tylko trzy, cztery neutrony! Te anomalne jądra są źródłem nowych informacji o własnościach sił jądrowych.

Źródłem interesującej informacji z dziedziny kosmologii są oddziaływania jąder padających na Ziemię w postaci promieni kosmicznych. W promieniowaniu kosmicznym występują oprócz protonów również cięższe jądra (wśród nich np. stosunkowo

Wymienione przykłady ilustrują niektóre tylko wyniki badań w tej szybko rozwijającej się dziedzinie. W najbliższym czasie zasięg badań rozszerzy się w związku z tym, że kilka akceleratorów protonowych (Dubna, Saclay, Berkeley) przyspiesza obecnie cięższe jądra do bardzo wielkiej energii. Na przykład w Berkeley w Stanach Zjednoczonych akcelerator „Bevatron" przyspiesza obecnie jądra lekkie (do neonu) do energii 3 gigaelektronowoltów na nukleon, a synchrofazotron ZIBJ w Dubnej aż do energii 5 GeV na nukleon. W ten sposób powstała nowa dziedzina fizyki — relatywistyczna fizyka jądrowa.

relatywistyczna fizyka jądrowa

J. Bartke *Eksperymenty z jądrami cięższymi od wodoru przyspieszanymi do bardzo wysokiej energii*, Post. Fiz. 24, 423 (1973); W. Czyż *Rozpraszanie cząstek wysokich energii na jądrach atomowych* w: *Cząstki elementarne, Jądro atomowe, Promieniotwórczość, W Hołdzie Marii Skłodowskiej-Curie*, Warszawa 1967; A. Jachołkowski *Zastosowania modelu Glaubera do opisu rozpraszania cząstek elementarnych i jąder atomowych w zakresie wielkich energii*, Post. Fiz. 26, 167 (1975); M. Mięsowicz *Fireball Model of Meson Production* w *Progress in Elementary Particle and Cosmic Ray Physics*, ed J.G. Wilson, S.A. Wouthuysen, vol. 10, 103 (1971); M. Mięsowicz, R. Sosnowski *Cluster production in high energy reactions*, Nukleonika 20, 24 (1975); *Nukłotron i relatiwistskaja jadiernaja fizika*, Sbornik statiej, O.I.J.I., Dubna 1974

Hiperjądra

Jerzy Pniewski

W latach trzydziestych sądzono, że protony i neutrony, nazywane nukleonami, są jedynymi składnikami jąder atomowych. Wraz z rozwojem fizyki cząstek elementarnych wykryto trzeci składnik, który daje się związać siłami jądrowymi z nukleonami w jedną strukturę stanowiącą nowy rodzaj materii jądrowej. Jest nim hiperon Λ (lambda), a nowa struktura jądrowa, zawierająca poza nukleonami przynajmniej jeden hiperon Λ, zwana jest hiperjądrem. Hiperjądra z powłokami odpowiednio obsadzonymi przez elektrony tworzą nowe izotopy znanych atomów. Hiperjądra zostały odkryte w 1952 roku przez M. Danysza i J. Pniewskiego.

Własności hiperjąder

Neutron i hiperon Λ są cząstkami elektrycznie obojętnymi, zatem, tak jak w przypadku zwykłego jądra, o liczbie atomowej hiperjądra decyduje liczba

nomenklatura hiperjąder

jego protonów. Najprostsze hiperjądro, zawierające wszystkie trzy składniki (p, n, Λ) występujące pojedynczo, jest hiperjądrem wodoru 3. Nazwy hiperjąder wywodzą się od nazw odpowiednich pierwiastków chemicznych, mówimy więc o hiperwodorze 3 lub hipertrycie. Jeśli do jądra helu 4, czyli cząstki α, dołączyć hiperon Λ, powstanie najczęściej spotykane hiperjądro — hiperhel 5. Przy zapisie hiperjąder korzysta się z oznaczeń typowych dla zwykłych jąder, zaopatrzonych dodatkowo w symbol hiperonu Λ. Na przykład: hipertryt — ${}^{3}_{\Lambda}H$, hiperhel 5 — ${}^{5}_{\Lambda}He$. Dotychczas zidentyfikowano 22 hiperjądra, czyli jednoznacznie wyznaczono ich liczby masowe i atomowe, od hiperwodoru do hiperazotu (tabela). Stwierdzono istnienie

Znane hipernuklidy

Z \ A	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	${}^{3}_{\Lambda}H$	${}^{4}_{\Lambda}H$												
2		${}^{4}_{\Lambda}He$	${}^{5}_{\Lambda}He$	${}^{6}_{\Lambda}He$	${}^{7}_{\Lambda}He$	${}^{8}_{\Lambda}He$	·							
3					${}^{7}_{\Lambda}Li$	${}^{8}_{\Lambda}Li$	${}^{9}_{\Lambda}Li$	·		·				
4					${}^{7}_{\Lambda}Be$	${}^{8}_{\Lambda}Be$	${}^{9}_{\Lambda}Be$	${}^{10}_{\Lambda}Be$	·	·	·			
5							${}^{9}_{\Lambda}B$	${}^{10}_{\Lambda}B$	${}^{11}_{\Lambda}B$	${}^{12}_{\Lambda}B$	·	·	·	·
6								·	·	${}^{12}_{\Lambda}C$	${}^{13}_{\Lambda}C$	${}^{14}_{\Lambda}C$	·	·
7											·	·	${}^{15}_{\Lambda}N$	·
8												·	·	·

Kropkami zaznaczono położenia hipernuklidów dotychczas nie obserwowanych ($Z \leq 8$, $A \leq 16$), których istnienia można oczekiwać na podstawie danych jądrowych (hipernuklidy zidentyfikowane za pomocą emulsji jądrowych w laboratoriach Europejskiej Współpracy K^-).

Systematyczne, wieloletnie badania własności hiperjąder prowadzone były przez międzynarodowy zespół laboratoriów zwany Europejską Współpracą K^-. Współpraca ta ostatecznie skoncentrowała się w laboratoriach w Belgradzie, Berlinie, Brukseli, Dublinie, Londynie i Warszawie. Prace tego zespołu objęły całą tematykę hiperjądrową dostępną techniką emulsji jądrowych. W poszczególnych ośrodkach brali w niej udział m.in.: D. H. Davis (Londyn), A. Montwill (Dublin), U. Krecker (Berlin), J. Sacton (Bruksela), J. Zakrzewski (Warszawa). Ogółem opublikowano blisko 100 prac, wiele z nich dotyczyło B_Λ.

ciężkich hiperjąder o masach zbliżonych do masy atomu srebra, a nawet większych. Nie udało się jednak ustalić ani dokładnych wartości ich liczb masowych, ani atomowych (J. Zakrzewski i inni — prace zespołowe od 1961).

Z ogólnej liczby ponad 200 000 hiperjąder, indywidualnie zaobserwowanych w emulsjach jądrowych, zidentyfikowano kilkanaście tysięcy. Znane są dwa przypadki tzw. hiperjąder podwójnych, zawierających po dwa hiperony Λ; są nimi hiperberyl 10 i hiperhel 6:

$$({}^{8}Be+\Lambda+\Lambda) \rightarrow {}^{10}_{\Lambda\Lambda}Be; \qquad (\alpha+\Lambda+\Lambda) \rightarrow {}^{6}_{\Lambda\Lambda}He.$$

rozpady hiperjąder

Hiperon Λ jest cząstką nietrwałą, rozpadającą się w wyniku słabych oddziaływań, ze średnim czasem życia $2{,}63 \cdot 10^{-10}$ s. W rozpadzie tym traci dziwność, opisywaną przez liczbę kwantową S, której ujemne wartości stanowią charakterystyczną cechę wszystkich hiperonów. (Liczba kwantowa S cząstek niedziwnych, np.: nukleonów i pionów, jest równa 0. We wszystkich procesach wywołanych oddziaływaniami silnymi bądź elektromagnetycznymi liczba S jest zachowana). Ponieważ różnica mas hiperonu Λ i nukleonu jest większa od masy mezonu π, istnieją dwa główne kanały rozpadu swobodnych cząstek Λ zwane mezonowymi: $\Lambda \rightarrow p+\pi^-+37{,}75$ MeV oraz $\Lambda \rightarrow n+\pi^0+41{,}06$ MeV. Inne rozpady hiperonu Λ są niezwykle mało prawdopodobne i mogą tu nie być brane pod uwagę. Związanie hiperonu Λ w hiperjądrze, mimo wydzielonej energii wiązania, nie zahamowuje procesu jego rozpadu. Przeciwnie, słabe oddziaływania z nukleonami otwierają nowe kanały rozpadów niemezonowych, w których cząstka Λ tracąc dziwność zamienia się bezpośrednio w neutron: np: $\Lambda+N \rightarrow n+N+176{,}0$ MeV. W ten sposób wszystkie hiperjądra są nietrwałe — są one w pewnym sensie promieniotwórcze, ich średni czas życia jest porównywalny z czasem życia swobodnej cząstki Λ.

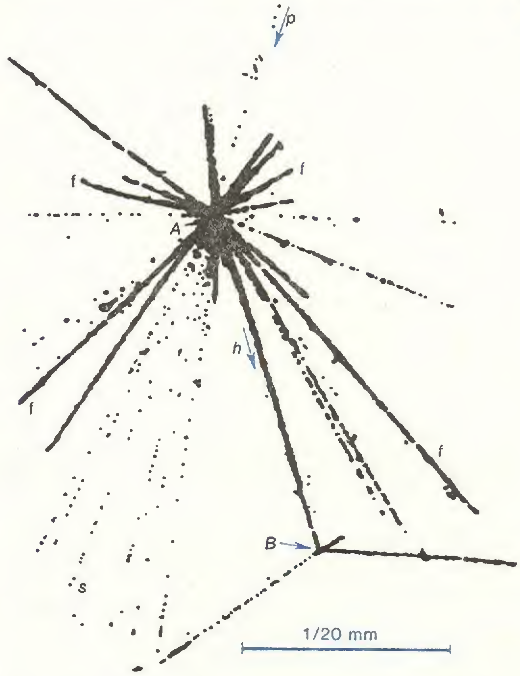
Przykłady najczęściej spotykanych rozpadów hiperhelu 5:

$$\left.\begin{array}{l} {}^{5}_{\Lambda}He \rightarrow \pi^- + p + \alpha + 34{,}63 \text{ MeV}, \\ {}^{5}_{\Lambda}He \rightarrow \pi^- + d + {}^{3}He + 16{,}28 \text{ MeV}, \\ {}^{5}_{\Lambda}He \rightarrow \pi^0 + n + \alpha + 37{,}94 \text{ MeV}, \end{array}\right\} \text{mezonowe}$$

$${}^{5}_{\Lambda}He \rightarrow n + \alpha + 172{,}91 \text{ MeV — niemezonowy.}$$

Charakterystycznym przykładem rozpadu hiperjądra jest hiperjądro zidentyfikowane w emulsji jądrowej, znalezione i rozpoznane po raz pierwszy w 1952 r.

pierwsze hiperjądro



Rys. 1. Mikroskopowy obraz produkcji i rozpadu pierwszego hiperjądra zidentyfikowanego w fotograficznej emulsji jądrowej naświetlonej promieniowaniem kosmicznym (M. Danysz i J. Pniewski, 1952). Oznaczenia: p tor cząstki pierwotnej promieniowania kosmicznego, A miejsce oddziaływania z jądrem bromu (lub srebra) napotkanym w emulsji fotograficznej, s pęk torów szybkich cząstek wtórnych, f tory fragmentów rozbitego jądra, h tor fragmentu hiperjądrowego, B miejsce rozpadu hiperjądra

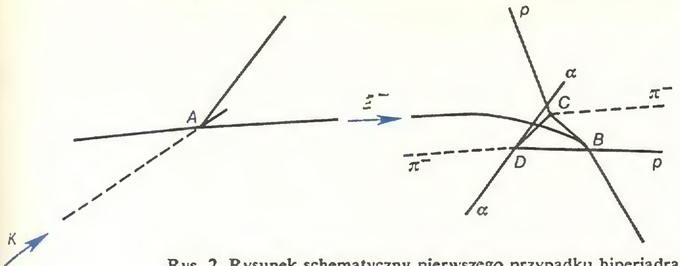
w Warszawie (rys. 1). Jądrowe emulsje fotograficzne, ze względu na dużą przestrzenną zdolność rozdzielczą, umożliwiają analizę indywidualnych przypadków hiperjąder. Analiza kinematyczna rozpadów mezonowych z emisją mezonu π^- stanowi główną podstawę identyfikacji hiperjąder.

produkcja hiperjąder

W celu otrzymania hiperjądra należy stworzyć warunki do wyprodukowania hiperonu Λ, a następnie związania go w jądrze atomowym. Początkowo obserwowano jedynie hiperjądra wytwarzane w zderzeniach cząstek promieniowania kosmicznego albo protonów i pionów wysokich energii uzyskiwanych z akceleratorów protonowych wielkich energii. Jądra atomów wchodzących w skład emulsji fotograficznej stanowiły tarczę do produkcji hiperonów i materiał wyjściowy do uzyskania hiperjąder w formie hiperfragmentów jądrowych. Wydajność tego procesu w emulsji jądrowej wynosiła około 0,1% wszystkich oddziaływań wielkich energii. Następnie zaczęto stosować znacznie bardziej wydajną metodę produkcji hiperjąder, jaką jest wychwyt zatrzymujących się mezonów K^- posiadających tę samą dziwność $S = -1$ co hiperony Λ. Wydajność produkcji hiperjąder tą metodą w tarczy węglowej sięga 10%, w srebrze zaś przekracza 50% chwytanych mezonów K^-.

dziwność hiperjąder

Hiperjądrom należy przypisać tę samą dziwność $S = -1$ co hiperonom Λ i mezonom K^-, natomiast hiperjądrom podwójnym (rys. 2) dziwność $S = -2$.



Rys. 2. Rysunek schematyczny pierwszego przypadku hiperjądra podwójnego zidentyfikowanego jako $^{10}_{\Lambda\Lambda}Be$ (odkrytego w Warszawie przez zespół: M. Danysz, K. Garbowska, J. Pniewski T. Pniewski, J. Zakrzewski, 1962).

Autorzy nie wykluczali możliwości interpretacji tego przypadku jako przykładu podwójnego hiperberylu 11, uważając ją jednak za mniej prawdopodobną.

Oznaczenia: A miejsce oddziaływania mezonu K^- (1,5 GeV/c) i produkcji hiperonu Ξ^-, B miejsce wychwytu hiperonu Ξ^- przez lekkie jądro emulsji (C, N, O) i produkcji hiperjądra $^{10}_{\Lambda\Lambda}Be$, BC tor $^{10}_{\Lambda\Lambda}Be$, C miejsce rozpadu $^{10}_{\Lambda\Lambda}Be$, CD tor hiperjądra $^{9}_{\Lambda}Be$, D miejsce rozpadu $^{9}_{\Lambda}Be$, linie przerywane są torami mezonów K^- i π^-

Hiperjądra podwójne mogą się tworzyć w reakcji wychwytu hiperonów Ξ^- o dziwności $S = -2$ (np.: $\Xi^- + {}^{12}C = {}^{10}_{\Lambda\Lambda}Be + d + n + 12{,}6$ MeV). Inne cząstki o dziwności ujemnej pod wpływem silnych oddziaływań z nukleonami w szybkim procesie przekazują swą dziwność tworzącej się cząstce Λ, najlżejszemu z hiperonów. W ten sposób poza nukleonami hiperony Λ mogą być jedynymi cząstkami wchodzącymi w skład materii typu jądrowego. Możliwość wiązania hiperonu Λ w materii jądrowej, powolny rozpad hiperonów i mezonów K na cząstki oddziaływające silnie, mimo że produkowane były w szybkim akcie zderzenia takich cząstek, wreszcie stwierdzenie skojarzonej produkcji hiperonów i mezonów K — stały się podstawą do wprowadzenia i ugruntowania pojęcia dziwności.

Fizyka hiperjąder zajmuje się badaniem własności hiperjąder, oddziaływań hiperonów Λ z nukleonami oraz dostarcza pewnych informacji o nietrwałych zwykłych strukturach jądrowych. Najcenniejszą informacją uzyskiwaną z analizy kinematycznej rozpadów mezonowych hiperjąder jest energia wiązania hiperonu Λ (B_Λ), zdefiniowana jako energia potrzebna do usunięcia go z hiperjądra. Wartości B_Λ rosną wraz z liczbą masową hiperjądra od (0,13±0,5) MeV dla $^{3}_{\Lambda}H$ do (13,59±0,15) MeV dla $^{15}_{\Lambda}N$ sięgając 23 MeV dla znacznie cięższych hiperjąder (tabela).

energia wiązania hiperonu Λ

Energia wiązania hiperonu Λ w różnych hiperjądrach

Hiperjądro	B_Λ, MeV	Hiperjądro	B_Λ, MeV	Hiperjądro	B_Λ, MeV
$^{3}_{\Lambda}H$	0,13±0,05	$^{7}_{\Lambda}Be$	5,16±0,08	$^{10}_{\Lambda}B$	8,89±0,12
$^{4}_{\Lambda}H$	2,04±0,04	$^{8}_{\Lambda}Li$	6,80±0,03	$^{11}_{\Lambda}B$	10,24±0,05
$^{4}_{\Lambda}He$	2,39±0,03	$^{8}_{\Lambda}Be$	6,84±0,05	$^{12}_{\Lambda}B$	11,37±0,06
$^{5}_{\Lambda}He$	3,12±0,02	$^{9}_{\Lambda}Li$	8,53±0,15	$^{13}_{\Lambda}C$	11,69±0,12
$^{6}_{\Lambda}He$	4,25±0,10	$^{9}_{\Lambda}Be$	6,71±0,04	$^{14}_{\Lambda}C$	12,17±0,33
$^{7}_{\Lambda}Li$	5,58±0,03	$^{9}_{\Lambda}B$	7,88±0,15	$^{15}_{\Lambda}N$	13,59±0,15

Ciężkie hiperjądra obserwowane w emulsji fotograficznej: $B_\Lambda \leq 23$ MeV (wg danych uzyskanych przez laboratoria Europejskiej Współpracy K^-. Energie B_Λ wyznaczone wg programu opracowanego przez W. Gajewskiego).

W przypadku rozpadów niemezonowych emisja nierejestrowanych w emulsji neutronów czyni te rozpady na ogół mało przydatnymi do identyfikacji i badania własności hiperjąder. Hipertryt rozpada się mezonowo niemal w 100%, w tym w $^2/_3$ z emisją mezonu naładowanego. W przypadku bardzo ciężkich hiperjąder dominują rozpady niemezonowe (> 99%). Wartości B_Λ wyznaczone dla różnych hiperjąder są podstawą do badania ich struktury i trwałości oraz poznania oddziaływań cząstki Λ z nukleonami.

Analiza oddziaływania hiperonu Λ z nukleonami oraz hiperonów Λ między sobą

hiperjądra s- i p-powłokowe

Hiperjądra o liczbie masowej $A \leq 5$ nazywane są s-powłokowymi, co zgodnie z powłokowym modelem jądra atomowego odpowiada założeniu, że wszystkie ich nukleony wraz z hiperonem Λ znajdują się w powłoce s. Pozostałe znane hiperjądra noszą nazwę p-powłokowych. Mają one zamknięte powłoki s dla nukleonów oraz pewną ich liczbę w powłoce p ($p_{3/2}$ lub $p_{1/2}$). Zakaz Pauliego nie wzbrania hiperonowi Λ, jako cząstce różnej od protonu i neutronu, lokować się w powłoce zamkniętej dla nukleonów. Dopiero w podwójnym hiperjądrze $^{6}_{\Lambda\Lambda}He$ powłoka s jest zamknięta dla wszystkich trzech składników hiperjądrowych. Cząstka Λ, wiązana kolejno z coraz cięższym rdzeniem jądrowym, może stale być lokowana w najgłębszej powłoce s, co uzasadnia wzrost jej energii wiązania wraz z liczbą masową A aż do wartości 23 MeV dla ciężkich hiperjąder obserwowanych w emulsji fotograficznej. W zwykłym jądrze związanie nowego nukleonu możliwe jest jedynie w jego zewnętrznej powłoce i nie prowadzi do wzrostu energii wiązania wraz ze wzrostem liczby A. Przedstawiając oddziaływanie hiperonu Λ z rdzeniem za pomocą uproszczonego średniego potencjału, mającego postać prostokątnej studni o szerokości zależnej od rozmiarów hiperjądra, można uzyskać w przybliżeniu właściwy związek B_Λ z liczbą A.

Dla kilku hiperjąder udało się wyznaczyć ich spiny na podstawie danych o względnej częstości występowania różnych rozpadów mezonowych. Fakt, że spiny trzech najprostszych hiperjąder ($^{3}_{\Lambda}H$, $^{4}_{\Lambda}H$, $^{4}_{\Lambda}He$) równe są różnicy spinów ich rdzeni jądrowych i spinu cząstki Λ (odpowiednio: $^1/_2 = 1 - ^1/_2$, $0 = ^1/_2 - ^1/_2$, $0 = ^1/_2 - ^1/_2$) wskazuje, że oddziaływanie hiperonu Λ z nukleonem w stanie singletowym (wypadkowy spin równy 0) dominuje nad oddziaływaniem trypletowym (wypadkowy spin równy 1).

Energie wiązania hiperonu Λ w hiperjądrach zwierciadlanych: $^{4}_{\Lambda}H$ (= p, $2n$, Λ) i $^{4}_{\Lambda}He$ (= $2p$, n, Λ) są nieco różne i wynoszą odpowiednio (2,04±0,04) MeV i (2,39±0,03) MeV. Poprawka na oddziaływanie kulombowskie dwóch protonów w $^{4}_{\Lambda}He$ powiększa jeszcze tę różnicę. Wytłumaczenia należy szukać w łamaniu tzw. symetrii ładunkowej w oddziaływaniu hiperonu Λ z nukleonem, tzn. występowaniu pewnej różnicy między oddziaływaniami Λn i Λp.

Energia B_Λ wyznaczona z rozpadu pierwszej cząstki Λ podwójnego hiperjądra $^{10}_{\Lambda\Lambda}Be$ jest o (4,3±0,4) MeV większa od wartości B_Λ uzyskanej w drugim rozpadzie zwykłego już hiperjądra $^{9}_{\Lambda}Be$. Można stąd wnioskować, że oddziaływanie między hiperonami Λ w stanie, w jakim są związane w podwójnym hiperjądrze, jest przyciągające. Zebrane dane o energii B_Λ umożliwiły ilościową charakterystykę oddziaływań hiperonu Λ z nukleonami (m.in. zajmowali się tym: R.H. Dalitz — od 1958 r., B.W.Downs, A.R. Bodmer, R.C. Herndon, Y.C. Tang, A. Gal, J. Dąbrowski, A. Deloff). Ustalono w ten sposób, że w strukturach hiperjądrowych dominują przyciągające siły dwuciałowe ΛN, oraz że zasięg tych sił jest mniejszy od zasięgu sił występujących między nukleonami. Zakłada się, że w pierwszym przybliżeniu zasięgi te odpowiadają wymianie dwóch mezonów π, w odróżnieniu od wymiany jednego mezonu uważanej za dominującą w opisie oddziaływań między nukleonami. Dokładniejsza analiza wskazuje na bardziej złożony charakter badanych oddziaływań, m.in. zakłada się obecnie pewien udział sił niecentralnych oraz sił trzyciałowych typu ΛNN. Oczekuje się, że energia wiązania hiperonów Λ w przypadku stanów

oddziaływanie Λn, Λp i ΛΛ

wzbudzonych najlżejszych hiperjąder lub równoważna tej informacji znajomość energii wzbudzenia winny stanowić cenny element w analizie oddziaływań hiperonu Λ z nukleonami.

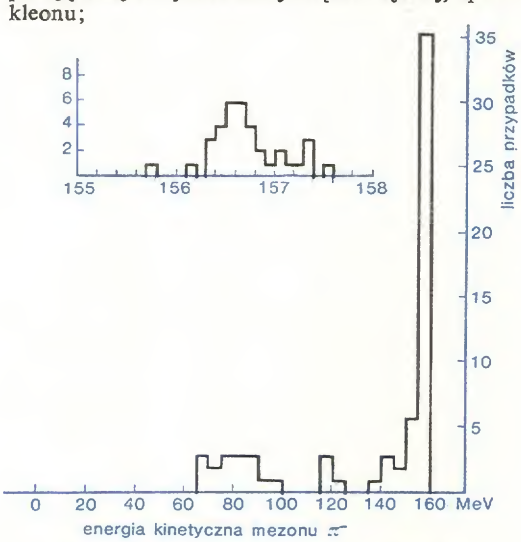
Spektroskopia hiperjądrowa

Duże intensywności wiązek mezonów K^-, jakie od 1969 r. zaczęto uzyskiwać przy użyciu wielkich akceleratorów, umożliwiły zastosowanie do badań hiperjądrowych techniki licznikowej. W ten sposób udało się wykryć emisję fotonów γ ze stanów wzbudzonych hiperjąder o liczbie masowej 4, co zapoczątkowało badania hiperjądrowej spektroskopii γ. Również metodami licznikowymi zaczęto identyfikować i wyznaczać energie mezonów π^- towarzyszących produkowanym hiperjądrom, co z kolei stworzyło podstawy dla spektroskopii hiperjądrowej innego typu, opartej na analizie energii mezonów π^-.

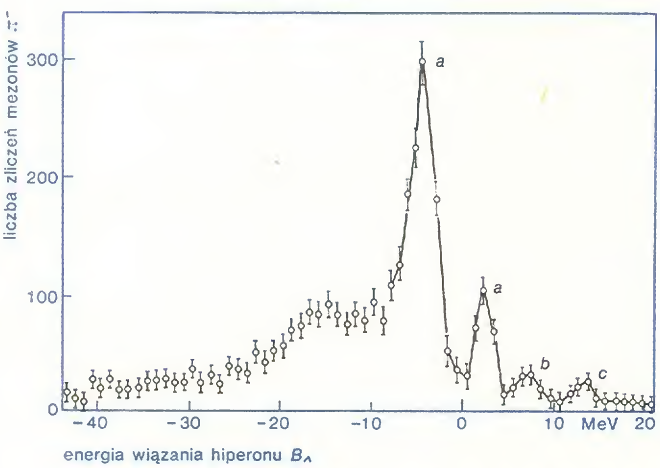
stany wzbudzone hiperjąder

Podział fenomenologiczny stanów wzbudzonych hiperjąder na trzy grupy jest następujący:

1) stany rezonansowe o krótkim czasie życia rozpadające się w wyniku emisji cząstki ciężkiej, np. nukleonu;



Rys. 3. Widmo energii kinetycznej mezonu π^- z wyraźnie widocznym maksimum rezonansowym odpowiadającym stanowi rezonansowemu ${}^{12}_{\Lambda}C^*$: $K^- + {}^{12}C = \pi^- + {}^{12}_{\Lambda}C^*$, ${}^{12}_{\Lambda}C^* \to p + {}^{11}_{\Lambda}B$ (wg danych Europejskiej Współpracy K^-, 1969)



Rys. 4. Widmo mezonów π^- uzyskane przy użyciu techniki licznikowej dla dwuciałowej reakcji: $K^- + {}^{16}O = \pi^- + {}^{16}_{\Lambda}O^*$. Maksima reprezentują widmo energetyczne ${}^{16}_{\Lambda}O$: *a* dwa stany wysokowzbudzone (uznawane za rezonanse), *b* stan wzbudzony nierezonansowy, *c* stan podstawowy. Obszar $B_\Lambda \leqslant 0$ odpowiada możliwości emisji hiperonu (wg B. Povh i inni, 1978)

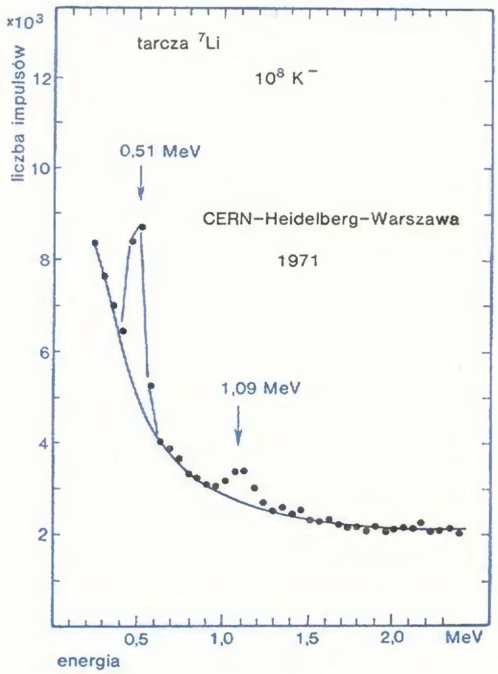
2) stany wzbudzone zanikające wraz z emisją fotonu;

3) stany izomeryczne o najdłuższym czasie życia, obserwowane gdy rozpad hiperonu konkuruje z emisją fotonu γ lub nawet staje się od niej bardziej prawdopodobny.

pierwszy zaobserwowany stan rezonansowy

Istnieją dane eksperymentalne wskazujące na możliwość tworzenia się stanów wzbudzonych hiperjąder wszystkich trzech rodzajów. Pierwszym niewątpliwym przykładem stanu rezonansowego hiperjądra był stan wzbudzony ${}^{12}_{\Lambda}C$, powstający w dwuciałowej reakcji wychwytu mezonu K^- przez jądro węgla 12: $K^- + {}^{12}C = \pi^- + {}^{12}_{\Lambda}C^*$, ${}^{12}_{\Lambda}C^* \to p + {}^{11}_{\Lambda}B$, ${}^{11}_{\Lambda}B \to \pi^- + {}^{11}C$. Produkcja stanu rezonansowego ${}^{12}_{\Lambda}C^*$ (rys. 3) w procesie dwuciałowym zaznacza się wyraźnie w widmie towarzyszących mezonów π^-. Efekt ten zaobserwowano najpierw w emulsji jądrowej, a następnie wykryto przy użyciu licznikowego spektrometru pionów. Dalszy postęp prac prowadzonych techniką licznikową doprowadził do odkrycia kilku nowych wysokowzbudzonych stanów i poznania na tej drodze nowych przedstawicieli struktur hiperjądrowych: hipertlenu 16, hipersiarki 32, hiperwapnia 40 (B. Povh i inni, 1975). Analiza reakcji dwuciałowych prowadzona za pomocą układu elektronicznego umożliwiła wykrycie wielu stanów wzbudzonych (rys. 4), jednak w ten sposób, w odróżnieniu od techniki emulsyjnej, ustalony został jedynie fakt ich produkcji, natomiast nie śledzono ich rozpadów. Mimo to stany te uważa się za rezonansowe, zanikające w wyniku emisji hiperonu Λ czy nukleonu.

W przypadku hipertlenu ${}^{16}_{\Lambda}O$ zaobserwowano dwa stany wysoko wzbudzone, jak również stan podstawowy tego hiperjądra oraz jeden ze stanów najprawdopodobniej zanikający w wyniku emisji fotonu γ, mimo że tego fotonu również nie obserwowano. Wszystkie stany badane na tej drodze znalazły dobre uzasadnienie w powłokowym modelu jądra atomowego. Według tego modelu cząstka Λ powstająca z neutronu lokuje się w jednej z powłok jądrowych,



Rys. 5. Hiperjądrowa linia 1,09 MeV w widmie γ uzyskanym z tarczy ${}^{7}Li$ naświetlonej zatrzymującymi się mezonami K^-. Widoczna dodatkowo linia anihilacji pozytonów odpowiadająca energii 0,51 MeV (wyniki zespołu CERN–Heidelberg–Warszawa. A. Bamberger, M.A. Faessler, U. Lynen, H. Piekarz, J. Piekarz, J. Pniewski, B. Povh, H.G. Ritter, V. Soergel, 1971). W następnym eksperymencie wykonanym przez zespół CERN–Lyon–Warszawa linia ta została rozszczepiona na dwie przypisane ${}^{4}_{\Lambda}H$ i ${}^{4}_{\Lambda}He$ (1979)

a po usuniętym neutronie powstaje luka w zajmowanej przez niego powłoce. Wyniki te dostarczyły informacji o wielkości sprzężenia spinu hiperonu Λ z jego momentem orbitalnym.

pierwsza hiperjądrowa linia widma γ

Pierwszą hiperjądrową linię (rys. 5), wyraźnie występującą w widmie γ, przypisano jednemu lub jednocześnie obu hiperjądrom ${}^{4}_{\Lambda}H$ i ${}^{4}_{\Lambda}He$, tworzonym jako wzbudzone hiperfragmenty w następstwie wychwytu mezonów K^- w tarczach ${}^{6}Li$ i ${}^{7}Li$. Wzbudzenie tych hiperjąder daje się tłumaczyć modelowo jako wynik odwrócenia spinu cząstki Λ związanej z rdzeniem jądrowym ${}^{3}H$ lub ${}^{3}He$. Ponieważ stwierdzono, że spiny ${}^{4}_{\Lambda}H$ i ${}^{4}_{\Lambda}He$ w stanie podstawowym są równe 0, należy sądzić, że stanom wzbudzonym odpowiadają wartości spinów równe 1. Wyznaczenie energii wzbudzenia obu hiperjąder o masie 4 stanie się szczególnie ważnym ogniwem w analizie oddziaływań hiperonu Λ z nukleonami. Obecnie wydaje się, że obserwowana poprzednio linia 1,09 MeV powstała w wyniku nałożenia się dwóch linii niezbyt od siebie odległych dających się przypisać obu hiperjądrom o masie 4. Może to stanowić przesłankę wskazującą na zbliżone wielkości sprzężenia spinu hiperonu Λ ze spinami protonu i neutronu.

izomeria hiperjądrowa

Hiperjądro ${}^{7}_{\Lambda}He$ jest przykładem hiperjądra, które — jak się wydaje — ma izomeryczne stany wzbudzone. Wraz z rozpadem zachodzącym w stanie wzbudzonym wyzwolona jest energia wzbudzenia, o którą zmniejsza się wyznaczana energia wiązania hiperonu Λ. Technika emulsyjna wydaje się z kolei być najlepsza do detekcji tych stanów. W przypadku ${}^{7}_{\Lambda}He$ o istnieniu takiego stanu można wnioskować na podstawie występowania odpowiedniego stanu wzbudzonego rdzenia hiperjądrowego będącego jądrem ${}^{6}He$ (M. Danysz, J. Pniewski, 1962).

wzbudzenie izomeryczne ${}^{7}_{\Lambda}He$

Dobra zdolność rozdzielcza widm γ powinna umożliwić wykrycie nisko położonych stanów wzbudzonych hiperjąder, natomiast widma energetyczne towarzyszących pionów wydają się być najbardziej przydatne do badania wyżej położonych stanów, przede wszystkim stanów rezonansowych, które po emisji nukleonu mogą prowadzić do powstania hiperjąder lżejszych przez analogię do procesów jądrowych. Można oczekiwać, że dalszy rozwój fizyki hiperjąder będzie głównie oparty na badaniach spektroskopowych obu typów. Mimo podjętych bezpośrednich badań nad rozpraszaniem swobodnych cząstek Λ na protonach, fizyka hiperjądrowa jest nadal głównym źródłem informacji o oddziaływaniach tych cząstek z nukleonami w obrębie małych energii.

M. Danysz, J. Pniewski *Delayed disintegration of a heavy nuclear fragment*, Philos. Mag. 44, 348 (1953); D.H. Davis, J. Sacton *Hypernuclear Physics, High Energy Physics*, vol. 2, p. 365, New York 1967; J. Pniewski, J. Zakrzewski *Hypernuclear spectroscopy: A new trend in hypernuclear physics*, Nukleonika 20, 43 (1975); J. Pniewski, D. Ziemińska *Present status of experimental research of hypernuclei*, Nukleonika 23, 797 (1978); J. Pniewski *Początki fizyki hiperjąder*, Post. Fiz. 30, 517 (1979).

Energia jądrowa

Janusz Mika

W 1905 r. Albert Einstein sformułował szczególną teorię względności, która obaliła wiele niewzruszonych, jak się wydawało wówczas, poglądów w fizyce. Jednym z najbardziej rewelacyjnych wniosków wypływających z nowej teorii była równoważność masy i energii. Równoważność tę wyraża słynny wzór Einsteina $E = mc^2$. Masę należy więc traktować jako jedną z form energii, która może przechodzić w inne formy energii.

$E = mc^2$

Masa spoczynkowa układu związanego (jak kryształ, cząsteczka, atom, jądro atomowe) jest mniejsza od sumy mas spoczynkowych składników o wartość równoważną energii wiązania układu, czyli energii potrzebnej na całkowite rozdzielenie składników. Dlatego procesom tworzenia układów związanych towarzyszy wydzielanie się energii (np. energii chemicznej podczas spalania lub energii jądrowej podczas syntezy lekkich jąder). Energia wydziela się również przy przejściu układu ze stanu słabo związanego, tzn. o mniejszej energii wiązania na jeden składnik, w stan związany silniej (np. przy rozszczepieniu ciężkich jąder). Ponieważ siły wiązania nukleonów w jądrze są znacznie potężniejsze niż siły wiązania elektronów w atomie — energie wydzielające się w reakcjach jądrowych są ok. milion razy większe od energii wydzielanych w reakcjach chemicznych (np. spalenie 1 g węgla najwyższej wartości opałowej daje zaledwie ok. 36 kJ, a „spalenie" 1 g uranu — 86 GJ).

energia jądrowa

O praktycznym wykorzystaniu energii zawartej w postaci masy w jądrze atomowym, czyli o energii jądrowej można było mówić dopiero po przeprowadzeniu pierwszej reakcji jądrowej w skali makroskopowej.

Zjawisko rozszczepienia jąder uranu pod wpływem neutronów zostało odkryte w 1938 r. Stwierdzono, że w reakcji rozszczepienia jądro prawie zawsze dzieli się na dwa fragmenty i neutrony swobodne, przy czym średnio na jeden akt rozszczepienia przypada ich ponad dwa. Dzięki temu pojawiła się możliwość uzyskania łańcuchowej reakcji rozszczepienia w skali makroskopowej. Istotnie, już po czterech latach w Chicago zbudowany został pierwszy reaktor jądrowy (atomowy), a wkrótce potem bomba jądrowa (atomowa). Po czterdziestu latach od odkrycia reakcji rozszczepienia znaczna część produkowanej w świecie energii elektrycznej pochodzi z reaktorów jądrowych.

Z energią termojądrową pochodzącą z reakcji syntezy lekkich jąder mamy do czynienia na co dzień w postaci energii słonecznej. Na razie nie udało się jednak, mimo ogromnych wysiłków, zbudować reaktora termojądrowego, w którym reakcja syntezy zachodziłaby w sposób kontrolowany, chociaż skonstruowano bombę termojądrową o mocy tysiąc razy przewyższającej moc bomby atomowej (→ Energia termojądrowa).

Reakcje jądrowe zachodzące w reaktorze

Neutron, jako cząstka obojętna, odgrywa szczególną rolę w reakcjach jądrowych, ponieważ w przeciwieństwie np. do cząstek α lub protonów wnikając do jądra nie musi pokonywać sił odpychania elektrostatycznego; neutrony o bardzo nawet małej energii kinetycznej mogą więc łatwo wywoływać reakcje jądrowe. Do najważniejszych reakcji jądrowych z neutronami zachodzących w reaktorze należą: rozpraszanie sprężyste, rozpraszanie niesprężyste, wychwyt radiacyjny i rozszczepienie (→ Reakcje jądrowe).

rozpraszanie sprężyste neutronu

Reakcja rozpraszania sprężystego nie przechodzi przez etap jądra złożonego i polega na zderzeniu neutronu z jądrem przy spełnieniu zasad zachowania energii kinetycznej i pędu. Rozpraszanie sprężyste zachodzi głównie na lekkich jądrach.

W pozostałych trzech reakcjach pierwszy etap jest taki sam i polega na utworzeniu jądra złożonego. Rozpraszanie niesprężyste zachodzi w wypadku rozpraszania neutronów o dużych energiach (powyżej kilkuset keV) na ciężkich jądrach. Energia wzbudzenia jądra złożonego jest wówczas na tyle duża, że następuje powtórna emisja neutronu o energii niższej

rozpraszanie niesprężyste neutronu

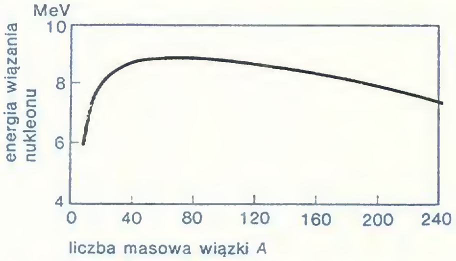
niż energia neutronu padającego, a nadwyżka energii jest wypromieniowana w postaci kwantu γ. W tej sytuacji nie jest zachowana energia kinetyczna.

wychwyt radiacyjny neutronu

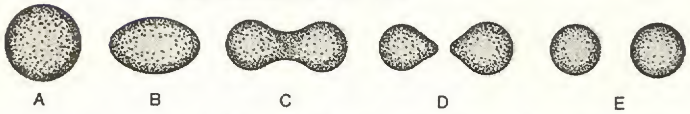
Jeśli energia wzbudzenia jest za mała, aby spowodować wyrzucenie neutronu z jądra, w drugim etapie reakcji jądro przechodzi do stanu podstawowego wypromieniowując kwant γ, a w wyniku powstaje jądro o liczbie masowej o jeden większej. Jest to reakcja wychwytu radiacyjnego.

reakcja rozszczepienia

Jeśli chodzi o reakcję rozszczepienia, to jak wynika z kształtu krzywej na rys. 1, wszystkie jądra o liczbie masowej większej niż 100 powinny, zgodnie z zasadą

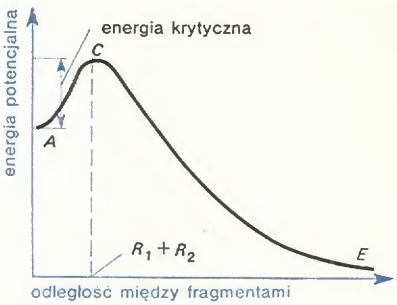


Rys. 1. Zależność wiązania nukleonu w jądrze od liczby masowej jądra (zob. też rys. 1 str. 175)



Rys. 2. Deformacje kropli materii jądrowej w procesie rozszczepienia

zachowania energii, ulegać spontanicznemu rozszczepieniu. Dlaczego jądra takie występują jednak w przyrodzie, wyjaśnia model kroplowy (→ Modele jądrowe). Rysunek 2 przedstawia deformacje kropli materii jądrowej w procesie rozszczepienia, a rys. 3 — zależność energii potencjalnej jądra od kształtu kropli.



Rys. 3. Zmiana energii potencjalnej jądra w procesie rozszczepienia; punkty *A C E* odpowiadają stanom kropli z rys. 2, R_1 i R_2 oznaczają promienie fragmentów

energia krytyczna

Z rysunków widać, że aby wywołać rozszczepienie, trzeba jądru dostarczyć energię równą co najmniej energii krytycznej (energii aktywacji), potrzebną na pokonanie krótkozasięgowych sił jądrowych. Energia krytyczna maleje wraz z liczbą masową jądra i osiąga zero przy liczbie równej ok. 260. Tak więc dopiero jądra o liczbach masowych większych niż 260 są rzeczywiście niestabilne ze względu na spontaniczne rozszczepienie.

Jądro pochłaniające neutron uzyskuje energię równą różnicy energii wiązania jądra złożonego i jądra pochłaniającego neutron (energia ta jest w przybliżeniu równa energii wiązania na 1 nukleon w powstałym jądrze złożonym) powiększoną o energię kinetyczną padającego neutronu. Jeśli energia krytyczna jądra złożonego jest mniejsza niż energia wiązania na 1 nukleon, wówczas rozszczepienie można wywołać za pomocą neutronów o dowolnie niskich energiach. Taka sytuacja zachodzi w wypadku jądra ^{235}U, które ulega rozszczepieniu zgodnie ze schematem: $^{235}U + ^{1}n \rightarrow$ $\rightarrow$ $^{236}U \xrightarrow{\text{rozszczepienie}}$. Jeżeli energia kinetyczna pochłoniętego neutronu jest równa zeru, to energia wzbudzenia jądra złożonego ^{236}U jest równa różnicy energii wiązania jąder ^{236}U i ^{235}U, tj. ok. 6,8 MeV, podczas gdy energia krytyczna ^{236}U wynosi jedynie 6,6 MeV.

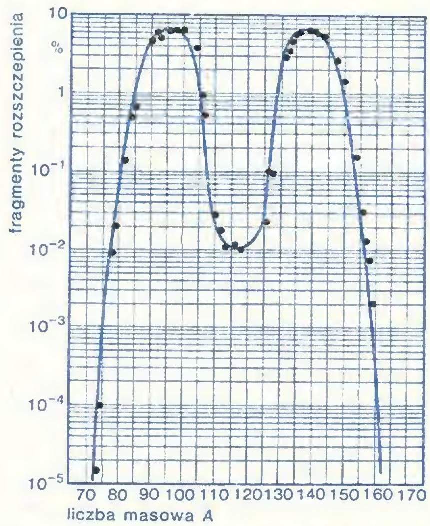
rozszczepienie ^{235}U

Obok ^{235}U drugim izotopem uranu występującym w przyrodzie jest ^{238}U, który stanowi 99,3% uranu naturalnego. Energia krytyczna ^{239}U wynosi 7,0 MeV, natomiast różnica energii wiązania jąder ^{239}U i ^{238}U tylko 5,5 MeV. Tak więc minimalna energia kinetyczna neutronu potrzebna do rozszczepienia jądra ^{238}U, czyli tzw. próg rozszczepienia, powinna być równa 1,5 MeV. Ze względu na przybliżony charakter przytoczonego rachunku liczba ta nie jest dokładna, a w istocie, jak stwierdzono doświadczalnie, próg rozszczepienia ^{238}U wynosi ok. 1,1 MeV, tzn. rozszczepienie ^{238}U mogą wywołać neutrony, których energia kinetyczna jest co najmniej równa 1,1 MeV.

próg rozszczepienia

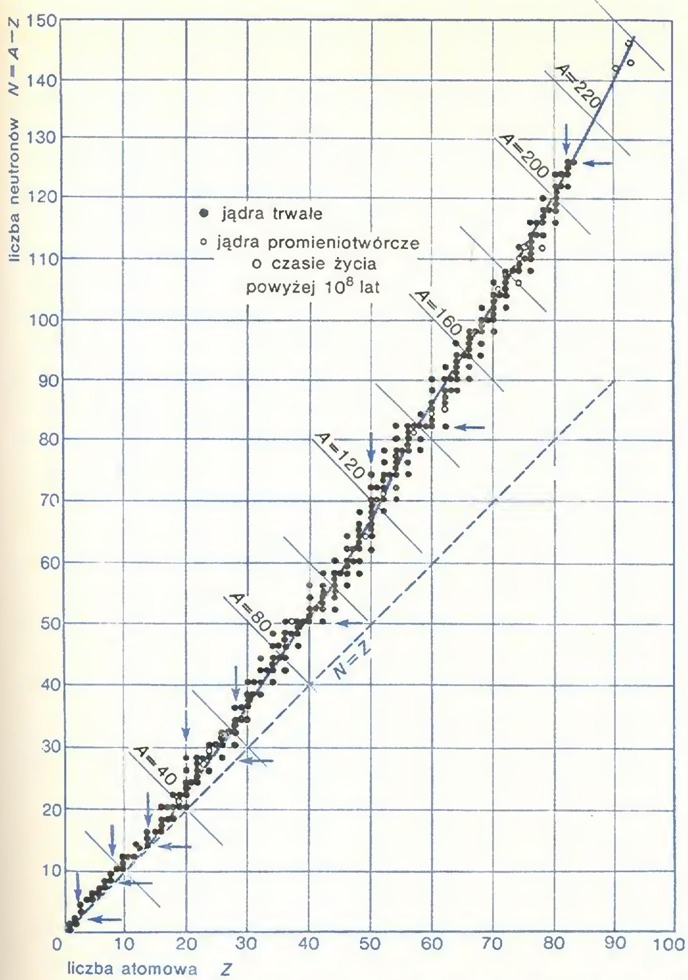
Obok dwóch izotopów uranu występujących w stanie naturalnym istnieje wiele nuklidów rozszczepialnych wytwarzanych sztucznie. Do najważniejszych z praktycznego punktu widzenia należy izotop plutonu ^{239}Pu, którego właściwości ze względu na rozszczepienie są podobne do właściwości ^{235}U.

Reakcja rozszczepienia jądra uranu czy też innego pierwiastka rozszczepialnego prowadzi prawie zawsze do podziału jądra na dwa fragmenty o mniej więcej równych masach. Rozkład mas fragmentów rozszczepienia przedstawia rys. 4. Jak widać, liczby masowe większości jąder powstających z rozszczepienia są

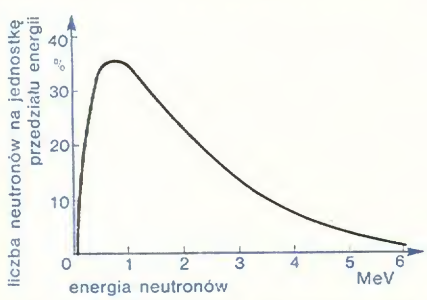


Rys. 4. Rozkład mas fragmentów rozszczepienia $^{235}_{92}U$

zawarte pomiędzy 80 a 110 i pomiędzy 125 a 155. Obserwuje się również tzw. rozszczepienia potrójne, ale ich udział jest niewielki, tak że nie mają one istotnego znaczenia w łańcuchowej reakcji rozszczepienia. Jak wynika z rys. 5, w miarę zwiększania się liczby masowej rośnie stosunek liczby neutronów do liczby protonów w jądrach trwałych, tak że podział jądra na dwa fragmenty prowadzi do nadmiaru neutronów i emisji swobodnych neutronów w liczbie od 1 do 6 na jeden akt rozszczepienia. Średnia liczba neutronów rozszczepieniowych dla ^{235}U wynosi 2,44, jeśli neutrony, które wywołały rozszczepienie, miały energie 0,025 eV. Wraz ze wzrostem energii neutronów liczba ta rośnie i osiąga wartość 2,50 przy energii neutronów 1 MeV. Rozkład energii neutronów rozszczepieniowych dla ^{235}U pokazuje rys. 6. Maksymalna energia neutronu równa się ok. 10 MeV, natomiast obliczona na podstawie tego rozkładu średnia energia wynosi ok. 2 MeV. Dla neutronów wtórnych pocho-



Rys. 5. Zależność między liczbą neutronów i liczbą atomową nuklidów trwałych (ścieżka trwałości nuklidów)



Rys. 6. Rozkład energetyczny neutronów pochodzących z rozszczepienia $^{235}_{92}U$

dzących od innych jąder rozszczepialnych wartości parametrów są podobne.

Emisja neutronów swobodnych w procesie rozszczepienia nie wyczerpuje ich nadmiaru, tak że fragmenty rozszczepienia również zawierają zbyt wielką w stosunku do liczby protonów liczbę neutronów, w związku z tym są z reguły β-promieniotwórcze, tzn. emitują elektrony. Poza tym niektóre fragmenty lub nuklidy powstające z fragmentów rozszczepienia emitują neutrony. Ze względu na to, że neutrony te nie są wydzielane w momencie rozszczepienia jądra, nazywają się neutronami opóźnionymi w przeciwieństwie do pozostałych neutronów rozszczepieniowych, zwanych neutronami natychmiastowymi. Udział neutronów opóźnionych jest bardzo mały i w wypadku ^{235}U wynosi ok. 0,7%, natomiast średnie opóźnienie,

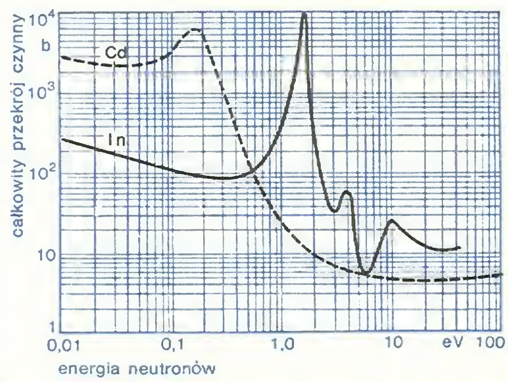
neutrony opóźnione

z jakim pojawiają się one w reakcji łańcuchowej, równe jest ok. 12,5 s. W skali procesów zachodzących w reaktorze jest to czas bardzo długi i neutrony opóźnione odgrywają zasadniczą rolę w praktycznej realizacji łańcuchowej reakcji rozszczepienia.

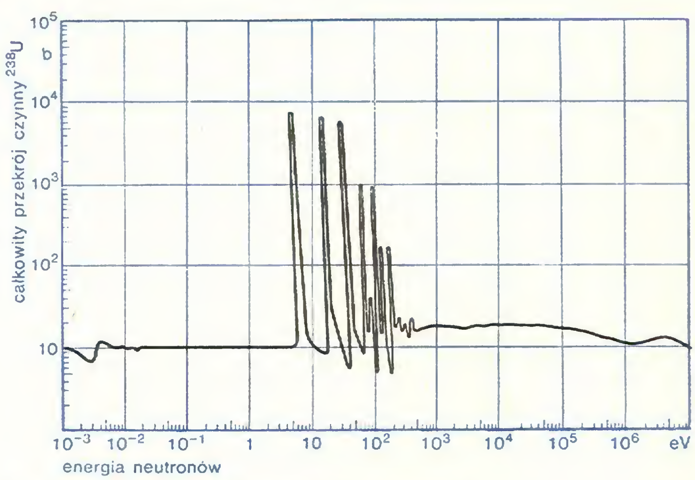
W reakcji rozszczepienia ^{235}U ok. 0,1% masy jądra zamienia się w energię dając ok. 200 MeV. Największa część tej energii przypada na energię kinetyczną fragmentów rozszczepienia, reszta zaś na energię promieniowania β i γ. Podobne relacje liczbowe występują w odniesieniu do rozszczepienia innych jąder rozszczepialnych.

przekrój czynny na pochłanianie

Bardzo ważną z praktycznego punktu widzenia rolę odgrywa zależność przekrojów czynnych od energii padającego neutronu. Dla większości nuklidów (wyjątek stanowi wodór) przekrój czynny na rozpraszanie sprężyste słabo zależy od energii i można go w praktyce uważać za stały. Natomiast przekrój czynny



Rys. 7. Zależność od energii neutronu przekroju czynnego kadmu i indu na pochłanianie neutronów



Rys. 8. Przekrój czynny $^{235}_{92}U$ na pochłanianie neutronów

na rozpraszanie niesprężyste (rys. 7) spada bardzo silnie wraz z energią i dochodzi do zera przy kilkuset keV; zależność od energii przekrojów czynnych na wychwyt radiacyjny i rozszczepienie jest podobna. Z tego względu wystarczy rozpatrywać sumę tych przekrojów, którą przyjęto nazywać przekrojem czynnym na pochłanianie.

W zakresie od zera do ok. 0,1 eV energii neutronów przekrój czynny na pochłanianie neutronów maleje wraz z energią odwrotnie proporcjonalnie do pierwiastka kwadratowego z energii. Od 0,1 eV do ok. 1 keV rozciąga się zakres energii rezonansowych, w którym występują ostre maksima i minima. Szczególnie silne pochłanianie rezonansowe neutronów obserwuje się na jądrach nuklidów rozszczepialnych, np. rys. 8 ukazuje przebieg całkowitego przekroju

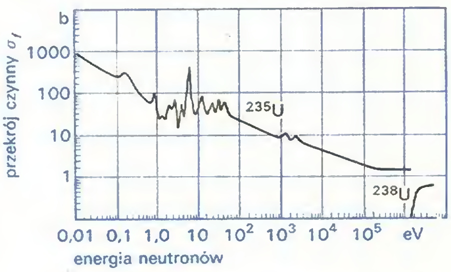
pochłanianie rezonansowe neutronów

czynnego ^{238}U. Wzrost przekroju czynnego na pochłanianie neutronów o małych energiach i efekt rezonansowy mają poważne konsekwencje praktyczne w reaktorze jądrowym.

Warunki pracy reaktora jądrowego

Spowalnianie neutronów

Przekrój czynny na rozszczepienie σ_f dla nuklidów rozszczepialnych rośnie bardzo szybko wraz ze zmniejszaniem się energii padającego neutronu. W wypadku ^{235}U (rys. 9) przy energii neutronów 10 keV

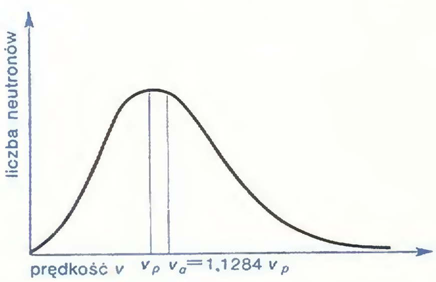


Rys. 9. Przekrój czynny $^{235}_{92}U$ i $^{238}_{92}U$ na rozszczepienie

wynosi on 4 b, natomiast przy energii 0,025 eV — 579 b (1 b = 10^{-28} m²). Tak więc zmniejszając energię neutronów wywołujących rozszczepienie można zwiększyć efektywność reakcji łańcuchowej. W związku z tym do reaktora wprowadza się moderator, którego zadaniem jest spowalnianie neutronów w reakcji rozpraszania sprężystego. Reakcja ta przebiega podobnie jak zderzenie dwóch kul doskonale sprężystych. Przy zderzeniu sprężystym kula poruszająca się (neutron) przekazuje część swej energii kinetycznej kuli nieruchomej (jądro). Ilość przekazanej energii zależy od kąta padania i średnio jest tym większa, im mniejsza jest masa kuli nieruchomej w stosunku do masy kuli ruchomej. Widać więc, że najefektywniej spowalnianie neutronów zachodzić będzie na lekkich jądrach. Dobry moderator musi spełniać poza tym dwa dodatkowe warunki. Po pierwsze, jego gęstość musi być dostatecznie duża (nie może on być np. w stanie gazowym), a po drugie, jego przekrój czynny na wychwyt neutronów musi być stosunkowo niewielki.

moderator

W obecności moderatora proces spowalniania trwa dopóty, dopóki neutrony nie osiągną energii porównywalnych z energią ruchu cieplnego i nie znajdą się w równowadze termicznej z otoczeniem. O takich



Rys. 10. Rozkład Maxwella prędkości neutronów; v_p — prędkość najbardziej prawdopodobna, v_a — prędkość średnia

neutronach mówi się, że mają energię termiczną lub że są neutronami termicznymi. Rozkład energii neutronów termicznych można z dość dużą dokładnością opisać za pomocą rozkładu Maxwella (rys. 10). W temperaturze 300 K średnia prędkość neutronów wynosi 2200 m/s, a średnia energia 0,025 eV.

neutrony termiczne

Bilans neutronów

Warunki, w których może zachodzić w sposób stacjonarny łańcuchowa reakcja rozszczepienia z udziałem neutronów termicznych, przeanalizujemy na przykładzie nieskończonego układu materialnego, zawierającego jednorodną mieszaninę paliwa (np. uran naturalny) i moderatora. Przyjmijmy, że w pewnej chwili w układzie znajduje się S swobodnych neutronów termicznych. Wprawdzie neutrony ulegają rozpadowi, jednak ich czas życia (rzędu 12 min) jest bardzo duży w porównaniu z czasem charakteryzującym łańcuchową reakcję rozszczepienia. Tak więc można uważać wszystkie neutrony termiczne w reaktorze za cząstki trwałe, które zostaną w rozpatrywanym układzie nieskończonym pochłonięte.

Ponieważ pochłanianie w moderatorze prowadzi jedynie do wychwytu radiacyjnego, tylko fS neutronów liczy się w dalszym bilansie, gdzie f oznacza prawdopodobieństwo pochłonięcia neutronu termicznego w paliwie, czyli tzw. współczynnik wykorzystania cieplnego. Zgodnie z definicją f jest zawsze mniejsze od jedności. Pochłonięcie neutronu w paliwie nie oznacza jeszcze, że wywoła on rozszczepienie. Jeśli α wyraża stosunek liczby neutronów powodujących rozszczepienie do liczby neutronów pochłoniętych w reakcji wychwytu radiacyjnego w paliwie, to ostatecznie liczba rozszczepień wywołanych przez S neutronów termicznych będzie równa αfS. Ponieważ w 1 akcie rozszczepienia powstaje średnio ν neutronów rozszczepieniowych, to liczba neutronów wtórnych przypadająca na S neutronów termicznych równać się będzie $\nu\alpha fS$ albo ηfS, gdzie $\eta = \nu\alpha$ oznacza średnią liczbę neutronów wtórnych na 1 akt pochłonięcia w paliwie.

Neutrony prędkie o energii większej niż 1,1 MeV mogą wywołać rozszczepienie ^{238}U (zob. rys. 8). Prowadzi to do zwiększenia liczby neutronów prędkich o tzw. współczynnik efektu prędkiego ε, który z definicji jest większy lub co najmniej równy jedności. Tak więc całkowita liczba neutronów prędkich pochodzących od S neutronów termicznych jest równa $\varepsilon\eta fS$.

Neutrony prędkie są spowalniane w moderatorze i mogą zostać pochłonięte w ^{238}U w zakresie energii rezonansowych. Jeśli przez p oznaczyć prawdopodobieństwo uniknięcia wychwytu rezonansowego przez neutrony o energiach wyższych niż energia termiczna, to ostatecznie całkowita liczba neutronów w pokoleniu następnym, pochodząca od S neutronów w pokoleniu poprzednim, równać się będzie $p\varepsilon\eta fS$.

Iloczyn czterech czynników $k = p\varepsilon\eta f$ nazywa się współczynnikiem mnożenia w układzie nieskończonym. Warunkiem tego, aby reakcja łańcuchowa przebiegała w sposób stacjonarny, jest stała liczba neutronów w kolejnych pokoleniach, czyli $k = 1$. Wtedy mówi się, że rozpatrywany układ jest krytyczny. Gdy $k < 1$, to w kolejnych pokoleniach liczba neutronów maleje i reakcja łańcuchowa zanika, a układ nazywany jest układem podkrytycznym. Natomiast gdy $k > 1$, to liczba neutronów rośnie, a układ jest nadkrytyczny. Układ krytyczny może działać w zasadzie przy dowolnej gęstości neutronów, a co za tym idzie — przy dowolnej mocy.

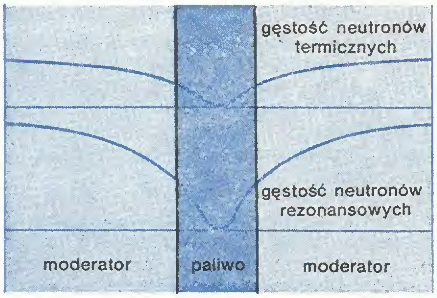
współczynnik mnożenia

układ krytyczny, podkrytyczny i nadkrytyczny

W praktyce każdy układ fizyczny jest skończony i neutrony mogą uciekać poza jego granicę. Prawdopodobieństwo uniknięcia tej ucieczki oznacza się zwykle przez P i do czterech czynników $p\varepsilon\eta f$ trzeba dołączyć piąty; otrzymuje się w ten sposób tzw. efektywny współczynnik mnożenia w układzie skończonym $k_{ef} = Pp\varepsilon\eta f = Pk$. Tak jak poprzednio — układ skończony jest podkrytyczny przy $k_{ef} < 1$, krytyczny przy $k_{ef} = 1$ i nadkrytyczny przy $k_{ef} > 1$.

Dotychczas rozpatrywany układ składający się z jednorodnej mieszaniny paliwa i moderatora byłby z praktycznego punktu widzenia zupełnie nieprzydatny. Zwykle w reaktorze paliwo rozmieszcza się w moderatorze w postaci regularnej siatki elementów paliwowych. Ponieważ w paliwie pochłanianie neutronów zarówno w zakresie energii termicznych jak

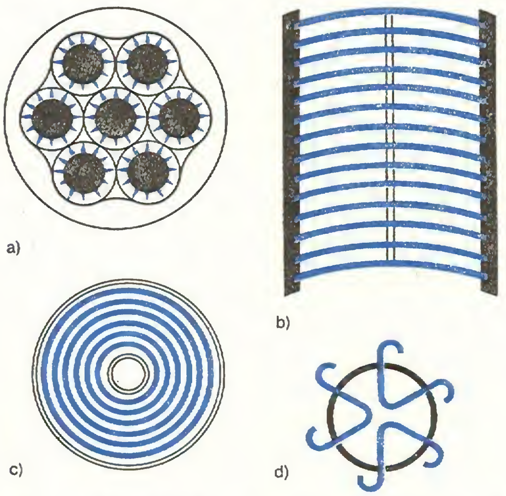
elementy paliwowe



Rys. 11. Rozkład gęstości neutronów termicznych i rezonansowych w reaktorze niejednorodnym

i rezonansowych jest znacznie silniejsze niż w moderatorze, to i gęstość neutronów jest mniejsza (rys. 11). Powoduje to zmniejszenie prawdopodobieństwa pochłonięcia neutronu w paliwie, co pociąga za sobą zmniejszenie współczynnika wykorzystania cieplnego f i zwiększenie prawdopodobieństwa uniknięcia wychwytu rezonansowego p. Ostatecznie jednak iloczyn pf w reaktorze niejednorodnym o odpowiednio dobranym stosunku ilości paliwa i moderatora może być znacznie większy niż w odpowiednim układzie jednorodnym.

Dążenie do najkorzystniejszego bilansu neutronów w reaktorze nie jest główną przyczyną stosowania elementów paliwowych i struktury niejednorodnej. Po pierwsze, większość nuklidów powstających w procesie rozszczepienia jest silnie radioaktywna i trzeba koniecznie izolować je od otoczenia (najwygodniejszym rozwiązaniem jest paliwo otoczone szczelną



Rys. 12. Przekroje elementów paliwowych: a) element wieloprętowy; b) element płytkowy; c) element rurowy; d) element z płytek profilowanych

koszulką metalową). Po drugie, w procesie rozszczepienia w paliwie wydziela się ciepło i trzeba je odprowadzać na zewnątrz reaktora. W tym celu przez reaktor przepuszcza się chłodziwo, które omywa elementy paliwowe, odbiera od nich ciepło i oddaje je poza reaktorem. Aby uzyskać jak najlepsze warunki odbioru ciepła stosuje się elementy paliwowe o bardzo silnie rozwiniętej powierzchni (na rys. 12 pokazano przekroje takich elementów).

Stany nieustalone reaktora

Omówione już zostały warunki, w których można zrealizować stacjonarną reakcję łańcuchową rozszczepienia w układzie krytycznym. Zrozumiałą jest jednak rzeczą, iż w praktyce reaktor jądrowy często znajduje się w stanie podkrytycznym lub nadkrytycznym, choćby w związku z koniecznością jego uruchamiania i zatrzymywania, a więc trzeba zmieniać moc reaktora.

Moc reaktora, w którym efektywny współczynnik mnożenia wynosi k_{ef}, a średni czas życia jednego pokolenia neutronów λ, w przybliżeniu rośnie lub maleje wykładniczo zgodnie z wzorem

$$M(t) = M_0 e^{t(k_{ef}-1)/\lambda},$$

gdzie M_0 oznacza początkową wartość mocy. Moc reaktora maleje lub rośnie tym szybciej, im bardziej k_{ef} różni się od jedności, dlatego też różnicę $\varrho = k_{ef}-1$ przyjęto nazywać reaktywnością. Im większa jest reaktywność danego reaktora (dodatnia lub ujemna), tym gwałtowniejsze zmiany mocy w nim zachodzą.

reaktywność

Czas życia neutronów (przy pominięciu neutronów opóźnionych) w reaktorze, λ (drugi parametr, który określa szybkość zmian mocy reaktora), zawarty jest w granicach 10^{-7}–10^{-3} s. Przy takich wartościach nawet niewielkie zmiany reaktywności powodowałyby tak gwałtowny spadek lub wzrost mocy reaktora, że w praktyce niemożliwa byłaby bezpieczna jego eksploatacja. Udział neutronów opóźnionych zwiększa jednak czas życia neutronów o kilka rzędów, tak że λ wynosi ok. 0,1 s, a to już umożliwia regulację reaktora.

czas życia neutronów w reaktorze

Wzór na moc reaktora obowiązuje jedynie wtedy, gdy reaktywność $\varrho < 0{,}007$, tzn. jest mniejsza od udziału neutronów opóźnionych. Po przekroczeniu przez reaktywność tej granicy, neutrony opóźnione przestają już praktycznie wpływać na przebieg reakcji łańcuchowej, a o zachowaniu się reaktora decydują jedynie neutrony natychmiastowe o bardzo krótkim czasie życia. Tak więc eksploatację reaktora trzeba prowadzić tak, aby nigdy reaktywność nie zbliżyła się do granicy 0,007.

Zasadniczym czynnikiem regulującym pracę reaktora są sprzężenia temperaturowe, które polegają na tym, że wzrost mocy powoduje wzrost temperatury, ten z kolei wywołuje zmianę przekrojów czynnych, a co za tym idzie zmianę reaktywności. W większości reaktorów występują ujemne sprzężenia temperaturowe i wzrost gęstości neutronów powoduje spadek reaktywności. Ujemne sprzężenie temperaturowe stanowi oczywiście naturalny czynnik stabilizujący pracę reaktora.

sprzężenie temperaturowe

W czasie normalnej eksploatacji reaktora zachodzą powolne zmiany parametrów jądrowych, związane z tworzeniem się produktów rozszczepienia oraz zmianą składu paliwa. Na ogół zmiany te powodują stopniowy spadek reaktywności. Dlatego też zwykle okres eksploatacji paliwa w reaktorze zależy od czasu, w ciągu którego reaktor staje się podkrytyczny. Niekiedy na okres eksploatacji wpływają również zmiany strukturalne zachodzące w paliwie.

Produkty rozszczepienia często pochłaniają neutrony, wywierają więc ujemny wpływ na reaktywność i z tego względu są nazywane truciznami. Rozróżnia się dwie grupy trucizn: stałe i przejściowe. Pierwsza grupa obejmuje kilkadziesiąt nuklidów trwałych i długożyjących, z których najważniejszy jest izotop samaru ^{149}Sm o przekroju czynnym na pochłanianie neutronów termicznych równym $5 \cdot 10^4$ b. W grupie trucizn przejściowych istotną rolę odgrywa właściwie jedynie izotop ksenonu ^{135}Xe, który odznacza się niezwykle dużym przekrojem czynnym na pochłanianie neutronów termicznych ($2 \cdot 10^6$ b). Czas życia ^{135}Xe wynosi 19,2 h, natomiast wydajność w procesie rozszczepienia ok. 6,3%. Z tego pewna część powstaje

trucizny jądrowe

bezpośrednio z rozszczepienia, reszta zaś w wyniku reakcji rozpadu:

$$^{235}_{92}U + ^{1}_{0}n \xrightarrow{\text{rozszczepienie}} ^{135}_{52}Te \xrightarrow[243\,s]{\beta} ^{135}_{53}J \xrightarrow[13,8\,h]{\beta}$$

$$^{135}_{54}Xe \xrightarrow[19,2\,h]{\beta} ^{135}_{55}Cs.$$

Czas życia ^{135}Cs jest rzędu $3 \cdot 10^6$ lat, tak że z praktycznego punktu widzenia jest on nuklidem trwałym.

jama jodowa

Fakt, że ^{135}Xe powstaje głównie z rozpadu izotopu jodu ^{135}J, powoduje bardzo ważne z punktu widzenia eksploatacji reaktorów zjawisko tzw. jamy jodowej. Polega ono na tym, że po upływie określonego czasu nieprzerwanej pracy reaktora przy ustalonej mocy ilość ksenonu osiąga wartość nasycenia w wyniku ustalenia się równowagi pomiędzy powstawaniem jąder ^{135}Xe z rozpadu ^{135}J a ich przemianą w ^{136}Xe wywołaną pochłanianiem neutronów. Wzrost ilości ksenonu w reaktorze zmniejsza reaktywność, czyli wywołuje zatrucie. Nagłe wyłączenie reaktora powoduje wzrost zatrucia ksenonem, który w dalszym ciągu powstaje z rozpadu jodu, a nie jest usuwany z reaktora przez pochłanianie neutronów. Dopiero po upływie ok. 10 h po wyłączeniu zatrucie ksenonem zaczyna się zmniejszać wskutek rozpadu zarówno jodu jak i ksenonu. Wielkość jamy jodowej zależy od mocy reaktora, a samo zjawisko występuje nie tylko przy wyłączeniu, ale także przy zmianach mocy reaktora.

Zmiany składu paliwa w czasie eksploatacji reaktora polegają przede wszystkim na wypalaniu się nuklidów rozszczepialnych. Gdy paliwem jest uran naturalny lub wzbogacony izotopem ^{235}U, to stopniowo zmniejsza się koncentracja tego izotopu. Z drugiej strony pochłanianie neutronów w ^{238}U prowadzi do powstawania plutonu ^{239}Pu

produkcja plutonu

$$^{238}_{92}U + ^{1}_{0}n \rightarrow ^{239}_{92}U \xrightarrow[48\,min]{\beta} ^{239}_{93}Np \xrightarrow[4,7\,dnia]{\beta} ^{239}_{94}Pu.$$

Okazuje się, że jądro ^{238}Pu ma właściwości bardzo zbliżone do właściwości ^{235}U. Ma ono bardzo duży przekrój czynny na rozszczepienie pod wpływem neutronów termicznych, a średnia liczba neutronów rozszczepieniowych na 1 akt rozszczepienia wynosi 2,91. Jest więc pluton doskonałym paliwem reaktorowym.

współczynnik przemiany paliwa

Zjawisko powstawania nuklidów rozszczepialnych z materiałów rodnych, takich jak np. ^{238}U, charakteryzuje współczynnik przemiany paliwa określony jako stosunek liczby jąder uzyskanych nuklidów rozszczepialnych do liczby wypalonych jąder paliwa przy końcu okresu eksploatacji. W reaktorach termicznych z uranem naturalnym współczynnik przemiany sięgać może 0,7. Jeśli współczynnik przemiany jest większy od jedności, to zachodzi tzw. powielanie paliwa. W praktyce powielanie paliwa udało się dotychczas zrealizować jedynie w reaktorach prędkich.

powielanie paliwa

Oprócz cyklu uranowo-plutonowego istnieje również cykl torowo-uranowy, w którym nuklidem rodnym jest ^{232}Th, a powstającym z niego nuklidem rozszczepialnym pod wpływem neutronów termicznych jest ^{233}U. Cykl torowo-uranowy jest bardzo atrakcyjny, gdyż istniejc możliwość powielania paliwa również w reaktorach termicznych.

Omówione wyżej zmiany izotopowe nie wyczerpują wszystkich zjawisk zachodzących w paliwie w czasie eksploatacji reaktora. Paliwo staje się mieszaniną wielu nuklidów z grupy aktynowców. Część z tych nuklidów ma właściwości materiałów rodnych, a część — materiałów rozszczepialnych.

Sterowanie reaktorem

Jak każde urządzenie wytwarzające energię, reaktor musi być wyposażony w układ sterowania, ponieważ stabilizujące działanie ujemnego sprzężenia temperaturowego jest niewystarczające, a poza tym zachodzi konieczność uruchamiania i wyłączania, a także zmiany mocy reaktora.

pręty sterownicze

Metody sterowania polegają na zmianie objętości paliwa, moderatora, reflektora lub substancji pochłaniającej neutrony. W reaktorach termicznych najczęściej stosuje się ruchome pręty sterownicze wykonane z materiałów silnie pochłaniających neutrony termiczne, takich jak kadm lub bor. Można je podzielić na trzy grupy:

1) Pręty bezpieczeństwa, wprowadzające do układu dużą reaktywność ujemną, służące do gwałtownego wyłączania reaktora. Są one połączone z mechanizmem napędowym — zwykle elektrycznym, lub też hydraulicznym albo pneumatycznym. Pręty bezpieczeństwa w wypadku awarii spadają do rdzenia pod wpływem siły ciężkości. Aby przyspieszyć ich spadanie mogą być zastosowane wyrzutnie stalowe lub ładunki wybuchowe.

2) Pręty kompensacyjne, służące do zmniejszania reaktywności po wstępnym okresie eksploatacji paliwa.

3) Pręty regulacyjne, powodujące niewielkie stosunkowo zmiany reaktywności, służące do kompensacji przypadkowych odchyleń mocy reaktora od stanu równowagi, a także do uruchamiania i zatrzymywania reaktora (il. 43, tabl. 12).

Często w reaktorach energetycznych zamiast prętów kompensacyjnych stosuje się trucizny, które wypalają się w trakcie eksploatacji. Zjawisko pochłaniania neutronów w prętach sterowniczych stosuje się również do wytwarzania nuklidów promieniotwórczych. W reaktorach prędkich, ze względu na małe pochłanianie neutronów o dużych energiach, zwykle do sterowania reaktorem stosuje się ruchome elementy paliwowe lub ruchome części reflektora.

Teoria transportu neutronów i obliczanie reaktorów

Wyznaczenie wielkości określających efektywny współczynnik mnożenia oraz innych wielkości istotnych z punktu widzenia eksploatacji reaktora wymaga znajomości przekrojów czynnych i rozkładu neutronów w rozpatrywanym układzie. Pomiary przekrojów czynnych należą do dziedziny fizyki jądrowej, natomiast do fizyki reaktorowej należą pomiary lub teoretyczne przewidywanie rozkładu neutronów.

teoria transportu neutronów

Opisu zachowania się swobodnych neutronów w ośrodku materialnym dostarcza dział fizyki statystycznej zwany teorią transportu neutronów. Podstawowym równaniem tej teorii jest liniowe równanie Boltzmanna, które swą liniowość zawdzięcza założeniu, że zderzenia pomiędzy neutronami są niezwykle mało prawdopodobne i można je pominąć. Poza tym zakłada się, że neutrony poruszają się zgodnie z zasadami mechaniki klasycznej, a efekty relatywistyczne i kwantowe są pomijalne.

równanie Boltzmanna

Rozkład neutronów opisuje funkcja $N(\vec{r}, \vec{v}, t)$, w której argumentami są: wektor położenia neutronu $\vec{r}$, wektor prędkości $\vec{v}$ i czas t. Jej sens fizyczny jest taki, że $N(\vec{r}, \vec{v}, t)\Delta\vec{r}\Delta\vec{v}$ oznacza średnią liczbę neutronów znajdujących się w chwili t w elemencie objętości $\Delta\vec{r}$ wokół punktu $\vec{r}$ o prędkościach zawartych w „elemencie" $\Delta\vec{v}$ wokół prędkości $\vec{v}$. Równanie Boltzmanna wyraża bilans neutronów wchodzących i wychodzących z elementu $\Delta\vec{r}\Delta\vec{v}$ i jest równaniem różniczkowo-całkowym.

Stany statyczne reaktorów

W rzeczywistych sytuacjach fizycznych rozwiązywanie równania Boltzmanna jest, nawet przy użyciu komputerów, bardzo uciążliwe; a niekiedy, przy istniejących możliwościach obliczeniowych, niewyko-

przybliżenie dyfuzyjne i wielogrupowe

nalne. Z tego względu w obliczeniach reaktorowych stosuje się z reguły metody przybliżone. Do najważniejszych z nich należą: przybliżenie dyfuzyjne, w którym zakłada się, że funkcja $N(\vec{r}, \vec{v}, t)$ zależy tylko od wartości bezwzględnej wektora $\vec{v}$, tzn. tylko od energii neutronu, a nie od kierunku jego prędkości oraz przybliżenie wielogrupowe, w którym przyjmuje się, że energia neutronów może przybierać tylko wartości dyskretne.

Obliczenia reaktora jądrowego są wykonywane zwykle w kilku etapach za pomocą programów numerycznych na maszynę cyfrową, czyli tzw. kodów reaktorowych. Pierwszy etap stanowią obliczenia komórkowe. Polegają one na znajdowaniu uśrednionych parametrów charakteryzujących rozkład neutronów w komórce elementarnej reaktora niejednorodnego. Przez komórkę elementarną reaktora rozumie się element paliwowy, tzn. paliwo wraz z koszulką, otoczone warstwą chłodziwa i moderatora. Dla uproszczenia obliczeń zwykle zakłada się, że komórka elementarna ma symetrię cylindryczną, a reaktor zastępuje się układem nieskończonym składającym się z takich komórek.

obliczenia komórkowe

Obliczenia komórkowe są prowadzone w przybliżeniu wielogrupowym dla kilkudziesięciu grup energetycznych. Na ogół nie stosuje się przybliżenia dyfuzyjnego, gdyż w komórce, która ma stosunkowo niewielki rozmiar i zawiera materiały silnie pochłaniające neutrony termiczne lub rezonansowe, prowadziłoby to do zbyt dużych błędów. W wyniku obliczeń komórkowych otrzymuje się tzw. stałe wielogrupowe, charakteryzujące energetyczny rozkład neutronów uśredniony na cały obszar komórki.

obliczenia widmowe

Drugi etap, tzw. obliczenia widmowe, polega na redukcji kilkudziesięciu grup energetycznych do kilku (zwykle nie więcej niż czterech) przez odpowiednie procedury uśredniania. W widmowych kodach reaktorowych uwzględnia się w sposób przybliżony to, że rozkład neutronów zmienia się od komórki do komórki.

Na podstawie wyników obliczeń komórkowych i widmowych można określić czynniki wchodzące w skład wzoru na współczynnik mnożenia k. Tak więc obliczenia komórkowe i widmowe odpowiadają obliczeniu współczynnika mnożenia nieskończonego układu niejednorodnego.

obliczenia krytyczne

Trzeci etap stanowią obliczenia krytyczne. Z reguły polegają one na rozwiązywaniu układu równań dyfuzyjnych dla kilku grup energetycznych przy użyciu stałych znalezionych w poprzednim etapie obliczeń. W wyniku otrzymuje się efektywny współczynnik mnożenia k_{ef}. Tak więc obliczenia krytyczne w istocie rzeczy sprowadzają się do znalezienia prawdopodobieństwa uniknięcia ucieczki neutronów z układu.

zestaw wykładniczy i zestaw krytyczny

Obliczenia komórkowe, widmowe i krytyczne można sprawdzać doświadczalnie na zestawach wykładniczych lub krytycznych. Zestawy takie zbudowane są z takich samych materiałów co reaktor i mają tę samą strukturę komórki podstawowej. Zestaw wykładniczy ma zbyt małe rozmiary, aby osiągnąć stan krytyczny. Jego nazwa wynika stąd, że rozkład neutronów pochodzących z umieszczonego w nim źródła ma charakter wykładniczy. Zestaw krytyczny jest właściwie reaktorem o bardzo małej mocy, w którym chłodzenie zapewnia konwekcja naturalna powietrza lub wody. W związku z tym koszt budowy zestawu krytycznego jest znacznie niższy niż koszt budowy reaktora o pełnej mocy, szczególnie reaktora energetycznego. Poza tym zestaw krytyczny może być łatwo przebudowany, co pozwala na sprawdzanie różnorodnych rozwiązań konstrukcyjnych reaktora.

obliczenia eksploatacyjne

Dalszym krokiem w obliczeniach reaktorowych są obliczenia eksploatacyjne, których celem jest określenie, jak zmieniają się parametry reaktora, a przede wszystkim k_{ef}, w miarę eksploatacji określonej porcji paliwa. Zwykle obliczenia eksploatacyjne polegają na wykonywaniu obliczeń krytycznych w kolejnych chwilach, połączonym z rozwiązywaniem równań różniczkowych opisujących zmiany składu paliwa i powstawanie trucizn. Dzięki obliczeniom eksploatacyjnym można wybrać optymalny sposób wymiany paliwa i przedłużyć czas jego przebywania w reaktorze, a w ten sposób obniżyć koszt eksploatacji. Ma to szczególne znaczenie w reaktorach energetycznych.

Dotychczas omawiane obliczenia dotyczyły stanów statycznych reaktora lub bardzo powolnych zmian zachodzących w czasie jego eksploatacji. Przewidywaniem zachowania się reaktora w stanach nieustalonych zajmuje się tzw. dynamika reaktorów.

Dynamika reaktorów

Jak już wspomniano, podstawowe równanie Boltzmanna opisujące zachowanie się neutronów w reaktorze o niewielkiej mocy jest liniowe, gdyż nie zawiera członu opisującego zderzenia neutronów między sobą. Dotyczy to również wszelkich równań przybliżonych, opisujących stany nieustalone reaktora o małej mocy.

W reaktorze o większej mocy istotną rolę odgrywa sprzężenie temperaturowe i w równaniach opisujących stany nieustalone reaktora pojawiają się człony nieliniowe. Poza tym do równań opisujących gęstość neutronów i gęstość jąder, z których powstają neutrony opóźnione, trzeba dołączyć równania przewodnictwa ciepła i przepływu chłodziwa.

Najważniejszym zadaniem dynamiki jest ustalenie, czy reaktor jest stabilny w warunkach normalnej pracy i w stanach awaryjnych, od tego bowiem zależy bezpieczna jego eksploatacja. Sprowadza się to do badania stabilności równań nieliniowych, co na ogół jest problemem bardzo skomplikowanym z matematycznego punktu widzenia. Właściwe obliczenia dynamiczne polegają na rozwiązywaniu równań dynamiki reaktora przy zadanych warunkach początkowych. W pracach projektowych stosuje się kody dynamiczne, które uwzględniają zależność przestrzenną gęstości neutronów i in. wielkości wchodzących w skład równań dynamiki. Zwykle są one bardzo skomplikowane i obliczenia zajmują dużo czasu nawet przy użyciu szybkich komputerów.

kody dynamiczne

Podczas eksploatacji reaktora zachodzi również konieczność przewidywania jego stanów przejściowych, ale ze względu na bezpieczeństwo czas wykonywania obliczeń musi być stosunkowo krótki. W tym wypadku więc stosuje się znacznie prostsze kody dynamiczne, oparte na ogół na tzw. modelu jednopunktowym, w którym brak zależności przestrzennych. Aby uniknąć poważniejszych błędów, parametry wchodzące w skład równań dynamiki ustala się na podstawie ich pomiarów lub przez porównanie wyników obliczeń ze zmierzonym przebiegiem gęstości neutronów w zależności od czasu. W ten sposób można wybrać optymalną strategię sterowania łańcuchową reakcją rozszczepienia w reaktorze, zarówno z punktu widzenia bezpieczeństwa, jak i ekonomii.

Budowa i klasyfikacja reaktorów jądrowych

W różnych krajach zbudowano dotąd kilka tysięcy reaktorów, z czego kilkaset służy do produkcji energii elektrycznej (il. 41-45, tabl. 12). Ogromna większość pracujących reaktorów to reaktory termiczne, czyli takie, w których większość rozszczepień zachodzi pod wpływem neutronów termicznych. Istnieją również reaktory prędkie (nie zawierające moderatorów), w których rozszczepienia są wywoływane głównie przez neutrony o dużych energiach (średnia energia neutronów wynosi w nich ok. 100 keV). Zasadniczą zaletą reaktorów prędkich jest możliwość powielania paliwa

reaktory termiczne i prędkie

w cyklu uranowo-plutonowym. Z tego względu uważa się, że przyszłość energetyki opartej na reaktorach jądrowych leży w prędkich reaktorach powielających. Warto dodać, że bomba jądrowa oparta na reakcji rozszczepienia jest właściwie reaktorem prędkim, w którym reaktywność znacznie przekracza granicę 0,007 określoną przez udział neutronów opóźnionych i reakcja łańcuchowa rozwija się w sposób wybuchowy.

reaktor wodny

Rozróżnia się kilka typowych rozwiązań konstrukcyjnych reaktorów. Przede wszystkim reaktory wodne o małych mocach są zwykle budowane w dużych zbiornikach wodnych, często z otwartym lustrem wody (reaktory basenowe). Reaktory wodne o większych mocach są umieszczone z reguły w zbiorniku ciśnieniowym. Znane są również rozwiązania konstrukcyjne, w których każdy element paliwowy wraz z opływającą go warstwą wody chłodzącej jest zawarty w rurze stalowej i utrzymywany pod ciśnieniem. W reaktorach z moderatorem stałym, jak grafit lub beryl, sam moderator stanowi materiał konstrukcyjny rdzenia. Cały reaktor może być także umieszczony w zbiorniku ciśnieniowym.

rdzeń

reflektor

Najkosztowniejszym składnikiem reaktora jądrowego jest paliwo, natomiast znacznie tańszy jest moderator. Z tego względu reaktor jest zwykle podzielony na dwie strefy. Pierwsza — tzw. rdzeń, zawiera paliwo, moderator, chłodziwo i materiały konstrukcyjne oraz pręty regulacyjne, kompensacyjne i bezpieczeństwa. Druga — tzw. reflektor, jest zbudowana z moderatora i otacza rdzeń, a służy do zatrzymywania części neutronów uciekających z rdzenia. W wielu rozwiązaniach konstrukcyjnych rdzeń i reflektor znajdują się w szczelnych zbiornikach stalowych i otoczone są specjalnymi osłonami. Zbiornik ciśnieniowy utrzymuje odpowiednie ciśnienie w reaktorze i zapewnia znacznie większe bezpieczeństwo na wypadek awarii. W reaktorach wodnych ciśnienie wewnątrz zbiornika dochodzi do 14, a w gazowych — do 4 megapaskali. Osłony reaktora zmniejszają intensywność promieniowania w otoczeniu do wartości dopuszczalnych ze względu na zagrożenie obsługi. Są one również stosowane do ochrony niektórych urządzeń pomocniczych i aparatury przed zniszczeniem lub promieniotwórczością wzbudzoną.

osłona

Ponieważ największe pochłanianie neutronów w materii zachodzi w zakresie energii termicznych, osłona reaktorowa powinna zawierać moderator spowalniający neutrony prędkie wychodzące z reaktora oraz materiał odznaczający się znacznym przekrojem czynnym na pochłanianie neutronów termicznych. Poza tym powinna efektywnie pochłaniać promieniowanie γ wychodzące z reaktora, jak również powstające w osłonie. Promieniowanie α lub β ma bardzo niewielki zasięg i praktycznie nie wydostaje się poza pierwszą warstwę osłony.

Największa ilość ciepła powstającego w osłonie wydziela się w warstwie przylegającej do reaktora. Z tego względu warstwa ta, zwykle oddzielona od reszty osłony, jest chłodzona wodą lub powietrzem i nazywa się osłoną termiczną. Pozostała część pełni rolę osłony biologicznej. Osłona termiczna jest najczęściej wykonana z płyt stalowych z dodatkiem boru, natomiast biologiczna z ciężkich betonów, składających się z cementu, drobnego złomu stalowego oraz rudy barytowej, limonitowej lub magnetytowej.

paliwo

Reaktory mogą różnić się od siebie rodzajem paliwa (uran naturalny, uran wzbogacony izotopem ^{235}U lub nuklidem ^{239}Pu), postacią, w jakiej paliwo występuje w reaktorze (paliwo metaliczne, węgliki lub tlenki uranu itp.) oraz kształtem elementów paliwowych. Początkowo stosowano głównie uran naturalny, obecnie w większości reaktorów paliwem jest uran wzbogacony izotopem ^{235}U, przy czym wzbogacenie waha się od 1% do 93%. Dzięki temu uzyskuje się znacznie lepsze parametry krytyczne reaktora i lepszy rozkład neutronów. Obok uranu stosuje się również ^{239}Pu otrzymywany z przemiany jądrowej ^{238}U. W przyszłości coraz większą rolę będzie odgrywać paliwo plutonowe, a także ^{233}U. Paliwo uranowe lub plutonowe stosuje się w postaci metalicznej lub ceramicznej (tlenki lub węgliki), a także w postaci roztworu metalu w ceramice, a elementy paliwowe są koszulkowane stalą nierdzewną, cyrkonem lub aluminium.

moderator

Moderator jest bardzo istotnym elementem reaktora termicznego i często używanym kryterium klasyfikacyjnym. Rozróżnia się więc następujące zasadnicze typy reaktorów termicznych: wodne, ciężkowodne i grafitowe. W niektórych rozwiązaniach stosuje się również beryl w połączeniu z wodą, spełniającą jednocześnie rolę moderatora i chłodziwa. Istnieją również reaktory, w których moderatorem i chłodziwem są ciekłe związki organiczne, głównie polifenyle; takie reaktory nie mają jednak dużego znaczenia praktycznego. Najpowszechniej używanym moderatorem jest zwykła woda, która ma doskonałe właściwości jądrowe i termiczne, a jest przy tym bardzo tania.

chłodziwo

Chłodziwa reaktorowe można podzielić na 3 grupy: gazy, ciecze niemetaliczne i ciecze metaliczne, bardzo różniące się pomiędzy sobą właściwościami fizycznymi, które decydują o efektywności przejmowania i przenoszenia ciepła. Ważną rolę w wypadku chłodziw odgrywają również właściwości jądrowe, jak np. pochłanianie neutronów czy promieniotwórczość wzbudzona oraz właściwości korozyjne. Do najczęściej stosowanych chłodziw w reaktorach termicznych należy zwykła woda. Używa się również dwutlenku węgla, a także helu. W reaktorach prędkich chłodziwem jest zwykle ciekły sód lub jego stopy, ale również przeprowadzane są próby chłodzenia gazem dysocjującym, np. N_2O_4, który dysocjuje na NO_2 pochłaniając ciepło z reaktora i redysocjuje poza reaktorem, będąc w ten sposób bardzo efektywnym nośnikiem ciepła. Wadą sodu jest znaczna aktywność chemiczna zarówno w stosunku do powietrza jak i wody, a także silne właściwości korozyjne. Powoduje to bardzo poważne trudności konstrukcyjne i eksploatacyjne w reaktorach chłodzonych sodem.

W reaktorach termicznych chłodziwo spełniać również może funkcję moderatora. Z tego względu najczęściej jako chłodziwo stosowana jest zwykła woda, dwutlenek węgla, hel i niekiedy związki organiczne. Woda jako chłodziwo może znajdować się pod ciśnieniem atmosferycznym lub podwyższonym. W tym ostatnim wypadku mówi się o reaktorach wodnych ciśnieniowych. Niekiedy dopuszcza się do wrzenia, tak że chłodziwem w tzw. reaktorach wodnych wrzących jest mieszanina wody z parą.

Chłodzenie dwutlenkiem węgla jest stosowane w reaktorach grafitowych z uranem naturalnym, jednak ten typ chłodzenia nie ma wielkiej przyszłości, podobnie jak chłodzenie związkami organicznymi. W nowoczesnych reaktorach grafitowych dwutlenek węgla jest zastąpiony helem, co pozwala znacznie zwiększyć temperaturę chłodziwa na wyjściu reaktora (do 1000°C).

Zastosowanie reaktorów jądrowych

Z uwagi na przeznaczenie reaktory można podzielić w zasadzie na dwie kategorie: badawcze i energetyczne, przy czym te ostatnie obejmują reaktory stacjonarne i napędowe.

reaktory badawcze

Reaktory badawcze, do których zaliczyć należy również zestawy krytyczne, służą jako narzędzia badań w zakresie fizyki reaktorowej, chemii radiacyjnej, radiochemii, właściwości materiałów, energetyki jądrowej itp. oraz jako źródła promieniowania (głównie neutronów i promieniowania γ) do produkcji radioizotopów (nuklidów promieniotwórczych), stosowanych w medycynie, biologii, rolnictwie, przemyśle

itd. Niektóre reaktory badawcze z uranem naturalnym lub niskowzbogaconym o dużym współczynniku przemiany paliwa stanowią również źródło nowego paliwa plutonowego.

Reaktory badawcze są wyposażone w kanały doświadczalne, dochodzące do reflektora i rdzenia reaktora. Dąży się do tego, aby strumień neutronów i objętość kanałów doświadczalnych były jak największe. Szczególnie duży strumień neutronów jest niezbędny w reaktorach przeznaczonych do badań materiałowych, w których sprawdza się w praktyce zachowanie się elementów paliwowych projektowanych reaktorów energetycznych.

W Polsce w Instytucie Badań Jądrowych w Świerku znajduje się kilka zestawów krytycznych i dwa reaktory badawcze (il. 42, 44; 45, tabl. 12). Pierwszy polski

reaktor EWA

reaktor EWA (Eksperymentalny Wodny Atomowy) został uruchomiony w 1958 r. Jego paliwem jest uran wzbogacony w izotop ^{235}U do 36%. Jako moderator i reflektor służy zwykła woda i częściowo beryl, chłodziwem jest zwykła woda. Reaktor EWA jest umieszczony w zbiorniku betonowym, który służy jako osłona biologiczna dla osób pracujących w hali reaktora. Obecna moc reaktora wynosi 8 MW (początkowa 2 MW), a strumień neutronów termicznych w kanałach doświadczalnych dochodzi do $2 \cdot 10^{14}$ n/cm^2s. Reaktor służy do produkcji radioizotopów i do badań w zakresie fizyki i chemii.

reaktor MARIA

Drugi polski reaktor MARIA (nazwany tak na cześć Marii Skłodowskiej-Curie) osiągnął stan krytyczny w grudniu 1974 r. Jego moc nominalna wynosi 30 MW, a strumień maksymalny $5 \cdot 10^{14}$ n/cm^2 s. Paliwem jest uran wzbogacony w izotop ^{235}U do 80%. Jako moderator służy beryl, reflektor jest zbudowany z grafitu, chłodziwem i zarazem moderatorem jest woda pod ciśnieniem 1,7 MPa. MARIA jest reaktorem uniwersalnym, przeznaczonym głównie do badania materiałów i do produkcji radioizotopów.

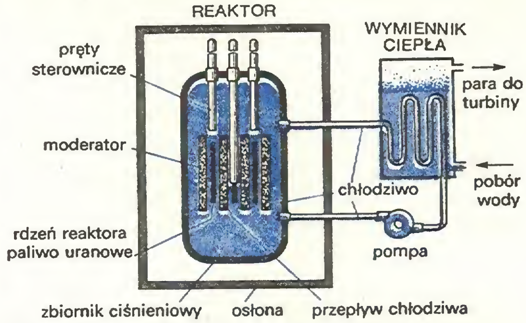
raaktory napędowe

Reaktory energetyczne są przeznaczone do produkcji energii w elektrowniach lądowych lub na statkach (reaktory napędowe). Reaktory napędowe nie różnią się w zasadzie od stacjonarnych, poza tym, że konstrukcyjnie są znacznie trudniejsze. Muszą one bowiem odznaczać się małymi rozmiarami, co dotyczy również osłon, i zapewniać bezpieczną eksploatację również w nie sprzyjających warunkach w czasie podróży. Dotychczas reaktory (wyłącznie wodne ciśnieniowe) znalazły zastosowanie do napędu łodzi podwodnych, okrętów wojennych i statków morskich.

Elektrownie jądrowe

Pierwsza elektrownia jądrowa (atomowa) o mocy 5 MW (elektrycznych) została zbudowana w 1954 r. w ZSRR, a już w 1976 r. na świecie pracowało kilkaset elektrowni jądrowych o mocach dochodzących do 1000 MW. Początkowo energia z elektrowni jądrowych była znacznie droższa niż z elektrowni klasycznych (węglowych lub wodnych). W miarę jednak rozwoju techniki reaktorów cena energii jądrowej spada i w 1976 r. była już porównywalna z ceną energii z innych źródeł. Dotychczasowe doświadczenia wskazują, że elektrownie jądrowe są bardzo bezpieczne wbrew dość rozpowszechnionym poglądom, iż każdy reaktor jądrowy stanowi potencjalną bombę. Poza tym elektrownie jądrowe nie powodują tak wielkiego zanieczyszczenia środowiska jak elektrownie węglowe, które wyrzucają ogromne ilości produktów spalania węgla do atmosfery. Zasadniczym problemem w wypadku elektrowni jądrowych jest tzw. zanieczyszczenie termiczne, które polega na wzroście temperatury wód wywołanym odprowadzanym z elektrowni ciepłem, a także gromadzenie się odpadów promieniotwórczych. Wydaje się jednak, że nauka potrafi w przyszłości uporać się z obydwoma problemami.

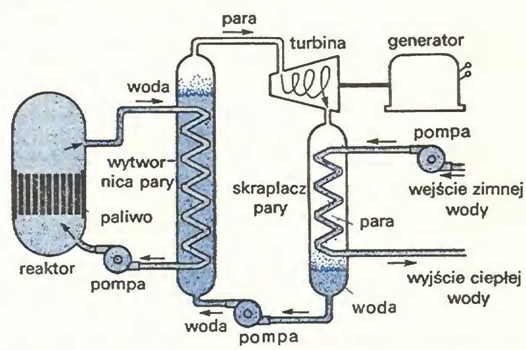
Najważniejszym jednak argumentem przemawiającym za rozwojem energetyki jądrowej jest gwałtowne kurczenie się zasobów paliw klasycznych, co bez udziału energii jądrowej doprowadziłoby już wkrótce do ostrego kryzysu energetycznego.



Rys. 13. Schemat reaktora energetycznego

reaktor energetyczny

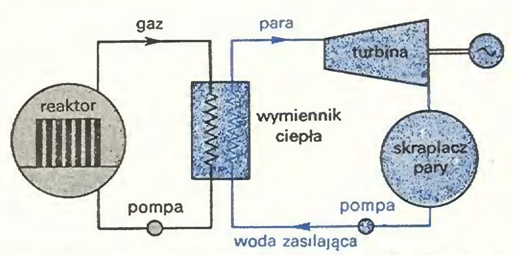
Na rys. 13 przedstawiono schemat typowego reaktora energetycznego stacjonarnego lub napędowego. Chłodziwo opuszczając zbiornik reaktora unosi ze sobą ciepło, które następnie oddane w wymienniku ciepła powoduje tworzenie się pary wodnej używanej do poruszania turbiny i wytwarzania energii elektrycznej odprowadzanej do sieci energetycznej lub, w reaktorach napędowych, używanej do napędu silników.



Rys. 14. Schemat elektrowni jądrowej z reaktorem wodnym ciśnieniowym

reaktory wodne

Najczęściej spotykanym typem reaktora energetycznego jest reaktor wodny ciśnieniowy (rys. 14). Elektrownie z tymi reaktorami pracują już w wielu krajach. W ZSRR i w innych krajach socjalistycznych działają reaktory wodne ciśnieniowe o rozwiązaniu konstrukcyjnym nazwanym WWER. Pierwsza polska elektrownia jądrowa, która ma być zbudowana nad Jeziorem Żarnowieckim w woj. gdańskim będzie wyposażona właśnie w reaktor WWER o mocy 440 WM (elektrycznych). Drugim typem reaktora wodnego jest reaktor wodny wrzący, w którym dopuszcza się do wrzenia wody w rdzeniu reaktora, co pozwala obniżyć ciśnienie w zbiorniku reaktora. Jednakże reaktory wodne wrzące są nieco trudniejsze konstrukcyjnie.



Rys. 15. Schemat elektrowni z reaktorem chłodzonym gazem

reaktory grafitowe

Pierwszymi reaktorami energetycznymi chłodzonymi gazem (dwutlenek węgla) były reaktory w Calder Hall (Wielka Brytania) oddane do eksploatacji w 1956 r. Moderatorem był grafit, a paliwem uran naturalny. Schemat działania reaktora z chłodzeniem gazowym pokazany jest na rys. 15. Obecnie prowadzi się intensywne prace nad ulepszonymi wersjami energetycznych reaktorów grafitowych z chłodzeniem gazowym, tzw. reaktorów wysokotemperaturowych, w których temperatura gazu na wyjściu osiąga 1000°C. Gaz o tej temperaturze pozwala osiągnąć znacznie wyższą sprawność cieplną elektrowni jądrowej. Szczególnie interesującym rozwiązaniem jest reaktor, w którym paliwem jest złoże usypane z kul uranowych pokrytych grafitem, a chłodziwem hel. Prototyp tego reaktora pracuje w Jülich (RFN).

Reaktory wysokotemperaturowe odznaczają się bardzo dużą sprawnością i pozwalają na dobre wykorzystanie paliwa. Dodatkową zaletą jest możliwość użycia tych reaktorów w przemyśle chemicznym.

Ze względu na fakt, że ogromna większość uranu znajdującego się w przyrodzie to ^{238}U, przyszłość energetyki jądrowej leży w prędkich reaktorach powielających. Reaktory prędkie są jednak znacznie trudniejsze zarówno z punktu widzenia konstrukcyjnego jak i eksploatacyjnego i ciągle jeszcze znajdują się w stadium prototypów. Pierwsze duże elektrownie jądrowe z reaktorami prędkimi pojawią się zapewne dopiero po 1990 r.

S. Glasstone *Podstawy techniki reaktorów jądrowych*, Warszawa 1958; S. Glasstone i M. C. Edlund *Podstawy teorii reaktorów jądrowych*, Warszawa 1957; J. Mika i A. Zmysłowski *Energia jądrowa i jej zastosowanie*, Warszawa 1962.

Energia termojądrowa

Lech Jakubowski i Marek Sadowski

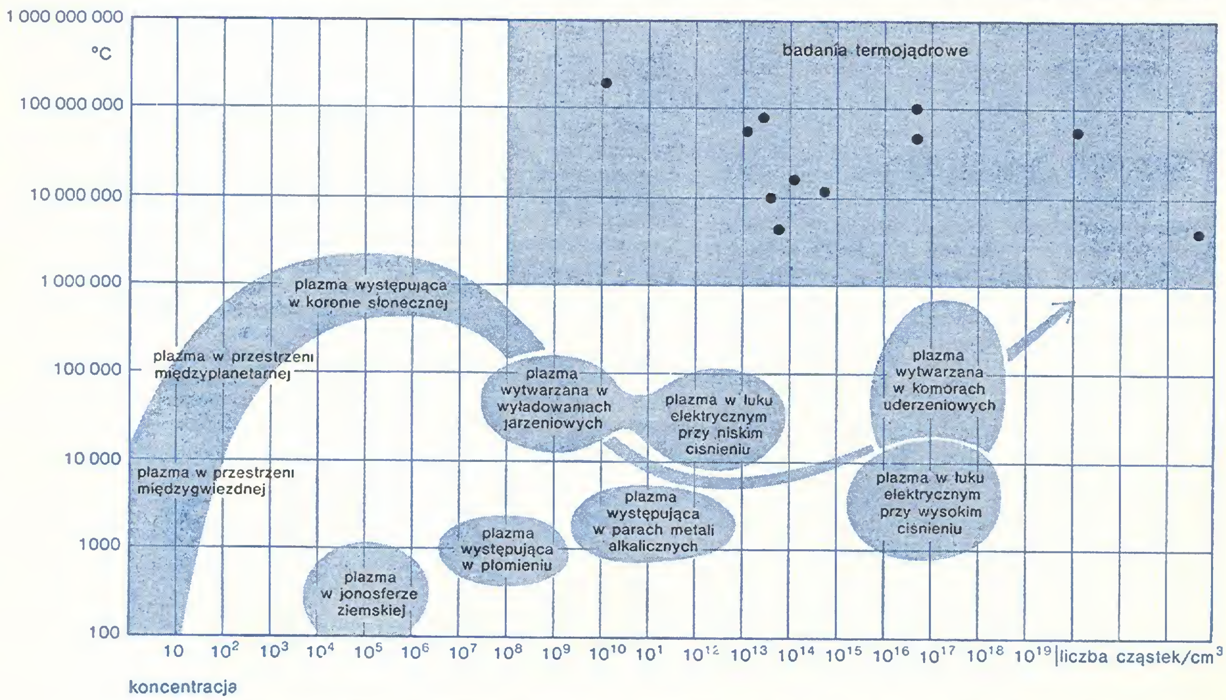
Energia jądrowa może się wyzwalać nie tylko w procesach rozszczepienia ciężkich jąder atomowych, ale również w reakcjach syntezy (tj. łączenia) jąder najlżejszych w jądra cięższe. Aby reakcja syntezy mogła dojść do skutku, jądra muszą pokonać siły wzajemnego odpychania elektrostatycznego i zbliżyć się na odległość rzędu 10^{-12} cm. W temperaturze pokojowej średnia energia kinetyczna ruchu cieplnego jąder atomowych jest zbyt mała, aby tego rodzaju reakcja mogła zachodzić. Natomiast w temperaturze wielu milionów stopni, jaka panuje np. we wnętrzu Słońca i innych gwiazd, energia ruchu cieplnego jest tak duża, że wystarcza do pokonania sił kulombowskich między jądrami, co umożliwia połączenie tych jąder. Procesy łączenia jąder atomowych, które zachodzą na skutek ruchów termicznych w bardzo wysokich temperaturach, nazywamy reakcjami syntezy termojądrowej, a wydzieloną w tych procesach energię — energią termojądrową.

W dostatecznie wysokich temperaturach następuje samorzutna jonizacja materii — atomy rozpadają się na dodatnio naładowane jony i swobodne elektrony, tworząc kwazineutralną mieszaninę — tzw. plazmę. Istotne jest, że dodatnie i ujemne cząstki występują w plazmie w takich proporcjach, że wypadkowy ładunek elektryczny jest równy zeru (kwazineutralność). W plazmie o niskiej temperaturze i małym stopniu jonizacji mogą występować atomy niezjonizowane (plazma niskotemperaturowa). W temperaturze wielu milionów stopni, w której mogą zachodzić reakcje termojądrowe, materia ulega całkowitej jonizacji, powstaje wówczas plazma wysokotemperaturowa — tzw. plazma gorąca.

plazma

W warunkach ziemskich plazmę spotyka się stosunkowo rzadko. Plazma o niskiej temperaturze występuje we wnętrzu płomienia, w iskrze lub łuku elektrycznym oraz w potężnych wyładowaniach atmosferycznych. Zewnętrzne warstwy atmosfery ziemskiej stanowi plazma o małej gęstości tworząca jonosferę. Ponad jonosferą istnieją obszary zawierające zjonizowaną materię, nazywane pasami van Allena. W obszarach międzygwiezdnych występuje również plazma

plazma we Wszechświecie



Rys. 1. Występowanie plazmy w przyrodzie oraz zakres badań nad plazmą. Prostokąt ukazuje zakres koncentracji i temperatur, którymi interesują się fizycy prowadzący badania termojądrowe; czarne punkty odpowiadają najważniejszym eksperymentom

o bardzo małej gęstości. Natomiast gorąca plazma występuje we wnętrzu miliardów gwiazd, w tym także naszego Słońca. Zatem tylko znikoma część materii we Wszechświecie występuje w trzech podstawowych stanach skupienia. Ponad 99,9% materii Wszechświata znajduje się w stanie zjonizowanym — stanowi plazmę, którą można uważać za odrębny stan materii (rys. 1).

W gorącej plazmie we wnętrzu gwiazd między cząstkami obdarzonymi wielką energią kinetyczną zachodzą reakcje syntezy termojądrowej. W reakcjach tych powstają nowe cięższe jądra atomowe, a jednocześnie wydzielane są ogromne ilości energii, której część dociera do nas w postaci promieniowania.

Reakcje syntezy termojądrowej i ich znaczenie

Reakcje przebiegające we wnętrzu gwiazd

Jako przykład zwykłej reakcji syntezy jądrowej rozpatrywać można przetworzenie czterech jąder wodoru (protonów) w jądro helu. Prawdopodobieństwo jednoczesnego zderzenia się czterech protonów jest jednak bardzo małe, nawet w warunkach jakie panują we wnętrzu gwiazd. Przetworzenie protonów w jądra helu możliwe jest poprzez szereg reakcji pośrednich, czyli w wyniku cyklicznej reakcji jądrowej, np. tzw. cyklu protonowo-protonowego, składającego się z następujących reakcji

cykl protonowo-protonowy

$$H^1+H^1 \rightarrow D^2+\beta^+ +\gamma$$
$$D^2+H^1 \rightarrow He^3+\gamma$$
$$2He^3 \rightarrow He^4+2H^1$$

lub

$$He^3+He^4 \rightarrow Be^7$$
$$Be^7+e^- \rightarrow Li^7$$
$$Li^7+H^1 \rightarrow 2He^7.$$

Obecnie uważa się, że energia wydzielona w reakcjach cyklu protonowo-protonowego jest głównym źródłem energii w gwiazdach, których temperatura wnętrza nie przekracza 15 mln K. Oprócz podanego wyżej cyklu protonowo-protonowego we wnętrzach gwiazd możliwy jest także cykl węglowo-azotowy (cykl Bethego)

cykl Bethego

$$C^{12}+H^1 \rightarrow {}^*N^{13}+\gamma$$
$${}^*N^{13} \rightarrow C^{13}+\beta^+ +\gamma$$
$$C^{13}+H^1 \rightarrow N^{14}+\gamma$$
$$N^{14}+H^1 \rightarrow {}^*O^{15}+\gamma$$
$${}^*O^{15} \rightarrow N^{15}+\beta^+ +\gamma$$
$$N^{15}+H^1 \rightarrow C^{12}+He^4,$$

gdzie * oznacza jądro wzbudzone.

„spalanie" wodoru

Korzystając z terminologii chemicznej, można powiedzieć, że powyższe reakcje opisują proces „spalania" wodoru na hel, w którym rolę „katalizatorów" spełniają jądra węgla, azotu i tlenu. Wydajność energetyczna podanych reakcji syntezy wynosi ok. 26 MeV na cykl, co oznacza, że ze spalenia 1 g wodoru można uzyskać ok. $6 \cdot 10^{11}$ dżuli energii. Omawiane cykle nie wyczerpują wszystkich możliwości. W temperaturach panujących we wnętrzach gwiazd przebiegać mogą również reakcje syntezy z udziałem innych pierwiastków lekkich, np. litu, berylu i boru. Produktem końcowym tych reakcji jest również hel. W temperaturze ok. 100 mln K zachodzą także inne reakcje, np. $C^{13}+He^4 \rightarrow O^{16}+n$. Powstające w wyniku takich reakcji swobodne neutrony mogą być pochłaniane przez jądra ciężkie, tworząc w ten sposób coraz cięższe jądra atomowe.

Znaczenie reakcji termojądrowych przebiegających na Słońcu

Omawiane procesy przebiegają bardzo powoli; we wnętrzu Słońca w ciągu jednego roku w procesach syntezy zużyte zostaje tylko jedno jądro na milion. Ze względu na ogromne rozmiary i masę Słońca, wystarcza to jednak do utrzymania odpowiedniej temperatury i pokrycia wszystkich strat energii związanych z promieniowaniem.

źródła energii na Ziemi

Aby ocenić znaczenie reakcji termojądrowych zachodzących we wnętrzu Słońca należy zauważyć, że prawie wszystkie zasoby energetyczne, z których korzystamy dotychczas na Ziemi, są pochodzenia słonecznego. Węgiel kamienny powstał ze szczątków roślin, które w minionych epokach wytworzone zostały w procesie fotosyntezy pod wpływem promieniowania. Z przemiany szczątków zwierzęcych i roślinnych, a więc pośrednio kosztem energii Słońca, powstała także ropa naftowa. Nawet energia wykorzystywana w hydroelektrowniach jest pochodzenia słonecznego, gdyż woda rzek pochodzi w większości z opadów atmosferycznych, a więc z pary wodnej, która powstaje głównie nad morzami, na skutek ogrzania ich przez Słońce. Podobnie jest z energią wiatrów. Można więc stwierdzić, że istnienie życia na Ziemi zależy w rzeczywistości od reakcji syntezy termojądrowej, które przebiegają nieustannie we wnętrzu Słońca — w ogromnym reaktorze termojądrowym o wymiarach liniowych rzędu 10^5 km, i temperaturze kilkunastu milionów K.

Reakcje syntezy termojądrowej możliwe do realizacji w warunkach ziemskich

paliwa termojądrowe

W dążeniu do poznania praw natury i opanowania nowych źródeł energii badacze zaczęli się zastanawiać, czy reakcje termojądrowe można realizować na Ziemi. Okazało się, że w warunkach ziemskich dysponujemy doskonałym paliwem termojądrowym — ciężkimi izotopami wodoru — deuterem i trytem:

$$D^2+D^2 \rightarrow T^3+p+4{,}0\ \text{MeV}$$
$$D^2+D^2 \rightarrow He^3+n+3{,}2\ \text{MeV}$$
$$D^2+T^3 \rightarrow He^4+n+17{,}6\ \text{MeV}$$
$$D^2+He^3 \rightarrow He^4+p+18{,}4\ \text{MeV}.$$

Interesujące są również reakcje syntezy z udziałem litu:

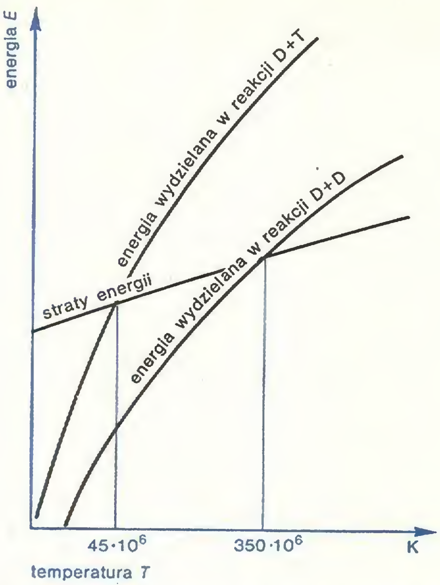
$$Li^6+n \rightarrow He^4+T^3+4{,}8\ \text{MeV}$$
$$Li^6+p \rightarrow He^4+He^3+3{,}9\ \text{MeV}$$
$$Li^6+D^2 \rightarrow 2He^4+22{,}3\ \text{MeV}.$$

temperatura „zapłonu"

Podstawowym warunkiem realizacji reakcji syntezy termojądrowej jest wytworzenie odpowiednio wysokiej temperatury. Występuje tu pewna analogia do temperatury zapłonu paliwa chemicznego. Oczywiście z tą różnicą, że temperatura „zapłonu" reakcji termojądrowych jest o wiele wyższa, wynosi dziesiątki, a nawet setki milionów stopni. W tak wysokich temperaturach materia może istnieć tylko w postaci całkowicie zjonizowanego gazu, czyli gorącej plazmy. Taka gorąca plazma intensywnie promieniuje, z czym związane są duże straty energetyczne. Jednak w miarę podwyższania temperatury energia wydzielana z reakcji termojądrowych wzrasta szybciej niż energia tracona na promieniowanie (rys. 2). W związku z powyższym za temperaturę zapłonu reakcji termojądrowych można uważać wartość tem-

temperatura zapłonu reakcji D–D i D–T

peratury, przy której energia wydzielana i tracona są sobie równe. W wypadku reakcji D–D temperatura ta wynosi ok. 350 mln K, a w wypadku reakcji D–T — ok. 45 mln K.



Rys. 2. Określenie temperatury zapłonu reakcji termojądrowych deuter-deuter (D+D) oraz deuter-tryt (D+T) przy założeniu, że gorąca plazma nie zawiera domieszek ciężkich atomów, które mogłyby powodować wzrost strat na promieniowanie

Warunki realizacji kontrolowanych reakcji termojądrowych

Warunek podstawowy, dotyczący ogrzania termojądrowego paliwa do temperatury zapłonu, omówiony został wcześniej. Dalsze warunki dotyczą odizolowania tej gorącej plazmy od otoczenia i utrzymania jej przez dostatecznie długi czas. Gorąca plazma musi przy tym posiadać odpowiednią gęstość. Równoczesne spełnienie tych warunków jest bardzo trudne. O skali trudności świadczą następujące dane: dla plazmy o koncentracji odpowiadającej normalnej gęstości powietrza (ok. $3 \cdot 10^{19}$ cząstek/cm^3) w temperaturze 100 mln K ciśnienie osiąga wartość ok. $2 \cdot 10^{11}$ Pa. W praktyce nie ma takich materiałów, które mogłyby wytrzymać działanie tak wysokich temperatur i tak wielkich ciśnień. Konieczne jest zatem obniżenie ciśnienia gorącej plazmy, np. przez zmniejszenie koncentracji cząstek. Ponieważ reakcja syntezy zachodzi tylko w trakcie zderzeń cząstek, zmniejszeniu koncentracji plazmy powinno jednak towarzyszyć wydłużenie czasu jej utrzymania w danym obszarze.

kryterium Lawsona

Jako kryterium może zatem służyć wartość iloczynu koncentracji n oraz czasu utrzymania τ (tzw. kryterium Lawsona). Według tego kryterium, przy określonych wyżej temperaturach zapłonu muszą być spełnione następujące warunki:

dla reakcji D–D $n\tau = 10^{16}$ cząstek·s/cm^3,
dla reakcji D–T $n\tau = 3 \cdot 10^{14}$ cząstek·s/cm^3.

W warunkach laboratoryjnych do utrzymywania gorącej plazmy wykorzystuje się silne pola magnetyczne. Kwazistacjonarne utrzymywanie plazmy jest jednak możliwe tylko wtedy, gdy ciśnienie wywierane przez pole magnetyczne jest równoważne lub większe od ciśnienia panującego w plazmie. Ciśnienie plazmy można przedstawić w postaci sumy dwóch składników:

$$p_g = n_e k T_e + n_i k T_i,$$

gdzie n_e oraz n_i — odpowiednio koncentracja elektronów i jonów, k — stała Boltzmanna, T_e oraz T_i — odpowiednia temperatura elektronów i jonów. Ciśnienie pola magnetycznego o indukcji B_0 wyraża się natomiast wzorem $p_m = B_0^2/8\pi$. Korzystając z podanych wyżej zależności łatwo obliczyć, ile dla plazmy o temperaturze $T_e = T_i = 400$ mln K wynosi jej krytyczna koncentracja, przy której można utrzymać plazmę polem magnetycznym (tabela).

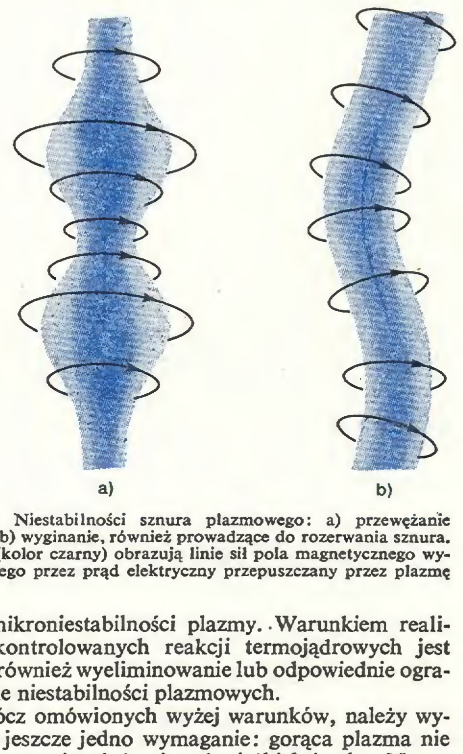
krytyczna koncentracja plazmy

B_0, T	n_{kryt}, cząstek/cm^3	B_0, T	n_{kryt}, cząstek/cm^3
1	10^{14}	50	$2,5 \cdot 10^{17}$
5	$2,5 \cdot 10^{15}$	100	10^{18}
10	10^{16}		

Za pomocą znanych dotychczas środków technicznych, można wytwarzać w sposób kwazistacjonarny pola do 100 kGs. Uzyskanie silniejszych pól magnetycznych jest możliwe tylko w urządzeniach pracujących impulsowo, a w wypadku pól powyżej 1 MGs — w sposób wybuchowy, niszczący (→ Najsilniejsze pola magnetyczne). Koncentracja plazmy, którą można utrzymać w sposób kwazistacjonarny, wynosi ok. 10^{15}–10^{16} cząstek/cm^3, a więc według kryterium Lawsona czas utrzymywania takiej plazmy powinien wynosić dla reakcji D–D — od 1 do 10 s, a dla reakcji D–T odpowiednio — od 0,03 do 0,33 s.

niestabilność plazmy

W praktyce największą przeszkodą w osiągnięciu czasu potrzebnego na utrzymanie gorącej plazmy są różne niestabilności magnetohydrodynamiczne (rys. 3)



Rys. 3. Niestabilności sznura plazmowego: a) przewężanie sznura; b) wyginanie, również prowadzące do rozerwania sznura. Okręgi (kolor czarny) obrazują linie sił pola magnetycznego wytworzonego przez prąd elektryczny przepuszczany przez plazmę

oraz mikroniestabilności plazmy. Warunkiem realizacji kontrolowanych reakcji termojądrowych jest zatem również wyeliminowanie lub odpowiednie ograniczenie niestabilności plazmowych.

wpływ domieszek

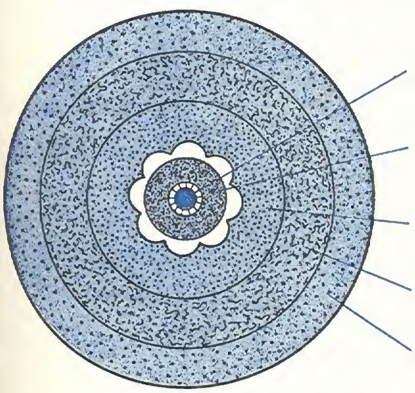
Oprócz omówionych wyżej warunków, należy wymienić jeszcze jedno wymaganie: gorąca plazma nie powinna zawierać domieszek ciężkich jonów. Nawet niewielkie zanieczyszczenie plazmy ciężkimi jonami wywołuje bowiem duży wzrost strat energii na promieniowanie, w związku z czym znacznie podwyższa się temperatura zapłonu reakcji termojądrowych. Należy więc dążyć do uzyskania bardzo czystej plazmy deuterowej (lub deuterowo-trytowej). W celu jak najszybszego osiągnięcia dodatniego bilansu energetycznego trzeba również dążyć do obniżenia strat energii we wszystkich urządzeniach pomocniczych: w uzwojeniach wytwarzających pole magnetyczne, w układach pomp próżniowych oraz w urządzeniach kontrolnych i pomiarowych.

Militarne wykorzystywanie energii termojądrowej

O możliwości wytworzenia gorącej plazmy i zrealizowania niekontrolowanych reakcji termojądrowych w warunkach ziemskich przekonano się ostatecznie w 1952 r., kiedy na atolu Eniwetok przeprowadzono eksplozję pierwszej amerykańskiej bomby termojądrowej. Paliwem w tej bombie był ciekły deuter i tryt. W rzeczywistości była to nie bomba, ale mała fabryka wyposażona w potężne urządzenia chłodnicze o wadze ponad 80 ton. Jako źródło energii inicjującej reakcje termojądrowe wykorzystano wówczas bombę atomową o mocy 5-krotnie większej od tej, która zniszczyła Hiroszimę. W efekcie eksplozja miała energię równoważną wybuchowi wielu milionów ton trotylu (TNT — trójnitrotoluen). W około 10 miesięcy później w ZSRR zrealizowano wybuch bomby termojądrowej, w której wykorzystano reakcje z litem.

niekontrolowane reakcje termojądrowe

Następnie, w 1954 r. Stany Zjednoczone przeprowadziły wybuch bomby litowej na atolu Bikini. Była to potężna eksplozja wyzwalająca ok. 15 megaton TNT energii, która prawie 10-krotnie przewyższała siłę wybuchową wszystkich bomb zrzuconych na terytorium Niemiec w latach II wojny światowej. Wtedy właśnie doszło do tragicznego wypadku z japońskimi rybakami, którzy ulegli napromieniowaniu w odległości ok. 200 km od miejsca eksplozji. Od czasu wybuchu tych pierwszych bomb termojądrowych aż do



Rys. 4. Schemat budowy bomby termojądrowej typu 3F

1958 r., kiedy to wstrzymano eksperymenty z bronią jądrową, przeprowadzono eksplozje wielu bomb wodorowych i litowych: były to eksplozje naziemne, powietrzne, podziemne i podwodne.

Współczesne bomby termojądrowe są przeważnie typu 3F, co oznacza „fission-fusion-fission" (rozszczepienie-synteza-rozszczepienie). Schemat ich budowy przedstawia rys. 4. W bombach tych materiałem podstawowym jest deuterek litu ^6LiD otoczony płaszczem z naturalnego uranu ^{238}U. Reakcję termojądrową inicjuje wybuch zwykłej bomby atomowej zawierającej uran ^{235}U lub pluton ^{239}Pu. Potem rozwija się reakcja syntezy termojądrowej z udziałem litu ^6Li. Wytwarzane przy tym neutrony prędkie przenikają do płaszcza z uranu ^{238}U, powodując reakcję rozszczepieniową i zwiększając wypadkowy efekt wybuchowy. Warto tutaj zestawić następujące dane:

1 kg uranu w procesie zwykłego rozszczepienia wyzwala energię równoważną wybuchowi 20 kiloton TNT,
1 kg ^6LiD w reakcji syntezy wyzwala energię 68 kiloton TNT,
1 kg mieszaniny D+T w reakcjach syntezy wyzwala energię rzędu 80 kiloton TNT.

Dla porównania należy przypomnieć, że do kompletnego zniszczenia Hiroszimy wystarczyła bomba atomowa o energii równoważnej 20 kilotonom TNT.

Łatwo wyciągnąć stąd wniosek, że w arsenałach wielu państw (broń termojądrową mają obecnie również Anglia, Francja i Chiny) drzemią dziś ogromne siły, których niszczycielskie użycie groziłoby całkowitym zniszczeniem naszej cywilizacji.

Możliwości pokojowego wykorzystania energii termojądrowej

Potężne wybuchy termojądrowe można wykorzystywać do wielkich robót ziemnych, do kopania ogromnych kanałów lub drążenia podziemnych zbiorników. Istnieją także projekty uzyskiwania energii z podziemnych eksplozji termojądrowych. W warunkach laboratoryjnych i przemysłowych nie można jednak jako zapalnika stosować bomb atomowych. Konieczne jest zatem opanowanie kontrolowanych reakcji termojądrowych przebiegających w sposób regulowany przez człowieka. Program ten napotyka jednak duże trudności i wymaga wielkich nakładów finansowych. Zanim szczegółowo omówimy te zagadnienia warto się zastanowić, czy cel, jakim jest opanowanie kontrolowanej syntezy termojądrowej, wart jest tych zabiegów?

W pierwszej kolejności warto poznać następujące dane:

1 kg węgla daje do 7 kWh energii,
1 kg uranu daje od 70 000 kWh (praktycznie) do 12 000 000 kWh (teoretycznie),
1 kg deuteru w reakcjach syntezy mógłby wytworzyć 24 000 000 kWh.

zasoby węgla i uranu

Zasoby węgla są eksploatowane bardzo intensywnie i obecnie ocenia się, że wystarczy ich jeszcze na około 100 lat. Podobna sytuacja jest ze złożami ropy naftowej. Zasoby uranu, natomiast, przy obecnej sprawności przetwarzania (gwarantującej zużycie zaledwie 0,7% masy paliwa) mogą wystarczyć zaledwie na kilkadziesiąt lat. Uran naturalny może stanowić trwałe źródło energii tylko w reaktorach powielających, których wykorzystanie w praktyce napotyka jeszcze trudności (→ Energia jądrowa). Poważny problem energetyki jądrowej stanowią również odpady promieniotwórcze.

zalety deuteru jako paliwa

W porównaniu z zasobami paliw tradycyjnych i uranu zapasy paliwa termojądrowego — deuteru — są właściwie niewyczerpalne. Deuter występuje w wodach mórz i oceanów w ilości 1 atom izotopu na 5000 atomów wodoru zwykłego. Łączne zasoby deuteru wynoszą około 10^{17} kg, co odpowiada zapasom energetycznym rzędu 10^{24} kWh. Przyjmując, że zużycie energii utrzyma się na poziomie dzisiejszym (co odpowiada wykorzystaniu mocy rzędu 5 miliardów kW) deuteru wystarczy na 20 miliardów lat. Nawet przy najbardziej intensywnym rozwoju energetyki termojądrowej paliwa tego nie zabraknie. Koszty wydobycia tego paliwa są przy tym bardzo małe. Współczesnymi metodami produkcji, kosztem 3 zł z 1 l wody wydziela się ok. 1/30 g deuteru. W reakcjach termojądrowych z takiej ilości deuteru wyzwolić można energię równoważną spaleniu 300 l benzyny. Wykorzystanie tych ogromnych zasobów energetycznych jest jednak uzależnione od opanowania kontrolowanych reakcji termojądrowych, a w pierwszej kolejności — od poznania własności materii w bardzo wysokich temperaturach.

Własności plazmy

Kwazineutralność plazmy

Do podstawowych własności plazmy zaliczyć należy wspomnianą na wstępie kwazineutralność elektryczną. W rzeczywistości w plazmie występuje pewne rozdzielenie ładunków. Koncentracja dodatnich i ujem-

nych cząstek jest jednak tak duża, zwłaszcza w plazmie o dużej gęstości, że nawet małe rozdzielenie ładunków wywołuje bardzo silne pole elektryczne (E), które przeciwdziała dalszym zmianom. W plazmie wówczas mogą powstawać oscylacje elektrostatyczne. Natomiast w obecności zewnętrznego pola magnetycznego H_0 w plazmie mogą występować także drgania o charakterze elektromagnetycznym. Do najważniejszych należą podłużne fale magnetohydrostatyczne, czyli tzw. fale Alfvéna (dla których $\vec{k} \| \vec{H}_0$, $\vec{E} \perp \vec{H}_0$, gdzie $\vec{k}$ jest wektorem falowym — reprezentującym kierunek rozchodzenia się fali), oraz poprzeczne fale magnetoakustyczne, czyli tzw. fale magnetodźwiękowe ($\vec{k} \perp \vec{H}_0$, $\vec{E} \perp \vec{H}_0$).

Promieniowanie plazmy

Inną podstawową własnością plazmy jest jej promieniowanie. Przy niskich temperaturach i małym stopniu jonizacji plazma emituje przede wszystkim promieniowanie o widmie dyskretnym — każda linia widmowa odpowiada przejściom elektronów między określonymi poziomami energetycznymi atomów lub jonów. Ze wzrostem temperatury (i jonizacji) plazmy wzrasta udział promieniowania o widmie ciągłym, pochodzącego z procesów rekombinacji jonów i elektronów oraz z procesu hamowania swobodnych elektronów w polu elektrycznym jonów. Zgodnie ze znaną zasadą, emisji promieniowania towarzyszy jego absorpcja (pochłanianie). Im większa jest gęstość i grubość warstwy plazmowej, tym silniej pochłania ona promieniowanie. Warto zwrócić uwagę na tę własność plazmy, ponieważ od niej zależy istnienie życia na Ziemi. Gdyby gorąca plazma Słońca była zupełnie przezroczysta, to promieniowanie, które docierałoby do powierzchni naszej planety, zamieniłoby wszystko w popiół. W rzeczywistości ze Słońca dociera do nas promieniowanie tylko z zewnętrznej warstwy plazmowej.

pochłanianie plazmy

Zgodnie z tym, co powiedziano wyżej, gorąca plazma może emitować promieniowanie o bardzo szerokim widmie — od fal radiowych do twardego promieniowania rentgenowskiego. Przez plazmę może przenikać tylko promieniowanie, którego częstość ω jest większa od częstości plazmowej $\omega_0 = \sqrt{4\pi n e^2/m_e}$ gdzie n, e, m_e są odpowiednio koncentracją, ładunkiem i masą elektronu. Analiza tego promieniowania umożliwia określenie podstawowych parametrów plazmy — jej koncentracji i temperatury.

W świetle molekularno-kinetycznej teorii budowy materii temperatura jest miarą średniej energii ruchu cieplnego, która jest proporcjonalna do kwadratu średniej prędkości ruchu cząstek. Należy przy tym zauważyć, że nie wszystkie cząstki mają taką samą energię kinetyczną. Istnieją cząstki poruszające się bardzo wolno i takie, które obdarzone są dużymi prędkościami i ulegają większej liczbie zderzeń. W wyniku tych zderzeń ustala się zwykle stan równowagi termodynamicznej, w którym rozkład prędkości cząstek (a tym samym ich rozkład energetyczny) można opisać funkcją maxwellowską. Pojęcie temperatury jest dobrze zdefiniowane jedynie dla stanu równowagi termodynamicznej.

temperatura plazmy

W praktyce pojęcie temperatury stosuje się również w wypadkach odchyleń od rozkładu maxwellowskiego rozpatrując np. oddzielnie funkcje rozkładu dla ruchów poprzecznych lub podłużnych (w stosunku do wyróżnionego kierunku). Często rozkład energetyczny elektronów w plazmie różni się od rozkładu energetycznego jonów lub resztek atomów neutralnych. (Wprowadza się wówczas pojęcie temperatury elektronowej T_e oraz temperatury jonowej T_j.) Przy dostatecznie dużej gęstości plazmy, w wyniku zderzeń między cząstkami, ustala się w plazmie po pewnym czasie stan równowagi termodynamicznej (wówczas $T_e = T_j$). Czas ten nazywa się czasem relaksacji. Należy pamiętać, że „temperatura” oznacza wówczas tylko miarę średniej energii kinetycznej i nie charakteryzuje w pełni stanu cieplnego danego ośrodka.

temperatura elektronowa i jonowa

Ze względu na proporcjonalność temperatury i średniej energii kinetycznej cząsteczek, temperaturę wyraża się czasem w jednostkach energetycznych — elektronowoltach. Energii 1 eV odpowiada przy tym temperatura 11 600 K (w przybliżeniu 1 eV $\triangleq$ 10^4K.

Inne własności plazmy

Występowanie w plazmie dużej ilości swobodnych elektronów i jonów powoduje, że w odróżnieniu od normalnego gazu plazma posiada zdolność przewodzenia prądu elektrycznego. W przeciwieństwie do zwykłych przewodników metalicznych wraz ze wzrostem temperatury opór elektryczny plazmy maleje, zmieniając się proporcjonalnie do $T^{-3/2}$. W rezultacie gorąca plazma wykazuje dużą przewodność elektryczną i w temperaturze kilkunastu mln K jej opór właściwy jest mniejszy od oporu najlepszych przewodników metalicznych.

opór elektryczny plazmy

Ze względu na występowanie w plazmie dużej liczby swobodnych cząstek naładowanych plazma podlega również działaniu sił pola elektrycznego i magnetycznego, co można wykorzystać do jej utrzymania w określonym obszarze w tzw. pułapce magnetycznej. Prawa, które rządzą ruchem cząstek naładowanych w polach elektrycznych i magnetycznych, dają wytłumaczenie wielu zjawisk zachodzących w plazmie. Konieczne jest jednak uwzględnienie, że w plazmie ruch swobodnych cząstek naładowanych zakłócony jest przez zderzenia. Ponieważ naładowane cząstki wytwarzają pola elektryczne o stosunkowo dużym zasięgu, w plazmie o dużej gęstości występują oddziaływania kolektywne. Polega to na tym, że plazma wykazuje skłonność do rozprzestrzeniania zaburzeń lokalnych na cały obszar plazmowy. Jeżeli takie zaburzenia nie są dostatecznie tłumione, występuje niestabilność plazmy (zob. rys. 3). Zjawiska te utrudniają utrzymywanie plazmy w pułapkach magnetycznych przez dłuższy czas.

oddziaływania kolektywne

niestabilność plazmy

Plazmę o małej gęstości można traktować jako zbiór pojedynczych cząstek, z których każda oddziałuje niezależnie — jest to tzw. model jednocząstkowy. Plazmę o większej gęstości można rozpatrywać jako przewodzący ośrodek ciągły, który opisują równania hydrodynamiki i elektrodynamiki — jest to tzw. model magnetohydrodynamiczny (model mhd).

Omówione wyżej własności plazmy są bardzo istotne zarówno z punktu widzenia badań podstawowych, jak również badań zmierzających do opanowania energii termojądrowej.

Metody wytwarzania gorącej plazmy

Z wytwarzaniem plazmy związany jest problem pomiaru wartości bardzo wysokich temperatur. Do pomiarów takich temperatur nie przydatne są oczywiście ani termometry cieczowe, ani gazowe. Nie można też stosować termometrów oporowych lub termoelektrycznych. Specjalne sondy kalorymetryczne pozwalają wprawdzie mierzyć temperaturę rzędu kilku tysięcy kelwinów, ale ulegają szybkiemu zużyciu, a prócz tego zaburzają badany ośrodek. Dlatego do pomiarów wysokich temperatur najbardziej przydatne są metody optyczne. Powyżej 1000 K powszechnie stosowany jest np. pirometr optyczny, którego działanie oparte jest na porównaniu promieniowania badanego obiektu z promieniowaniem włókien żarówki — wzorca. Za pomocą takiego przyrządu można np. określić temperaturę płomienia, która może sięgać ok. 2000 K.

metody optyczne pomiaru temperatury

metody spektroskopowe pomiaru temperatury

Do pomiarów jeszcze wyższych temperatur wykorzystuje się fakt, że gorąca plazma wysyła promieniowanie o różnych długościach fal, od fal radiowych do krótkofalowego promieniowania nadfioletowego a nawet rentgenowskiego. Badania tego promieniowania można przeprowadzić za pomocą spektrografów i monochromatorów. Na podstawie pomiaru stosunku natężeń linii widmowych emitowanych przez jony tego samego pierwiastka (ale o różnych stopniach wzbudzenia) można np. ocenić temperaturę T_e, nawet jeśli wynosi ona wiele milionów kelwinów. Do określenia temperatury T_j można natomiast wykorzystać zależność szerokości linii widmowych od ruchu jonów (tzw. dopplerowskie rozszerzenie linii).

Metody wytwarzania plazmy o bardzo wysokich temperaturach można podzielić na dwie zasadnicze kategorie. Do pierwszej należą metody polegające na wytworzeniu plazmy (lub wiązki cząstek) o wysokiej energii w specjalnych urządzeniach akceleracyjnych, a następnie — na iniekcji tej plazmy (lub cząstek) do wnętrza pułapki magnetycznej. Do drugiej kategorii zaliczają się metody, które polegają na wytworzeniu chłodnej plazmy od razu wewnątrz pułapki magnetycznej, a następnie — na ogrzewaniu tej plazmy do bardzo wysokich temperatur. Można tu podać pewną analogię: metody pierwszej kategorii przypominają napełnianie termosu gorącym płynem, a metody drugiej kategorii odpowiadają nagrzewaniu płynu wewnątrz zamkniętego kociołka. Klasyfikacja metod wytwarzania gorącej plazmy według miejsca jej powstawania jest jednak zbyt powierzchowna i dlatego konieczne jest rozważenie różnych procesów fizycznych, na których te metody się opierają. Poniżej podamy bardziej szczegółowy opis niektórych metod wytwarzania gorącej plazmy i porównamy ich możliwości.

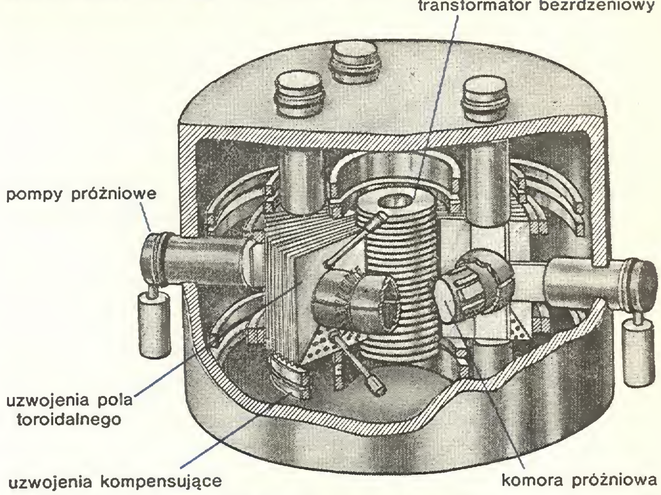
Metoda grzania omowego

Jest to najprostsza metoda wytwarzania gorącej plazmy polegająca na przepuszczeniu przez zjonizowany gaz bardzo silnych prądów elektrycznych, w wyniku czego następuje dalsza jonizacja gazu i wzrost jego temperatury. Wydzielanie ciepła wiąże się przy tym ze zderzeniami między nośnikami prądu (elektronami) i innymi cząstkami gazu. Ponieważ przy podwyższaniu temperatury opór plazmy szybko maleje (przeciwnie niż w przypadku zwykłych przewodników metalicznych), możliwości grzania omowego są ograniczone. Stosując omawianą metodę można osiągnąć temperatury ok. kilku keV (rzędu 10^7 K), ale ze wzrostem temperatury wydajność tej metody maleje. Nie pomaga również zwiększanie natężenia przepuszczanych prądów, ponieważ w pewnych warunkach może to wywołać niestabilność plazmy.

Metoda grzania omowego jest jednak bardzo wygodna i może być wykorzystana w różnych układach eksperymentalnych. Grzanie omowe można stosować w układach typu otwartego (patrz rozdział następny) wykorzystując zewnętrzne elektrody i przepuszczając przez plazmę impulsy prądu o natężeniu rzędu milionów A. Uzyskuje się wówczas plazmę o temperaturze ok. kilkuset eV (kilku milionów K) i koncentracji rzędu 10^{16} cząstek/cm³, ale o bardzo krótkim czasie trwania (rzędu 10^{-5} s), co spowodowane jest niestabilnością sznura plazmowego.

Grzanie omowe stosuje się również w zamkniętych (toroidalnych) pułapkach magnetycznych, w których prądy w plazmie wytwarzane są metodą indukcyjną za pomocą odpowiednich transformatorów. Przykład toroidalnego układu eksperymentalnego, wyposażonego w transformator do grzania omowego, przedstawiono na rys. 5. Ze względu na swoją prostotę, metoda grzania omowego stosowana jest w większości układów zamkniętych. Przy pomocy tej właśnie metody w urządzeniach typu tokamak udało się uzyskać plazmę o koncentracji $6 \cdot 10^{13}$ cząstek/cm³ i temperaturze $T_j \approx 700$ eV (ok. 7 milionów K) i utrzymać ją przez stosunkowo długi czas (ok. 30 ms).

Alcator



Rys. 5. Schemat układu toroidalnego Alcator (zbudowany w Stanach Zjednoczonych), w którym stosuje się grzanie omowe plazmy prądem indukowanym przez transformator bezrdzeniowy. Maksymalne natężenie tego prądu osiąga wartość 650 kA. Mała średnica toroidalnej komory układu wynosi 25 cm, duża — ok. 100 cm. Do utrzymywania plazmy wykorzystuje się silne zewnętrzne pole magnetyczne $B_{max} = 12$ T

Metody grzania turbulencyjnego

Grzanie turbulencyjne występuje przy oddziaływaniu z plazmą wiązki elektronów, gdy natężenie tej wiązki przekracza pewną wartość krytyczną. Można je zrealizować dwoma sposobami. Przy stosowaniu pierwszego sposobu najpierw wytwarza się gęstą, ale stosunkowo zimną plazmę, wykorzystując inne metody (np. wstępną jonizację gazu resztkowego w komorze lub iniekcję plazmy z zewnątrz). Następnie przez umieszczoną w polu magnetycznym plazmę przepuszcza się krótkie impulsy prądu o bardzo dużym natężeniu przykładając do elektrod impulsy wysokiego napięcia. Przy pomocy takiej metody udało się w pułapkach typu zwierciadlanego otrzymać plazmę o koncentracji $2 \cdot 10^{13}$ cząstek/cm³ i temperaturze T_j wynoszącej 3–5 keV (rzędu 10^7 K). W pułapkach zamkniętych uzyskano omawianą metodą zbliżone wartości temperatury, ale mniejszą koncentrację plazmy.

Przy stosowaniu drugiego sposobu plazmę wytwarza od razu wiązka elektronowa przechodząca przez gaz niezjonizowany. Wymagane są wówczas mniejsze prądy, niż przy omawianej wyżej metodzie, ale impulsy znacznie dłuższe. Przy zastosowaniu omawianego sposobu w pułapce typu zwierciadlanego udało się uzyskać plazmę o koncentracji 10^{11}–10^{12} cząstek/cm³, temperaturze $T_j \approx 1$ keV i bardzo wysokiej temperaturze $T_e \approx 150$ keV. Podobne wyniki (z wyjątkiem tak wysokiej temperatury elektronowej) uzyskano ostatnio także w pułapkach toroidalnych. Należy podkreślić, że procesy powodujące grzanie plazmy przy oddziaływaniu wiązka–plazma są bardzo skomplikowane i nie mają jeszcze pełnej interpretacji teoretycznej.

Metody grzania rezonansowego

Do wytwarzania gorącej plazmy w różnych pułapkach magnetycznych stosowane jest również grzanie rezonansowe, które występuje przy oddziaływaniu z plazmą silnych fal elektromagnetycznych o odpowiednio dobranej częstości. Ponieważ w polu magnetycznym naładowane cząstki plazmy (elektrony i jony) wykonują m.in. ruchy wirowe z częstością cyklotronową —

f_{ce} oraz f_{ci}, to dobierając odpowiednio częstość fali elektromagnetycznej f można wywołać elektronowy rezonans cyklotronowy (gdy $f = f_{ce}$), lub jonowy rezonans cyklotronowy (gdy $f = f_{ci}$); prowadzi to do wzbudzenia intensywnego ruchu cząstek kosztem energii fali. Dla przykładu — w polu o indukcji $B = 5T$ częstość cyklotronowa elektronów wynosi $f_{ce} = 140$ GHz, co odpowiada fali o długości 2 mm, a częstość cyklotronowa protonów wynosi f_{ci} — 78 MHz, co odpowiada fali o długości ok. 4 m.

wzbudzanie elektronowego rezonansu cyklotronowego

Elektronowy rezonans cyklotronowy można uzyskać stosując zewnętrzny generator mikrofalowy i doprowadzając falę elektromagnetyczną do plazmy przez odpowiednie falowody. Należy przy tym zauważyć, że tłumienie fali elektromagnetycznej i przekazywanie energii plazmie może następować nie tylko w wyniku rezonansu cyklotronowego, ale również przy jednakowych prędkościach fazowych fali i cząstek plazmy (występuje wtedy tzw. tłumienie Landaua). Metoda elektronowego rezonansu cyklotronowego może być stosowana zarówno w zamkniętych jak i otwartych pułapkach magnetycznych. W stellaratorach za pomocą tej metody otrzymano plazmę o koncentracji 10^{10}–10^{12} cząstek/cm³ i temperaturze $T_e \approx$ $\approx$ 30 eV. W otwartej pułapce typu zwierciadlanego z dodatkowym polem stabilizacyjnym (eksperyment INTEREM) udało się uzyskać w ten sposób plazmę o koncentracji 10^{12} cząstek/cm³ i bardzo wysokiej temperaturze $T_e \approx 100$ keV.

wzbudzanie jonowego rezonansu cyklotronowego

Do wzbudzenia jonowego rezonansu cyklotronowego nie można wykorzystać falowodów ze względu na znaczną długość fali (rzędu kilku metrów), tak że fale elektromagnetyczne muszą być doprowadzane do plazmy innym sposobem. Stosuje się do tego celu układy specjalnych cewek (tzw. cewki Stixa), zasilane z silnych generatorów pracujących na częstości radiowej. Metoda jonowego rezonansu cyklotronowego jest także wykorzystywana zarówno w pułapkach zamkniętych jak i otwartych. Udało się za jej pomocą wytworzyć w Stellaratorze C plazmę o koncentracji 10^{12}–10^{13} cząstek/cm³ i temperaturze $T_j \approx 550$ eV. W eksperymentach z otwartymi pułapkami zwierciadlanymi przy zastosowaniu omawianej metody uzyskano plazmę o koncentracji 10^{14} cząstek/cm³ i temperaturze $T_j \approx 1{,}5$ keV (przy absorpcji ok. 30% dostarczanej energii).

stochastyczne grzanie elektronów

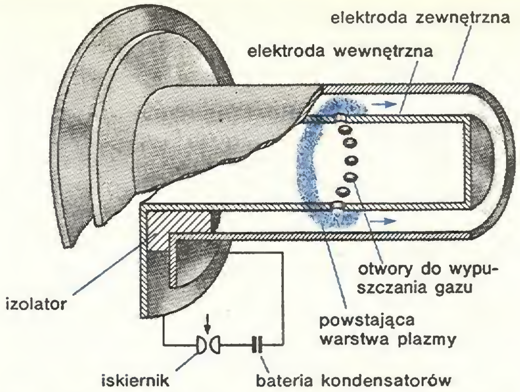
Do metod rezonansowych zalicza się również tzw. stochastyczne grzanie elektronów, które można zrealizować w następujący sposób. Do szczeliny w metalowej przegrodzie komory eksperymentalnej przykłada się napięcie o częstości rzędu kilku MHz. Elektrony, które powstają w wyniku jonizacji gazu w tej komorze, dostają się do obszaru wytwarzanego pola i uzyskują wówczas dodatkową energię. Osiągana temperatura elektronowa jest proporcjonalna do wielkości $(fU)^{2/3}$, gdzie f — częstość stosowanego napięcia, U — jego amplituda. Wykorzystując powyższą metodę w pułapce typu stellarator uzyskano plazmę o koncentracji 10^9 cząstek/cm³ i temperaturze $T_e \approx 300$ eV.

Metody iniekcyjne

Do napełniania pułapek magnetycznych (zwłaszcza pułapek typu otwartego) bardzo często stosowane są metody iniekcyjne, które polegają na wstrzeliwaniu do wnętrza pułapki strumienia gorącej plazmy, wiązki wysokoenergetycznych jonów lub szybkich atomów neutralnych.

iniektory współosiowe

Do iniekcji plazmy stosowane są różnego typu iniektory plazmowe (tzw. działa plazmowe). Najczęściej w tym celu wykorzystuje się iniektory współosiowe typu Marshalla. Schemat budowy takiego iniektora przedstawiono na rys. 6. Jest on wyposażony w dwie cylindryczne i współosiowe elektrody. Jego zasada działania jest następująca: najpierw do obszaru między elektrodami wpuszcza się impulsowo określoną ilość gazu pod normalnym ciśnieniem (np. wodoru lub deuteru). Następnie elektrody iniektora przyłącza się do baterii naładowanych do wysokiego napięcia kondensatorów. Wyładowanie elektryczne, które rozwija się w obszarze międzyelektrodowym, powoduje jonizację wpuszczonego wcześniej gazu i utworzenie się warstwy plazmy, przez którą płynie prąd wyładowania. Pole magnetyczne, które towarzyszy przepływowi tego prądu, wywołuje przesuwanie i stopniowe przyspieszanie utworzonej warstwy plazmy w kierunku wylotu iniektora. W rezultacie z iniektora wyrzucany jest strumień wysokoenergetycznej plazmy lub oddzielne zgęstki plazmowe — tzw. plazmoidy.



Rys 6. Schemat współosiowego iniektora plazmowego (nazywanego także działem plazmowym lub koaksjalnym akceleratorem plazmy)

plazmoidy

Oprócz współosiowych iniektorów plazmowych, do wytwarzania strumieni gorącej plazmy stosuje się również inne typy iniektorów, np. bezelektrodowe iniektory indukcyjne oraz iniektory typu Halla.

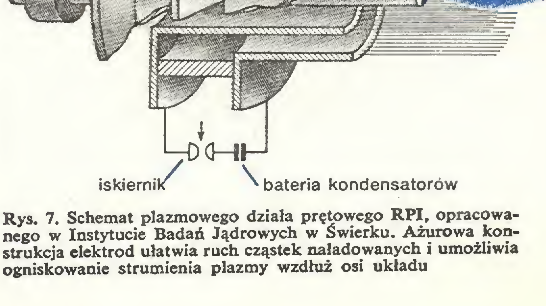
iniektory indukcyjne

W iniektorach indukcyjnych plazmę wytwarza się przez szybkie rozładowanie baterii kondensatorów przez spiralne uzwojenie, które otacza stożkową komorę szklaną lub ceramiczną. W czasie przepływu prądu przez uzwojenie pojawia się szybkozmienne pole magnetyczne, które indukuje wirowe pole elektryczne, wywołujące z kolei jonizację gazu i powstanie plazmy. Oddziaływanie z zewnętrznym polem magnetycznym prowadzi w rezultacie do przyspieszenia wytworzonej plazmy.

iniektory Halla

W iniektorach typu Halla wyładowanie plazmowe inicjowane jest w polu elektrycznym skrzyżowanym z kwazistacjonarnym polem magnetycznym, np. we współosiowych iniektorach Halla stosuje się podłużne pole elektryczne E_z oraz radialne pole magnetyczne B_r.

Interesującą odmianą iniektora plazmowego jest tzw. działo prętowe RPI, które opracowano w Instytu-



Rys. 7. Schemat plazmowego działa prętowego RPI, opracowanego w Instytucie Badań Jądrowych w Świerku. Ażurowa konstrukcja elektrod ułatwia ruch cząstek naładowanych i umożliwia ogniskowanie strumienia plazmy wzdłuż osi układu

działo prętowe RPI

cie Badań Jądrowych w Świerku. Działo RPI ma również cylindryczne elektrody, ale — w odróżnieniu od zwykłych iniektorów współosiowych — zbudowane są one z dużej liczby cienkich prętów ułożonych współosiowo na obwodzie cylindrycznych pierścieni (rys. 7). W czasie wyładowań elektrycznych cząstki naładowane mogą poruszać się swobodnie między prętami elektrod i ulegać przyspieszeniu. Ponieważ powierzchnia czynna elektrod jest w tym wypadku mniejsza niż w iniektorze z elektrodami pełnymi, wytwarzana plazma może zawierać mniej zanieczyszczeń pochodzących z powierzchni elektrod iniektora.

Omawiane wyżej iniektory mogą wytwarzać plazmoidy o koncentracji 10^{13}–10^{15} cząstek/cm³, zależnie od typu iniektora i warunków jego pracy. Przy odpowiednim doborze warunków pracy iniektora wytwarzane plazmoidy mogą mieć bardzo duże prędkości rzędu 10^7–10^8 cm/s, a ich średnia energia kinetyczna może osiągać wartość kilku keV. Ze względu na to, że znaczna część tej energii związana jest z ruchem uporządkowanym, a nie z chaotycznym ruchem cieplnym cząstek, temperatura wytwarzanej plazmy nie przekracza kilkuset eV (tj. kilku milionów kelwinów).

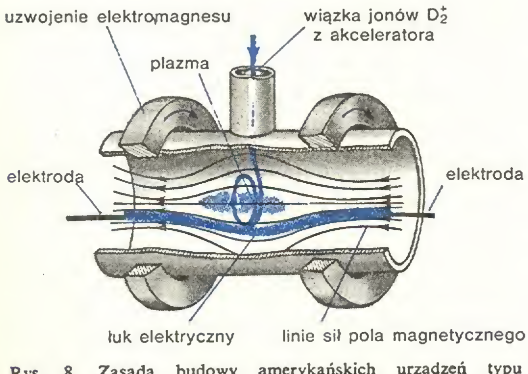
Przy iniekcji plazmy do wnętrza pułapki magnetycznej wstrzeliwana plazma musi pokonać barierę pola magnetycznego, co wpływa na obniżenie jej koncentracji i prędkości, a w niektórych wypadkach — także temperatury. Efekty te zależą od konfiguracji i natężenia pola magnetycznego pułapki. Stosując iniekcję plazmy w pułapkach otwartych typu Probkotron uzyskano plazmę o koncentracji 10^9 cząstek/cm³ i temperaturze $T_j \approx 5$ keV. W pułapkach typu stellarator za pomocą iniektorów plazmowych uzyskano plazmę o koncentracji 10^9–10^{11} cząstek/cm³ i temperaturze jonowej 30–100 eV.

Iniekcja wysokoenergetycznych jonów

Gorącą plazmę można również wytwarzać przez iniekcję do pułapek magnetycznych wiązek wysokoenergetycznych jonów, które uzyskuje się za pomocą akceleratorów cząstek naładowanych. Bardzo dobre rezultaty daje przy tym metoda polegająca na wstrzeliwaniu jonów cząsteczkowych D_2^+, które wewnątrz pułapki ulegają dysocjacji na neutralne atomy deuteru i jony atomowe D^+. Neutralne atomy uciekają wówczas z pułapki, unosząc pewną część energii i pędu, a jony D^+ są zatrzymywane w polu magnetycznym, powodując jonizację (tzw. „wypalanie") resztek gazu w komorze eksperymentalnej.

układ DCX

Omówiona metoda stosowana była w amerykańskich urządzeniach DCX, których zasadę budowy przedstawiono na rys. 8, a widok ogólny na ilustracji 46 (tabl. 12). W urządzeniach tych wstępna dysocjacja jonów D_2^+ następowała w łuku elektrycznym, który palił się między dwiema elektrodami umieszczonymi w pułapce magnetycznej typu zwierciadlanego. Wysokoenergetyczne jony D^+ zderzając się z cząsteczkami pozostałego w komorze gazu wytwarzały obłok gorącej plazmy. Energii wstrzeliwanych jonów odpowiadała temperatura rzędu 6 miliardów K. Tak wysoką temperaturę uzyskiwała jednak tylko część cząstek plazmy; pozostałe cząstki miały temperaturę znacznie niższą.



Rys. 8. Zasada budowy amerykańskich urządzeń typu DCX. Wyładowanie łukowe powoduje dysocjacje wstrzeliwanych jonów molekularnych D_2^+, a do utrzymywania gorącej plazmy służy pułapka magnetyczna typu zwierciadlanego

układ OGRA

Iniekcja jonów była również stosowana w jednym z największych urządzeń radzieckich — w układzie OGRA. Komora tego układu miała średnicę ok. 2 m i długość ok. 20 m. W odróżnieniu od układów DCX, w układzie tym nie stosowano pomocniczego wyładowania łukowego, a dysocjacja jonów molekularnych następowała na skutek zderzeń tych jonów z neutralnymi cząsteczkami pozostałego w komorze gazu. Podobnie jak w urządzeniach DCX w urządzeniu OGRA uzyskiwano temperatury setek milionów K.

Iniekcja wysokoenergetycznych atomów neutralnych

Plazma o bardzo wysokiej temperaturze może być wytwarzana także za pomocą intensywnych wiązek prędkich atomów neutralnych. Wiązki takie można uzyskać przepuszczając wiązkę wysokoenergetycznych jonów przez komorę zawierającą znaczną liczbę cząsteczek gazu. Część jonów wychwytuje elektrony i biegnie dalej jako prędkie atomy neutralne. Uzyskaną w ten sposób wiązkę wysokoenergetycznych atomów neutralnych można następnie wprowadzić do pułapki magnetycznej, gdzie przy odpowiednio dobranych warunkach, atomy te ulegają ponownej jonizacji tworząc plazmę o odpowiednio wysokiej temperaturze.

Metoda ta wykorzystana została w amerykańskich urządzeniach Baseball, w angielskim układzie Phoenix, a także w radzieckim urządzeniu OGRA II. W urządzeniach tych uzyskano plazmę o koncentracji 10^8–10^9 cząstek/cm³ i maksymalnej temperaturze $T_j \approx 20$ keV (tj. 200 mln K).

Metody kompresji adiabatycznej

W celu podwyższenia temperatury i koncentracji plazmy, która została wytworzona w procesie jonizacji wstępnej lub iniekcji i znajduje się wewnątrz pułapki magnetycznej, można zastosować metodę kompresji adiabatycznej. Polega ona na ściśnięciu plazmy w wyniku przyłożenia dodatkowego rosnącego pola magnetycznego. Jeżeli zmiany pola i objętości plazmy następują w sposób adiabatyczny, następuje wówczas wzrost gęstości i temperatury plazmy (rys. 9).

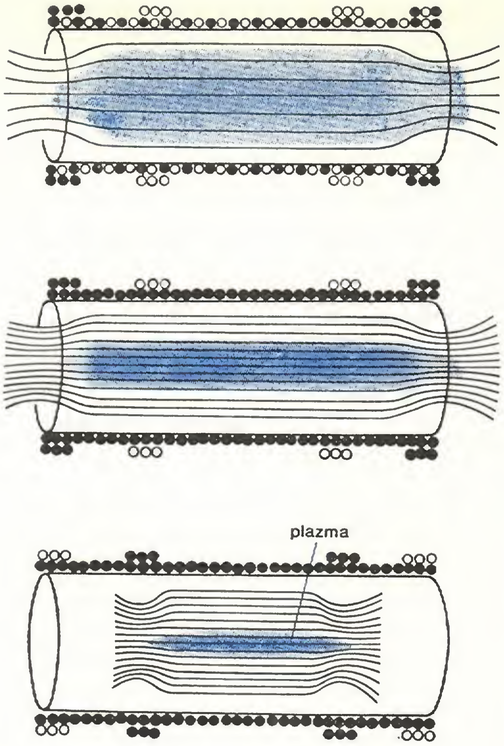
układ ATC

Metoda kompresji adiabatycznej stosowana była dotychczas głównie w układach typu otwartego, np. w pułapkach zwierciadlanych uzyskano tą metodą plazmę o koncentracji 10^{11} cząstek/cm³ i temperaturze $T_e \approx$ ≈ 10 keV. Ostatnio kompresja adiabatyczna stosowana jest coraz częściej także w pułapkach typu zamkniętego, np. w amerykańskim układzie toroidalnym ATC przy pomocy tej metody udało się uzyskać z plazmy o koncentracji $4 \cdot 10^{13}$ cząstek/cm³ i temperaturach $T_e \approx 1$ keV i $T_j \approx 200$ eV — plazmę o koncentracji $2 \cdot 10^{14}$ cząstek/cm³ oraz temperaturach $T_e \approx 2$ keV i $T_j \approx 600$ eV.

Inne metody grzania plazmy

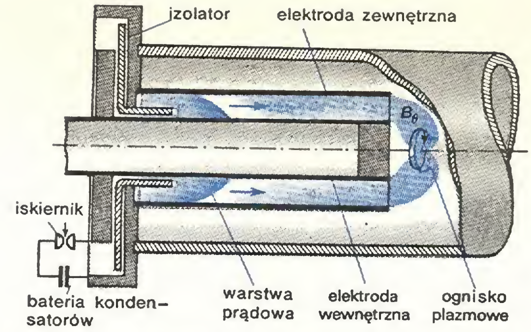
wykorzystanie fal uderzeniowych

Do nagrzewania plazmy wykorzystuje się również zjawiska, które występują przy rozchodzeniu się fal o bardzo dużej amplitudzie — tzw. fal uderzeniowych (→ Fale uderzeniowe). Do bardziej znanych zalicza się przy tym metodę grzania plazmy za pomocą fal uderzeniowych, których grubość frontu jest mniejsza niż długość drogi swobodnej cząstek plazmy. Metoda ta różni się od omówionej wyżej metody grzania turbulencyjnego tym, że nie wymaga wymuszania

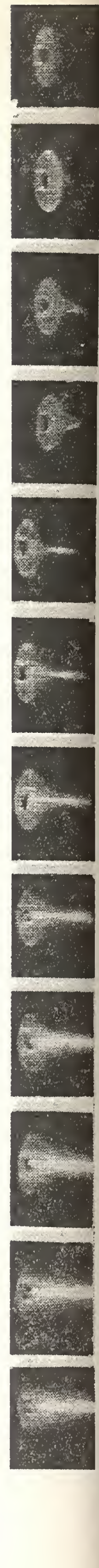


● uzwojenia, w których aktualnie płynie prąd
○ uzwojenia okresowo niezasilane

Rys. 9. Adiabatyczna kompresja plazmy przez narastające pole magnetyczne. Wzrostowi natężenia pola towarzyszy zmniejszenie obszaru zajmowanego przez plazmę oraz wzrost koncentracji i temperatury ściskanej plazmy. Rysunki przedstawiają kolejne fazy zjawiska



Rys. 10. Metoda „plasma focus". Mechanizm formowania ogniska plazmowego w pobliżu wylotu iniektora współosiowego. Warstwa zjonizowanego gazu, przez którą przepływa prąd wyładowania, po wyrzuceniu z obszaru iniektora ulega odkształceniu i radialnej kompresji. Na osi układu tworzy się wówczas obszar gęstej i bardzo gorącej plazmy. Obok seria zdjęć wylotu współosiowego iniektora plazmowego ukazująca różne fazy ogniska plazmowego



przepływu prądów wzdłuż linii sił pola magnetycznego. Brak jeszcze pełnej interpretacji teoretycznej tych zjawisk, eksperymentalnie jednak stwierdzono, że przy prędkościach znacznie większych od prędkości dźwięku w plazmie występuje anomalna lepkość, co prowadzi do grzania jonów. Przy odpowiednim doborze liczby Macha i innych parametrów można tą metodą wytwarzać plazmę o określonym stosunku temperatur T_e/T_i i bardzo wysokiej temperaturze elektronowej.

metody mieszane

W badaniach eksperymentalnych stosowane są bardzo często metody mieszane. Na przykład w niektórych układach zamkniętych wykorzystuje się najpierw grzanie omowe i rezonansowe, a następnie realizuje się kompresję adiabatyczną. Pozwala to poprawić parametry plazmy, zwiększyć jej koncentrację i temperaturę lub wydłużyć czas jej utrzymania.

Ograniczone możliwości omówionych wyżej sposobów wytwarzania plazmy wysokotemperaturowej powodują, że nadal poszukuje się nowych, bardziej efektywnych metod. W ostatnich latach szczególnie intensywnie rozwinęły się badania nad wyładowaniami typu plasma focus (ognisko plazmowe) oraz badania nad wytwarzaniem gorącej plazmy za pomocą laserów.

Wyładowanie typu plasma focus

Metoda oparta na wyładowaniach typu plasma focus jest właściwie odmianą metody iniekcyjnej, w której stosuje się iniektory współosiowe o specjalnej konstrukcji, przystosowane do pracy przy stosunkowo wysokim ciśnieniu gazu roboczego i dużych prądach wyładowania. W takich warunkach w obszarze międzyelektrodowym iniektora może uformować się wyraźna warstwa zjonizowanego gazu (tzw. warstwa prądowa), przez którą przepływa prawie cały prąd wyładowania. Pod wpływem sił elektrodynamicznych warstwa ta przesuwa się wzdłuż iniektora, a po dojściu do końca elektrod ulega wygięciu i radialnej kompresji (rys. 10).

Największa kompresja i nagrzanie plazmy występuje na osi symetrii układu, w pobliżu końca elektrody wewnętrznej. W obszarze tym powstaje „ognisko plazmowe", które w pewnych przypadkach może mieć bardzo małe rozmiary — średnicę rzędu 1 mm i długość rzędu kilku mm (efekt takiego ogniskowania ukazuje seria fotografii przedstawiona na rys. 10). W ognisku plazmowym można uzyskać plazmę o bardzo dużej koncentracji i bardzo wysokiej temperaturze. W wielu eksperymentach typu plasma focus uzyskuje się plazmę o koncentracji $5 \cdot 10^{19}$ cząstek /cm^3 i temperaturze rzędu 2,5 keV. W największych układach osiągane są rekordowe koncentracje 10^{20} cząstek/cm^3 i temperatury ok. 6 keV. Mimo że czas życia otrzymywanej w ten sposób plazmy jest bardzo krótki (10^{-8}–10^{-7} s), istnieją w niej warunki do wytwarzania neutronów z reakcji syntezy jądrowej. Dlatego podczas wyładowań typu plasma focus obserwuje się silne impulsy promieniowania neutronowego. Średnia wydajność neutronów jest przy tym w przybliżeniu proporcjonalna do kwadratu energii kumulowanej w układzie przed wyładowaniem. W największych działających obecnie układach typu plasma focus obserwuje się wytwarzanie 10^{10}–10^{12} neutronów na jedno wyładowanie.

Uzyskanie bardzo dużych koncentracji i temperatur plazmy, a także silnych impulsów neutronowych, spowodowało duży wzrost zainteresowania wyładowaniami plasma focus. Badania w tym kierunku prowadzone są w wielu krajach — w Stanach Zjednoczonych, w Związku Radzieckim, Wielkiej Brytanii, Republice Federalnej Niemiec i we Włoszech, a także w Polsce.

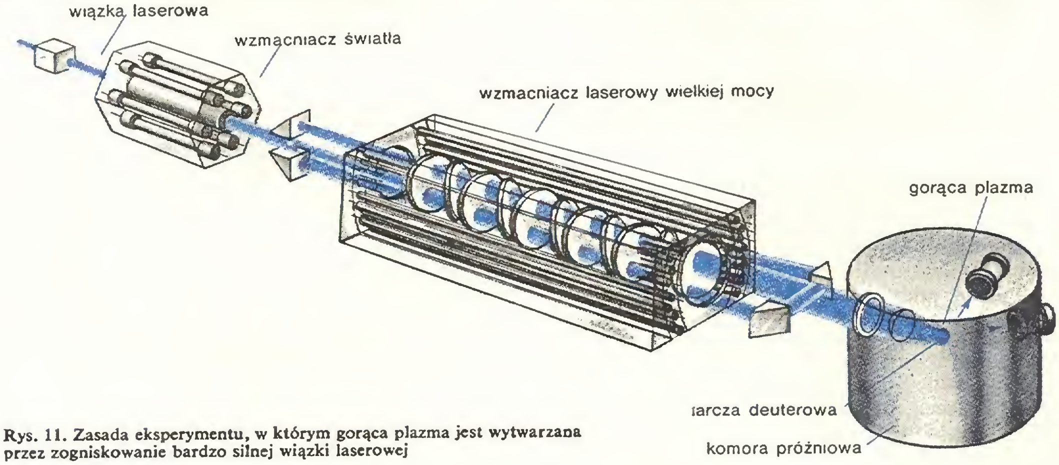
Metody laserowe

Kierunek badawczy, któremu poświęca się obecnie wiele uwagi, stanowią prace nad wytwarzaniem gorącej plazmy za pomocą laserów. Metody takie opierają się głównie na wykorzystaniu impulsowych laserów wielkiej mocy. Jeżeli wiązka z takiego lasera jest ogniskowana w obszarze gazowym lub na powierzchni ciała stałego i gęstość strumienia energii świetlnej w ognisku osiąga wartość 10^{10}–10^{12} W/cm^2, następuje proces jonizacji i rozwija się wyładowanie plazmowe. Zjawisko to może być wykorzystane do wytwarzania plazmy o wysokiej temperaturze, ponieważ część promieniowania lasera może być pochłaniana przez plaz-

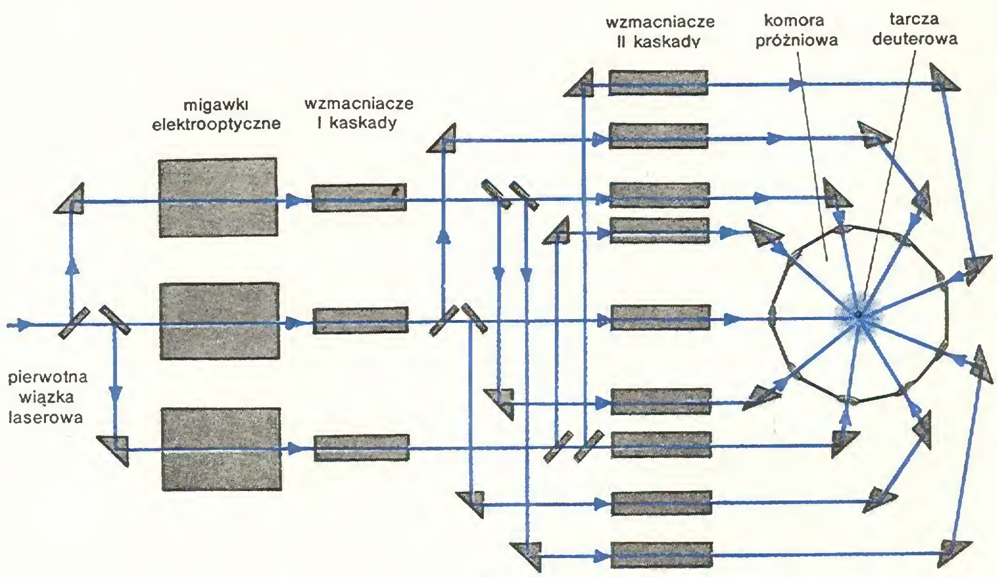
mę powstającą w czasie rozwoju wyładowania. Kiedy stopień jonizacji gazu w obszarze przed nagrzaną plazmą osiąga dostatecznie dużą wartość, nowa warstwa plazmy staje się nieprzepuszczalna dla rozpatrywanego promieniowania. W ten sposób strefa pochłaniania przemieszcza się w kierunku lasera. Proces ten przebiega bardzo szybko (z prędkością rzędu 10^7 cm/s) i uniemożliwia wydzielenie się całej energii impulsu w małym obszarze początkowego wyładowania, co utrudnia osiągnięcie bardzo wysokich temperatur.

Utrzymywanie gorącej plazmy

Plazmę jako zbiór cząstek naładowanych najłatwiej jest utrzymać w określonym obszarze przez otoczenie jej polem magnetycznym o odpowiedniej konfiguracji. Silne pola magnetyczne potrzebne do tego celu wytwarza się za pomocą elektrycznych prądów przepuszczanych przez specjalne uzwojenie zewnętrzne lub za pomocą prądów płynących przez badaną plazmę. Możliwe jest również połączenie obu tych sposobów.



Rys. 11. Zasada eksperymentu, w którym gorąca plazma jest wytwarzana przez zogniskowanie bardzo silnej wiązki laserowej



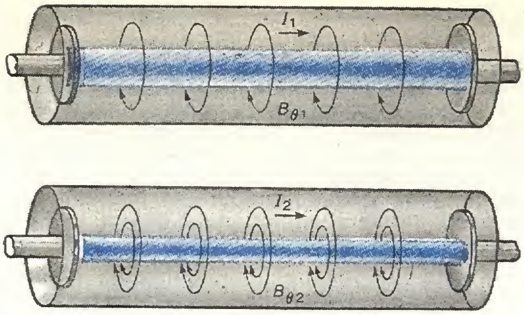
Rys. 12. Schemat eksperymentu laserowego, w którym wysokotemperaturowa plazma była wytwarzana przez 9 wiązek laserowych Wiązki te oświetlały jednocześnie małą kulkę deuteryzowanego polietylenu (o promieniu 0,25 mm). Łączna energia impulsów laserowych osiągała 1300 J, a czas ich trwania wynosił 2–16 ns, co zapewniało bardzo dużą moc promieniowania

Wielka moc współczesnych układów laserowych (10^{12}–10^{13} W przy czasie trwania impulsu 10^{-11}–10^{-9} s) umożliwia jednak skuteczną jonizację cienkich folii metalowych i organicznych, drobnych cząsteczek deuterku litu lub małych kulek zamrożonego wodoru (deuteru). W eksperymentach tego typu (rys. 11) przeciętnie na jedno wyładowanie uzyskuje się 10^{16}–10^{17} jonów o średniej energii 100–1000 eV. W celu zwiększenia energii i mocy dostarczanej do plazmy, a także dla uzyskania lepszej symetrii wyładowania największe eksperymenty laserowe realizowane są w układach wielowiązkowych (rys. 12). W eksperymentach tego typu uzyskuje się plazmę o rekordowej koncentracji $5 \cdot 10^{22}$ cząstek na 1 cm³ i temperaturze jonowej $T_j \approx$ ≈ 500 eV.

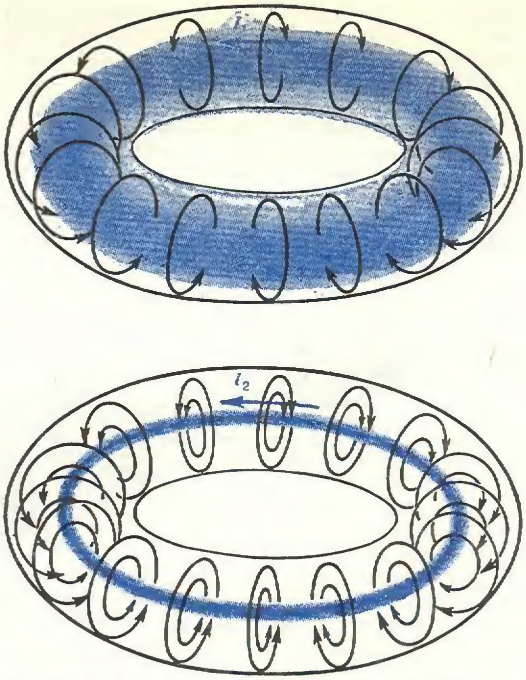
Zjawisko pinchu

Rozpatrzmy najpierw utrzymywanie plazmy za pomocą pola magnetycznego wytwarzanego przez prąd elektryczny, który przepływa wzdłuż osi cylindrycznej kolumny plazmowej. Pole magnetyczne, które towarzyszy przepływowi tego prądu, ma wówczas kierunek azymutalny. Oddziaływanie takiego pola z poosiowym prądem elektrycznym powoduje radialne ściśnięcie plazmy w cienki sznur plazmowy (rys. 13). Omawiane zjawisko nazywa się pinchem liniowym lub *Z*-pinchem (ang. *pinch* 'ściskać'). Wykorzystując zjawisko *Z*-pinchu zbudowano wiele urządzeń do badań plazmowych, np. przedstawiony na il. 47 (tabl. 12) Columbus II.

Z-pinch



Rys. 13. Zjawisko pinchu liniowego. Przepływowi prądu przez plazmę towarzyszy azymutalne pole magnetyczne. Zwiększenie natężenia prądu powoduje wzrost natężenia pola magnetycznego oraz ściśnięcie kolumny plazmowej w cienki sznur plazmowy



Rys. 15. Zjawisko pinchu toroidalnego. Wzrostowi indukowanego w plazmie prądu towarzyszy zwiększenie pola magnetycznego, które powoduje ściskanie pierścienia plazmowego

Ściśnięcie sznura plazmowego może być również wywołane przez szybko narastające zewnętrzne pole magnetyczne B_z. Takie pole można wytworzyć drogą rozładowania dużej baterii kondensatorów przez odpowiednio masywną cewkę. Natężenie prądu elektrycznego w takiej cewce może osiągnąć miliony amperów. W plazmie indukowane są wówczas silne prądy elektryczne I_z o kierunku azymutalnym, które w oddziaływaniu z zewnętrznym polem B_z wywołują ściśnięcie plazmy ku osi symetrii układu (rys. 14). Ze względu na kierunek prądów omawiany proces nazywa się theta-pinchem (θ-pinch). Wykorzystując zjawisko θ-pinchu zbudowano wiele urządzeń badawczych, m.in. amerykańskie układy Scylla oraz angielski układ Thetatron. Zdjęcie ilustrujące ściskanie plazmy w takim układzie przedstawia rys. 14b. W układach z θ-pinchem można wytwarzać plazmę o koncentracji 10^{16}–10^{17} cząstek/cm^3 i stosunkowo wysokich temperaturach $T_e \approx 0{,}3$–$1{,}5$ keV i $T_j \approx 4$ keV, ale czas jej utrzymywania jest bardzo krótki — rzędu 10 µs.

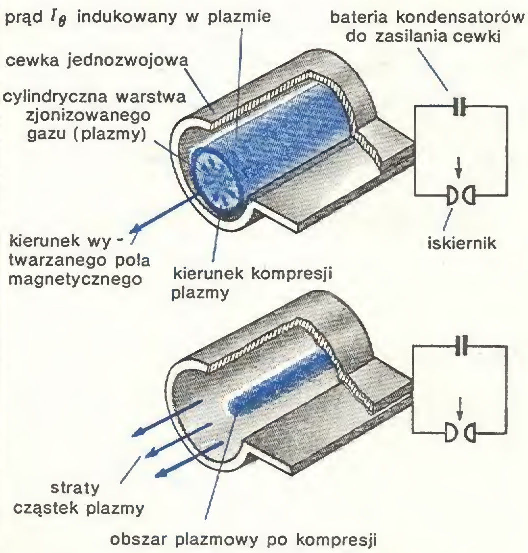
θ-pinch

Zasadniczą wadą omówionych wyżej układów typu Z-pinch lub θ-pinch jest krótki czas utrzymywania plazmy spowodowany szybką ucieczką cząstek naładowanych przez oba końce komory eksperymentalnej. Aby wyeliminować te straty w ostatnich latach podjęto badania θ-pinchu toroidalnego w układach zamkniętych w kształcie pierścienia (rys. 15). W zwykłej komorze toroidalnej na plazmę działają dodatkowe siły radialne, gdyż linie sił pola magnetycznego mają większą krzywiznę przy wewnętrznej stronie komory. Siły te dążą do przesunięcia sznura plazmy w stronę zewnętrznej ścianki torusa. Ich działanie można wyeliminować stosując specjalną konfigurację sznura plazmowego, w której wewnętrzna powierzchnia plazmowego torusa jest odpowiednio „pofałdowana". Siły naruszające stabilność można również skompensować w układzie toroidalnym, w którym indukowany jest dostatecznie silny prąd I_z. Pole magnetyczne pochodzące od tego prądu nakłada się wtedy na zewnętrzne pole magnetyczne i daje pole wypadkowe, którego linie sił owijają po linii śrubowej sznur plazmowy. Wyładowania tego typu noszą nazwę „screw-pinch". Największe urządzenie do badań θ-pinchu toroidalnego stanowi obecnie układ Scyllac (USA), w którym duża średnica toroidalnej komory próżniowej wynosi 4 m, a do wytwarzania pola magnetycznego wykorzystana jest bateria kondensatorów o energii 15 MJ.

θ-pinch toroidalny

screw-pinch





Rys. 14. Zjawisko θ-pinchu. a) Szybko narastające osiowe pole magnetyczne indukuje azymutalne prądy, które nagrzewają zjonizowany gaz, a następnie (pod wpływem działania sił elektrodynamicznych) powodują ściskanie plazmy w kierunku osi układu. b) Zdjęcia kolejnych faz zjawiska wykonane wzdłuż osi cylindrycznej komory eksperymentalnej za pomocą ultraszybkiej kamery fotograficznej w odstępach co 2 µs

Pułapki megnetyczne typu otwartego

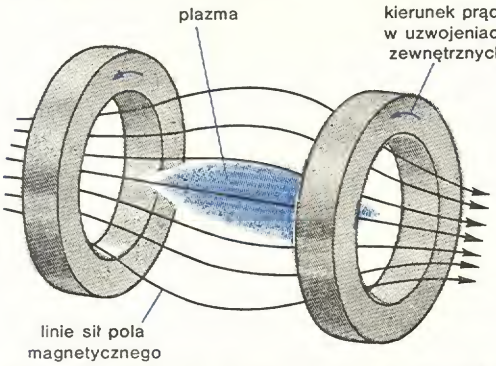
Do utrzymywania gorącej plazmy w pożądanym obszarze stosuje się również układy z kwazistacjonarnymi polami magnetycznymi — tzw. pułapki magnetyczne. Najprostszymi pułapkami magnetycznymi są pułapki zwierciadlane (ang. *mirror machines*; rys. 16). W urządzeniach tych stosuje się uzwojenia zewnętrzne wytwarzające osiowe pole magnetyczne słabsze w środkowej części komory, a silniejsze na obu jej końcach. Zagęszczające się na końcach komory linie sił pola magnetycznego tworzą wówczas tzw. zwierciadła magnetyczne, które odbijają część uciekających z plazmy cząstek z powrotem do wnętrza komory. Cząstki, które mają zbyt dużą składową prędkość wzdłuż osi komory, uciekają z układu.

pułapki zwierciadlane

W prostych pułapkach zwierciadlanych nie udało się jednak utrzymać plazmy o dużej gęstości, np. w największym tego typu układzie Ogra I (ZSRR) uzyskano plazmę o koncentracji $n \approx 10^8$ cząstek/cm^3 i czasie utrzymania $\tau \approx 0{,}3$ ms, a w układzie DCX II

(USA) plazmę o parametrach $n \approx 8 \cdot 10^9$ cząstek/cm³ oraz $\tau \approx 30$ ms. Inną wadą układów z pułapkami zwierciadlanymi było występowanie niestabilności magnetohydrodynamicznych. Niestabilności te można jednak wyeliminować przez zastosowanie pułapek typu „minimum-B", w których pole magnetyczne wzrasta we wszystkich kierunkach od środka układu, np. pułapka typu „karo" (ang. *cusp geometry*; rys. 17). Ze względu na duże straty cząstek czas utrzymywania

pułapki „minimum-B"

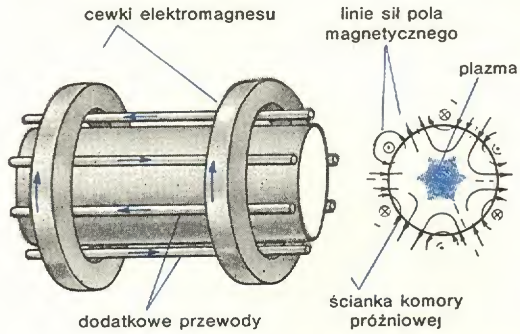


Rys. 16. Zasada budowy pułapki magnetycznej typu zwierciadlanego. Pole magnetyczne silniejsze na obu końcach komory powoduje zawracanie uciekających z plazmy cząstek naładowanych i działa jak zwierciadło magnetyczne

plazmy w pułapkach karo jest krótki — wynosi zwykle kilkadziesiąt µs. W układach tego typu można jednak metodą iniekcji wytworzyć plazmę o koncentracji 10^{13}–10^{14} cząstek/cm³, co w rezultacie daje porównywalne z pułapkami zwierciadlanymi wartości iloczynu $n \cdot \tau$. Pułapkami otwartymi typu minimum-B są również pułapki zwierciadlane z prętami stabilizacyjnymi (tzw. pułapki typu Joffe, rys. 18). Zbudowano szereg urządzeń eksperymentalnych tego typu, z których najbardziej znane są radzieckie Probkotrony. Osiągnięto w nich koncentrację plazmy rzędu $5 \cdot 10^9$ cząstek/cm³, temperaturę $T_j \approx 4$ keV oraz czas utrzymywania plazmy $\tau \approx 10$ ms, a więc znacznie dłuższy niż w układach typu karo. Pułapką tego typu był radziecki układ OGRA II (il. 50, tabl. 13).

Zmniejszając tylko liczbę prętów do 4 i łącząc odpowiednio ich końce, zbudowano pułapki minimum-B z uzwojeniami w kształcie szwu na piłce tenisowej. Pułapki takie zastosowane są w układach eksperymentalnych Baseball i Alice (USA), Phoenix (Anglia) oraz DECA II (Francja). W układach tego typu, stosując iniekcję wysokoenergetycznych cząstek neutralnych, uzyskano plazmę o koncentracji $2 \cdot 10^8$ cząstek/cm³ i czasie utrzymania 50 ms (Alice), a przy wykorzystaniu iniekcji strumieni plazmowych — plazmę o koncentracji 10^{13}–10^{14} cząstek/cm³, ale o krótszym czasie utrzymania 50 µs (DECA II).

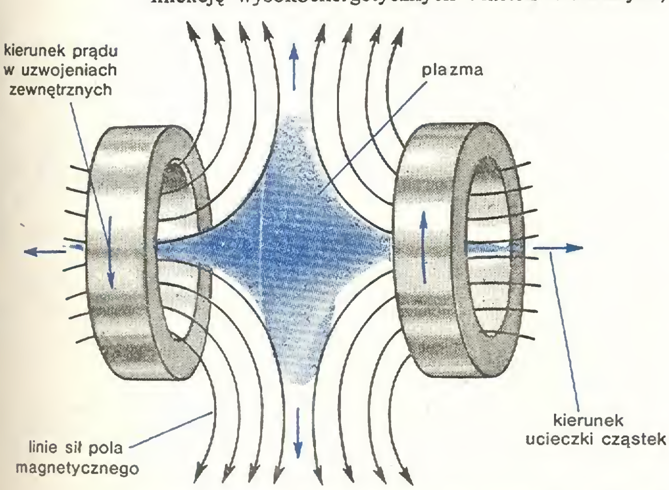
Omówione wyżej pułapki z minimum-B mają jednak istotną wadę; między równoległymi uzwojeniami stabilizacyjnymi występują liniowe szczeliny, przez które ucieka część cząstek plazmy. Wady tej pozba-



Rys. 18. Zasada budowy pułapki zwierciadlanej z prętami stabilizacyjnymi (pułapka typu Joffe). Prądy płynące przez przewody ułożone współosiowo na obwodzie cylindrycznej komory eksperymentalnej wytwarzają pola magnetyczne, które nakładają się na pole zwykłej pułapki zwierciadlanej. W rezultacie plazma wypełnia obszar pokazany w przekroju po prawej stronie rysunku

wiona jest pułapka typu SM (*Spherical Multipole*), którą opracowano w Instytucie Badań Jądrowych w Świerku. Pułapkę typu SM można zrealizować rozmieszczając symetrycznie na powierzchni kulistej komory próżniowej odpowiednio skierowane dipole magnetyczne, a w praktyce — silne elektromagnesy. Rozkład linii sił pola magnetycznego w takiej pułapce jest pokazany na ilustracji 51 (tabl. 13). Jak wynika z rozważań topologicznych i modelowych, w pułapce typu SM nie ma liniowych szczelin ucieczki plazmy, występuje natomiast pewna liczba wąskich stożków ucieczki. Przy odpowiednim doborze natężeń składowych pól magnetycznych wszystkim stożkom ucieczki w takiej pułapce może jednak odpowiadać kąt bryłowy mniejszy niż w innych pułapkach minimum-B. W przeprowadzonych dotychczas eksperymentach z pułapką SM uzyskano czas utrzymywania plazmy w przybliżeniu 3-krotnie dłuższy niż w konwencjonalnej pułapce typu karo. Prowadzi się obecnie badania w celu uzyskania jeszcze lepszych parametrów plazmy, np. przez wykorzystanie mieszanych pól magnetycznych — tzw. konfiguracji hybrydowych, oraz stosowanie różnych metod wytwarzania.

pułapka SM



Rys. 17. Zasada budowy pułapki magnetycznej typu karo. Przeciwnie skierowane pola magnetyczne tworzą obszar, który ma 2 stożki ucieczki oraz pierścieniową szczelinę ucieczki w płaszczyźnie symetrii układu. W środku tego obszaru występuje minimum pola magnetycznego (minimum-B)

Pułapki magnetyczne typu zamkniętego

W ostatnich latach szczególnie intensywnie prowadzone były badania nad pułapkami magnetycznymi typu zamkniętego, których typowymi przykładami są układ toroidalny przedstawiony na rys. 19 oraz angielski układ ZETA.

Wadą prostych układów toroidalnych jest występowanie radialnych nieskompensowanych gradientów pola magnetycznego, które wywołują przemieszczanie cząstek naładowanych w poprzek linii sił pola ograniczającego (ruchy dryfowe) i prowadzą do niestabilności plazmy. Ruchy dryfowe można ograniczyć przez odpowiednie skręcenie linii sił pola magnetycznego, np. przez nadanie komorze kształtu ósemki lub przez zastosowanie skręconych uzwojeń ułożonych na obwodzie komory toroidalnej. Koncepcję tę wykorzystano budując amerykańskie układy toroidalne, które otrzymały nazwę stellaratorów (rys. 20), szereg urządzeń w Anglii, we Francji i NRF, a także radziecki stellarator Uragan (il. 48, tabl. 13).

układy toroidalne

stellaratory

tokamaki

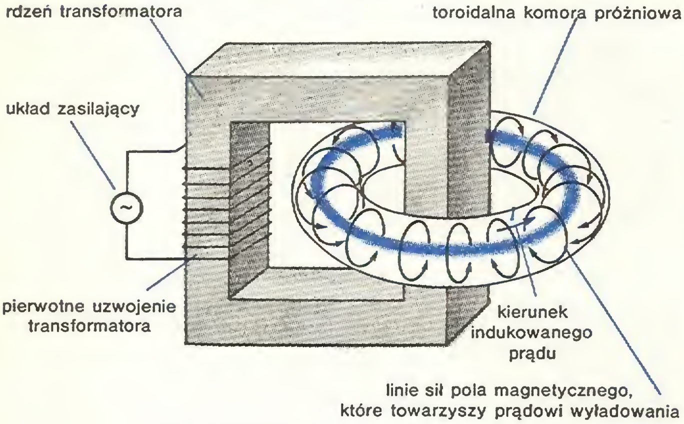
Innym rodzajem pułapek zamkniętych są układy toroidalne typu tokamak (il. 49, tabl. 13), których zasadę budowy ilustruje rys. 22. W odróżnieniu od omówionych wyżej prostych układów toroidalnych (typu ZETA), w układach typu tokamak pole stabilizujące B_z jest znacznie silniejsze od pola B_φ.

Wielką zaletą omówionych wyżej pułapek toroidalnych jest brak możliwości ucieczki cząstek naładowanych wzdłuż linii sił pola magnetycznego. W pułapkach tego typu, oprócz zwykłej dyfuzji cząstek spowodowanej zderzeniami binarnymi (dwucząstkowymi), występuje jednak dyfuzja anomalna, która powoduje znaczne straty cząstek i zmniejsza czas życia plazmy. W ostatnich latach udowodniono jednak, że można stworzyć warunki, w których uzyskuje się znacznie dłuższy czas życia plazmy. W eksperymentach przeprowadzonych w stellaratorze Wendelstein w Garching (NRF) otrzymano czas utrzymania $\tau \approx 100\tau_B$ — (gdzie τ_B — czas życia plazmy obliczony teoretycznie), a w urządzeniach tokamak zbudowanych w Instytucie Energii Atomowej w Moskwie utrzymano plazmę o koncentracji (3–6)$\cdot 10^{13}$ cząstek/cm^3, temperaturach $T_e \approx 1000$ eV (12 mln K) i $T_j \approx 700$ eV (8 mln K) przez czas $\tau = 30\tau_B$ (ok. 70 ms).

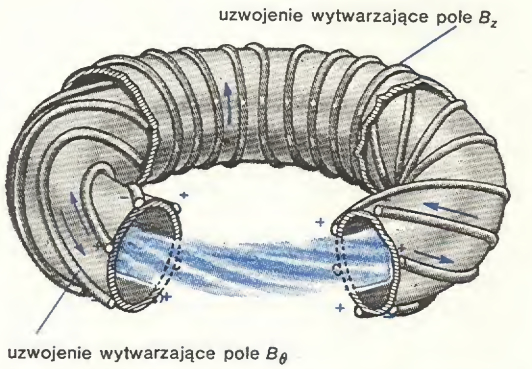
układ T-10

W Moskwie zbudowano bardzo wiele układów typu tokamak, z których największy obecnie jest układ T-10. Duży promień toroidalnej komory tego układu wynosi 1,5 m, maksymalne pole magnetyczne $B_z = 5$ T, a prąd wyładowania $I_{max} = 500$ kA. Parametry plazmy otrzymywanej w tokamakach są tak dobre, że nawet dla oszczędności czasu i środków w Princeton (USA) przebudowano na tokamak największy stellarator amerykański — Model C (il. 49, tabl. 13). Podjęta została również budowa nowych tokamaków w USA, Anglii, Francji, NRF oraz w Japonii. Część tych układów jest obecnie budowana w ramach współpracy międzynarodowej.

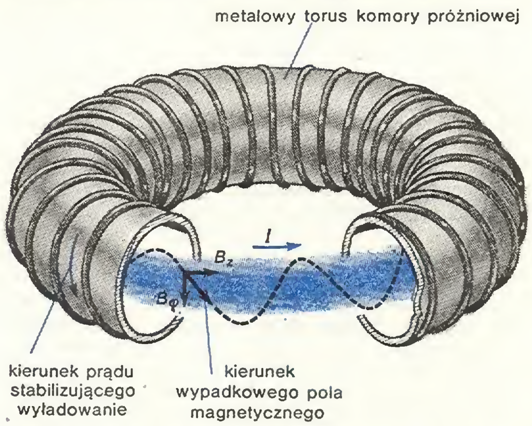
Równocześnie z pracami nad ulepszeniem pułapek toroidalnych wzrosło także zainteresowanie innymi pułapkami typu zamkniętego. Badania różnych układów zamkniętych wykazały jednak, że spełnienie warunków minimum-B jest topologicznie niemożliwe. W związku z tym zwrócono uwagę na możliwość za-



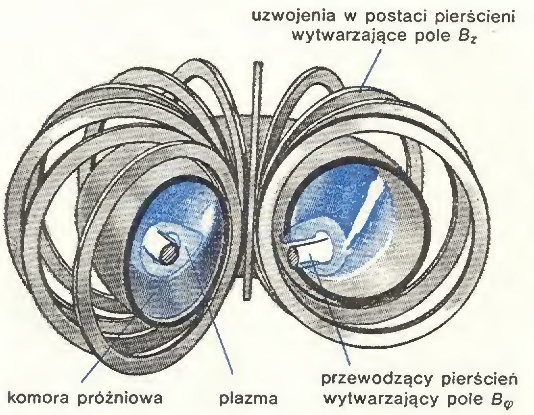
Rys. 19. Zasada budowy zamkniętych pułapek magnetycznych typu toroidalnego. Plazma w komorze eksperymentalnej stanowi wtórne uzwojenie transformatora, co umożliwia indukowanie prądów powodujących wytwarzanie i nagrzewanie plazmy



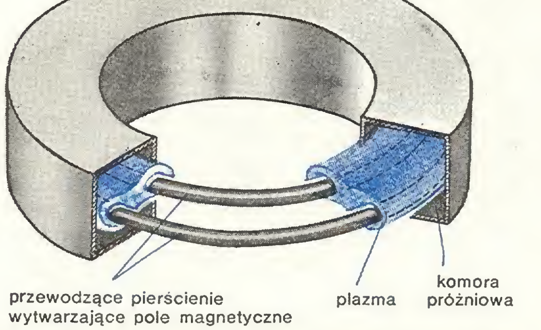
Rys. 20. Zasada budowy pułapek typu stellarator. Dodatkowe uzwojenia powodują skręcenia linii sił wypadkowego pola magnetycznego i wpływają stabilizująco na wyładowanie plazmowe. Na rysunku pominięto rdzeń transformatora, który się stosuje zwykle do zasilania wyładowań



Rys. 22. Zasada budowy pułapek typu tokamak. Sznur plazmowy utrzymywany jest polem magnetycznym B_φ pochodzącym od prądu wyładowania. Do stabilizacji wykorzystuje się silne pole magnetyczne B_z, wytwarzane przez cewki umieszczone na obwodzie toroidalnej komory eksperymentalnej. Linie sił wypadkowego pola magnetycznego są silnie skręcone



Rys. 21. Zasada budowy pułapki Levitron. Wewnątrz toroidalnej komory otoczonej uzwojeniami stabilizującymi umieszczony jest pierścień z prądem, który wytwarza dodatkowe pole magnetyczne, stabilizujące plazmę



Rys. 23. Zasada budowy pułapki typu kwadrupola toroidalnego. Plazma jest utrzymywana głównie przez pole magnetyczne pochodzące od prądów płynących w dwóch pierścieniowych uzwojeniach wewnętrznych. W analogiczny sposób można zbudować oktupol toroidalny lub układ jeszcze wyższego rzędu

warunek average minimum-B

pewnienia „średniego" minimum-B wzdłuż określonych torów ruchu cząstek plazmy (czyli tzw. average minimum-B). Spełnienie warunku average minimum-B można osiągnąć nie tylko w stellaratorach i tokamakach, ale również w pułapkach typu Levitron lub spherator, w których wewnątrz toroidalnej (lub elipsoidalnej) komory umieszczone jest dodatkowe pierścieniowe uzwojenie z odpowiednio silnym prądem. Pole magnetyczne B_φ pochodzące od tego prądu nakłada się na pole toroidalne B_z, wywołując „skręcenie" linii sił pola wypadkowego. W rezultacie plazma utrzymywana jest wokół pierścieniowego uzwojenia wewnętrznego (rys. 21). Inną odmianę pułapek zamkniętych, spełniających warunek average minimum-B, stanowią pułapki toroidalne quadrupolowe lub oktupolowe, w których wewnątrz toroidalnej komory umieszcza się odpowiednio 2 lub 4 pierścienie z prądem (rys. 23). W pułapkach tego typu plazma zajmuje obszar wokół i pomiędzy przewodzącymi pierścieniami, które zapewniają najlepszą stabilność magnetohydrodynamiczną.

Porównując różne pułapki typu zamkniętego można stwierdzić, że największy postęp w technice otrzymywania i utrzymywania gorącej plazmy został dotychczas osiągnięty w urządzeniach tokamak, dlatego też wysiłek wielu ośrodków badawczych skierowany jest obecnie na budowę i badanie dużych układów typu tokamak. Równocześnie prowadzi się badania w różnych innych kierunkach.

Perspektywy dalszych osiągnięć

Dotychczasowe osiągnięcia w dziedzinie badań termojądrowych ujmuje wykres, na którym podane są podstawowe parametry wytwarzanej plazmy: jej temperatura, gęstość oraz czas utrzymywania (rys. 24). Wykres umożliwia porównanie wyników, które udało się osiągnąć w różnych układach eksperymentalnych przy wykorzystaniu wyżej wymienionych metod wytwarzania i nagrzewania plazmy. Na szczególną uwagę zasługują rekordowe wartości iloczynu koncentracji i czasu utrzymania plazmy ($n\tau \approx 10^{13}$ cm^{-3} · s), które osiągnięto w ostatnich latach w niektórych urządzeniach typu plasma focus oraz w eksperymentach z laserem.

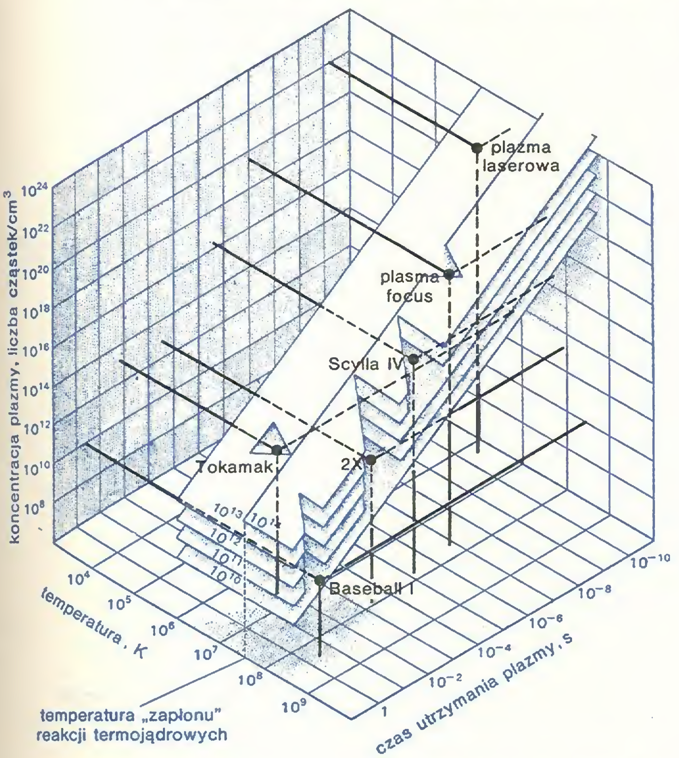
Należy przy tym zauważyć, że nie ma zasadniczych trudności z wytworzeniem jonów o jeszcze większych energiach i uzyskaniem jeszcze wyższych temperatur. Poważne trudności związane są natomiast z utrzymywaniem plazmy w izolacji w tak wysokich temperaturach i dlatego ciągle poszukuje się nowych bardziej efektywnych metod.

Potrzeba rozwoju badań termojądrowych zintensyfikowała postęp w technice i technologii. W dziedzinie techniki wysokiej próżni zbudowano np. układy pompujące o próżni granicznej rzędu 10^{-7} Pa i wydajności wielu tysięcy l/s oraz opanowano technologię wysokopróżniowych materiałów konstrukcyjnych, potrzebnych do budowy przyszłego reaktora termojądrowego. Bardzo duży postęp nastąpił także w dziedzinie budowy elektromagnesów, zwłaszcza nadprzewodzących, potrzebnych do utrzymywania gorącej plazmy. Wiele wysiłku włożono również w rozwój techniki wysokich napięć.

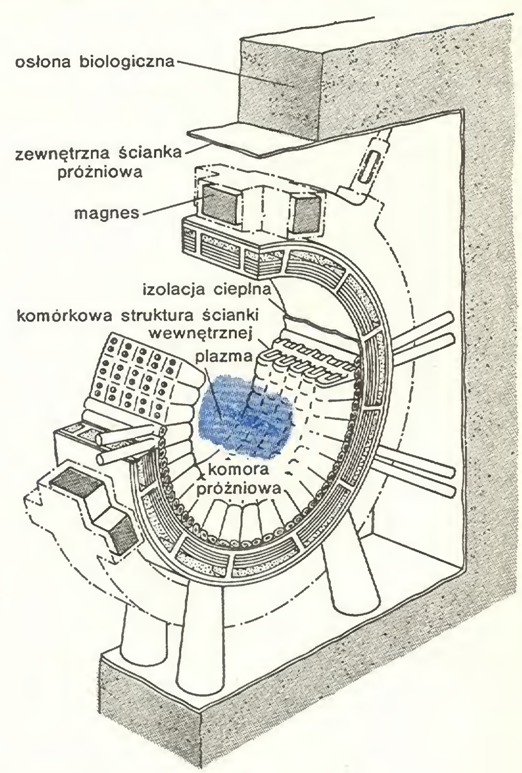
Zbudowano generatory udarowe, zdolne do kumulowania energii do kilkunastu milionów dżuli i wytwarzania impulsów prądowych o amplitudzie do kilku milionów amperów.

technologia reaktorów termojądrowych

W chwili obecnej istnieje już wiele wstępnych projektów technicznych reaktorów termojądrowych o różnych zasadach działania. Projekty te opierają się zwykle na założeniu, że w przyszłym reaktorze termojądrowym komora próżniowa, która zawierać będzie gorącą plazmę, otoczona zostanie grubą warstwą moderatora neutronów. Badania nad różnego typu materiałami wykazały, że ścianki takiej komory



Rys. 24. Wykres zestawiający najlepsze wyniki, uzyskane w różnych urządzeniach do badań termojądrowych. Zaznaczone na rysunku płaszczyzny odpowiadają określonym wartościom iloczynu koncentracji n i czasu utrzymania plazmy τ. Warunkiem uzyskania dodatniego bilansu energetycznego reakcji termojądrowej w mieszaninie deuter-tryt o temperaturze powyżej 45 mln K jest $n\tau \geqslant 10^{14}$ s/cm^3



Rys. 25. Projekt konstrukcji jednego z sektorów reaktora termojądrowego z komorą toroidalną o średnicy zewnętrznej rzędu kilkunastu metrów

mogą być wykonane z niobu i molibdenu lub niobu i wanadu, a moderator — z czystego litu, ze stopu lit-beryl lub z eutektycznego związku $(LiF)_2BeF_2$. Umieszczone na zewnątrz moderatora uzwojenia elektromagnesów mogą być natomiast wykonane z materiałów nadprzewodzących pracujących w odpowiednio niskich temperaturach. Poza tym, przyszły reaktor termojądrowy musi mieć oczywiście odpowiednią konstrukcję wspierającą i specjalne osłony biologiczne. Schemat ideowy takiego reaktora przedstawiono na rys. 25.

Z analizy różnych możliwych konfiguracji wynika, że dla przyszłego reaktora termojądrowego korzystna byłaby konfiguracja typu tokamak (lub stellarator), o ile uda się uzyskać odpowiednio długi czas utrzymywania plazmy i zwiększyć osiągane wartości iloczynu $n\tau$. Obliczenia wykazały również możliwość wykorzystania konfiguracji typu otwartego (np. pułapek zwierciadlanych), pod warunkiem, że uda się opracować odpowiednio wydajny system iniekcyjny oraz ograniczyć straty cząstek naładowanych. Istnieją także projekty impulsowych reaktorów termojądrowych opartych na wykorzystaniu metod laserowych.

Wszystkie przeprowadzone dotychczas obliczenia techniczne wskazują na to, że przyszłe reaktory termojądrowe będą prawdopodobnie miały stosunkowo duże rozmiary (promień samej komory próżniowej — rzędu kilku do kilkunastu metrów). Moc tych reaktorów będzie prawdopodobnie osiągać wartości od kilku do kilkunastu gigawatów.

Można śmiało powiedzieć, że ludzkość jest już bardzo blisko zdobycia nowego niewyczerpalnego źródła energii. Na drodze do tego celu piętrzą się jednak jeszcze poważne trudności, tak że nie sposób dokładnie określić termin uruchomienia pierwszego reaktora termojądrowego.

L.A. Arcimowicz *Czwarty stan materii*, Warszawa 1972; J. Crussard *Energia termojądrowa*, Warszawa 1967; A.W. Czernietski *Wstęp do fizyki plazmy*, Warszawa 1971; D.A. Frank-Kamieniecki *Plazma — czwarty stan materii*, Warszawa 1963; W.F. Kalinin *Termojądrowy reaktor przyszłości*, Warszawa 1968; S. Kaliski *Lasery — synteza termojądrowa*, Warszawa 1975; J.G.Linhart *Fizyka plazmy*, Warszawa 1963; M. Sadowski *Świat wysokich temperatur*, Warszawa 1975; W. Zonn *Astrofizyka ogólna*, Warszawa 1955.

Radioizotopy

Lech Stolarczyk

Jądro atomowe jest układem złożonym z protonów i neutronów silnie ze sobą związanych siłami jądrowymi. Jądro o określonych liczbach protonów p i neutronów n (liczba atomowa Z i liczba masowa A) może znajdować się w jednym z możliwych dla niego kwantowych stanów energetycznych. Najtrwalszym stanem danego jądra (czyli inaczej nuklidu) jest stan o najniższej możliwej energii, czyli tzw. stan podstawowy. Jądra w stanie podstawowym bez ingerencji z zewnątrz albo trwają niezmienione — są to izotopy trwałe, albo z określonym prawdopodobieństwem ulegają samorzutnej reakcji (przemianie) jądrowej — są to izotopy promieniotwórcze, inaczej radioizotopy (→ Rozpady jąder). Samorzutna przemiana radioizotopów w izotopy lub radioizotopy (na ogół innego pierwiastka) zachodzi przez emisję cząstek elementarnych lub jąder (np. helu). Najważniejszymi samorzutnymi przemianami są: emisja β^+, β^-, α, wychwyt K oraz samorzutna reakcja rozszczepienia. Rozszczepieniu ulegają tylko jądra najcięższe, nigdy nie jest to proces dominujący (np. dla izomeru kalifornu $^{252m}_{98}Cr$ stanowią one 3,1% samorzutnych reakcji jądrowych, a dla uranu ^{238}U — ok. 10^{-4} %).

nuklid

Jądra powstające w reakcjach jądrowych są bardzo często w stanie wzbudzonym, mogą one ulec przemianie jądrowej albo w jednym lub kilku etapach wypromieniować energię wzbudzenia w postaci fotonów (promieniowanie γ) i przejść do stanu podstawowego. Proces emisji promieniowania γ, zwany przemianą γ, jest na ogół bardzo szybki. Zwykle zachodzi on z prawdopodobieństwem bliskim 1 w ciągu czasu ok. 10^{-15} s. Zdarzają się jednak takie stany wzbudzone, w których przemiana γ jest znacznie wolniejsza (mniej prawdopodobna), i wtedy jądro może trwać w takim stanie znacznie dłużej. Na przykład izotop bizmutu ^{210}Bi, trwały w stanie podstawowym, posiada taki stan wzbudzony, dla którego nawet po upływie milionów lat prawdopodobieństwo przejścia do stanu podstawowego jeszcze znacznie różni się od 1. Takie względnie trwałe stany wzbudzone izotopów nazywamy stanami metatrwałymi. Jądra tego samego izotopu znajdujące się w różnych stanach (podstawowym i metatrwałym) nazywamy izomerami. Izomery różnią się nieco masą. Izomeria jądrowa nie jest zjawiskiem częstym, a tylko kilka izotopów ma więcej niż 2 izomery. Jądra w stanie wzbudzonym, w szczególności w stanie metatrwałym, są mniej trwałe od jąder w stanie podstawowym; mogą one ulegać takim reakcjom, które byłyby niemożliwe, gdyby jądra były w stanie podstawowym.

przemiana γ

stan metatrwały

izomer

Izotopy i radioizotopy oznaczamy symbolami literowymi odpowiedniego pierwiastka chemicznego dopisując przed tymi symbolami u dołu liczbę Z, a u góry — liczbę A. W skróconym zapisie opuszczamy liczbę Z, ponieważ jest ona zaszyfrowana w symbolu chemicznym. Wyjątkowo izotopy wodoru zapisujemy często innymi literami niż jego symbol chemiczny H: D od deuter i T od tryt. Symbole, w których po liczbie masowej dopisuje się literę m oznaczają izomery w stanie wzbudzonym metatrwałym (np. ^{210}Bi to stan podstawowy, a ^{210m}Bi to wspomniany, wyjątkowo trwały izomer wzbudzony). W wypadku izomerii wielokrotnej oraz na oznaczenie innych niż metatrwałe stanów wzbudzonych posługujemy się symbolami izotopu z odpowiednimi dopiskami, znakami (np. gwiazdka za symbolem) itp.

Dobrą miarą trwałości izotopu jest okres połowicznego rozpadu $t_{1/2}$, czyli czas, w ciągu którego połowa jąder danego izotopu (izomeru) ulega rozpadowi. Dla izotopów trwałych $t_{1/2} = \infty$. Im trwalszy radioizotop, tym mniejsze prawdopodobieństwo zajścia reakcji i tym dłuższe jest $t_{1/2}$. Radioizotopy o bardzo długich $t_{1/2}$ (np. dłuższych od 10^{17} lat) są niewykrywalne. Podział izotopów na izotopy trwałe i radioizotopy jest nieostry, zależny od dokładności metod pomiarowych. Istnieje nieliczna grupa izotopów, których nietrwałości nie można ani teoretycznie wykluczyć, ani doświadczalnie potwierdzić (zob. rys. 5, str. 221). Praktycznie granicą doświadczalnego zidentyfikowania nuklidu jako radioizotopu jest jego trwałość odpowiadająca nie krótszemu $t_{1/2}$ niż 10^{-6} s. Granica ta nie jest oczywiście ostra. W pewnym sensie umownie możemy więc radioizotopy określić jako nuklidy mające w stanie podstawowym (a w wypadku izomerów — w stanie metatrwałym) wartości $t_{1/2}$ leżące w przedziale od 10^{-6} s do 10^{17} lat.

okres połowicznego rozpadu

Jednak nawet w grupie izotopów trwałych możemy mówić o zróżnicowanej trwałości jąder. Miarą tej trwałości jest energia wiązania E_n przypadająca na jeden nukleon. Najtrwalszymi jądrami są jądra z liczbą masową $25 \leqslant A \leqslant 150$, mające największą energię E_n. Jądra o parzystych liczbach p i n (parzysto-parzyste) są na ogół trwalsze od jąder o jednej z tych liczb nieparzystej (nieparzysto-parzystych lub parzysto-nieparzystych), a jądra nieparzysto-nieparzyste są zawsze

energia wiązania

znacznie mniej trwałe. W tej ostatniej grupie spotykamy tylko cztery izotopy trwałe (D, ^{6}Li, ^{10}B oraz ^{14}N). Pierwiastki o Z parzystym mają często po kilka trwałych izotopów i wiele radioizotopów, a pierwiastki o Z nieparzystych tworzą znacznie mniej nuklidów o porównywalnej trwałości. Dla dużych liczb masowych energia E_n maleje do zera i tak dla $Z > 83$ nie ma już izotopów trwałych, a dla $Z \leqslant 83$ jedynie technet (nieparzyste $Z = 43$) nie ma żadnego trwałego izotopu.

Metody radiometryczne pomiaru radioizotopów

Najczęściej w pomiarach ilościowych wykorzystuje się zdolność cząstek naładowanych i promieni γ do jonizacji ośrodka, do którego cząstka i promienie trafiają (metody jonizacyjne). Cząstka α lub β przekazuje wzdłuż toru swego lotu energię cząsteczkom ośrodka, a część tej energii zużywa się na wytworzenie jonów. Fotony (a więc promienie γ) mogą całą swą energię lub jej część przekazać jakiemuś elektronowi ośrodka (zjawisko fotoelektryczne lub efekt Comptona), który wyrwany ze swego położenia przejmuje tę energię jako energię kinetyczną. W wypadku fotonów o energii większej od 1,022 MeV możliwe jest zjawisko tworzenia par, w którym foton znika, a powstaje para negaton-pozyton. Masa spoczynkowa powstałych cząstek równoważna jest energii 1,022 MeV, a nadwyżka energii fotonu ponad tę wartość pojawia się jako energia kinetyczna tych cząstek. Wytworzone cząstki jonizują ośrodek wzdłuż swych torów, a spowolniony pozyton anihiluje następnie z którymś z elektronów (negatonów) ośrodka — obie cząstki znikają, a na ich miejsce pojawiają się dwa fotony o energii 0,511 MeV. Fotony te mogą dalej oddziaływać z ośrodkiem. We wszystkich więc wypadkach ośrodek ulega jonizacji wzdłuż toru pierwotnych lub wtórnych cząstek.

oddziaływanie α, β lub γ z materią

Jeżeli jonizacja wywołana przez promienie α, β lub γ zajdzie w komorze z rozrzedzonym gazem i dołączonymi elektrodami, to przykładając do elektrod odpowiednie napięcie możemy „wyłapać" na elektrodach cały ładunek wytworzonych jonów (liczniki proporcjonalne). Ponieważ każdy rodzaj promieniowania musi przekazać ośrodkowi określoną energię na wytworzenie określonej liczby (ładunku) jonów, liczniki proporcjonalne pozwalają zmierzyć energię promieniowania przekazaną ośrodkowi. Gdy promieniowanie jest bardzo słabe, pomiar taki staje się jednak niemożliwy. Można wtedy przyłożyć do elektrod znacznie wyższe napięcie, tak aby wytwarzane przez promieniowanie jony ulegały przyspieszeniu w polu elektrycznym w komorze i wywoływały jonizacje wtórne (liczniki Geigera-Müllera — GM). Każde trafienie cząstki jonizującej w licznik spowoduje w tych warunkach przepływ prądu trwający od 1 µs do 1 ms (zależnie od konstrukcji licznika). Licznik GM umożliwia zliczanie cząstek (lub fotonów) jonizujących trafiających w licznik.

licznik proporcjonalny

licznik Geigera–Müllera

Zastąpienie gazu w komorze półprzewodnikiem znacznie rozszerza możliwości badawcze, a odpowiednia konstrukcja detektorów (komór, liczników) promieniowania umożliwia tworzenie zestawów z detektorów i współpracujących z nimi urządzeń elektronicznych (tabl. 14, il. 53). Zmiany te umożliwiają ilościową i jakościową analizę promieniowania, czyli określenie ilości i energii cząstek emitowanych przez radioizotopy.

Metody radiometryczne cechuje wysoka czułość i duża dokładność, dzięki czemu można szybko identyfikować i określać ilościowo różne radioizotopy. Warunkiem wykrywalności radioizotopu jest takie jego stężenie, aby w próbce zachodziło przynajmniej kilka rozpadów w ciągu sekundy. Izotop ^{226}Ra o $t_{1/2} = 1590$ lat można wykryć przy zawartości w próbce 10^{-9} g; gdy radioizotop ma $t_{1/2}$ krótszy — wykrywalna ilość substancji jest znacznie mniejsza.

warunek wykrywalności radioizotopu

Otrzymywanie radioizotopów

Znanych izotopów jest blisko 1800, a tylko ok. 270 to izotopy trwałe, resztę stanowią radioizotopy. Ale tylko nieliczne z nich występują w przyrodzie. W dużych skupiskach spotyka się tylko radioizotopy o długich $t_{1/2}$, należące do pokazanych na rys. 1 rodzin promieniotwórczych. Towarzyszą im w bardzo małych stężeniach krótkożyjące radioizotopy z tych rodzin. Wyodrębnienie poszczególnych radioizotopów ze źródeł naturalnych jest zawsze bardzo kosztowne i rzadko opłacalne. Wiele różnych radioizotopów, ale w bardzo małych stężeniach, powstaje w reakcjach jądrowych inicjowanych przez promieniowanie kosmiczne, np. izotop węgla ^{14}C występuje w atmosferycznym dwutlenku węgla CO_2 w stężeniu ok. 10^{-10} %. Izotop ^{14}C można wykryć w jeszcze znacznie mniejszym stężeniu, jednak otrzymywanie go ze środowiska naturalnego jest niemożliwe, a ściśle mówiąc byłoby ono niesłychanie kosztowne. Powszechnie w środowisku naturalnym występują też w małych stężeniach i inne radioizotopy o długim $t_{1/2}$, np. izotop potasu ^{40}K.

Przeprowadzane wybuchy jądrowe wprowadziły do środowiska naturalnego znaczne ilości różnych innych radioizotopów, z których trwalsze stanowią nawet poważne zagrożenie, głównie ze względu na wywołane ich promieniowaniem przyspieszanie degeneracyjnych mutacji, wzrost zachorowań na białaczkę i in. I te radioizotopy nie mogą jednak być otrzymywane ze środowiska naturalnego.

Aktywacja neutronowa

Ogromna większość wykorzystywanych radioizotopów jest celowo wytwarzana w specjalnie prowadzonych, kontrolowanych reakcjach jądrowych (→ Energia jądrowa). Z punktu widzenia techniki wytwarzania radioizotopów w większych ilościach najważniejsze są reakcje przebiegające w wyniku trafienia w jądro neutronem (aktywacja neutronowa) oraz reakcje należące do szeregu konsekutywnych procesów samorzutnych, inicjowanych takim trafieniem. Reakcje te przeprowadza się na wielką skalę wprowadzając do kanałów reaktora jądrowego różne substancje, zwane ogólnie materiałami tarczowymi.

aktywacja neutronowa

W reakcji $^{59}Co + n \rightarrow {}^{60}Co$ otrzymuje się z trwałego izotopu kobaltu ^{59}Co radioaktywny izotop ^{60}Co, stosowany na wielką skalę jako źródło użytecznego promieniowania γ. Radioizotopu ^{60}Co nie wyodrębnia się z masy macierzystego izotopu trwałego, ale wykorzystuje mieszaninę izotopów tym bogatszą w radioizotop, im większe było natężenie strumienia neutronów w reaktorze i im dłużej materiał tarczowy przebywał w kanałach reaktora. Podobnie w reakcji: ${}^{14}_{7}N + {}^{1}_{0}n \rightarrow {}^{15}_{7}N^{*} \rightarrow {}^{14}_{6}C + {}^{1}_{1}H$ powstaje promieniotwórczy izotop węgla, który można chemicznie oddzielić od macierzystego materiału tarczowego i przekształcić w potrzebny związek chemiczny.

otrzymywanie kobaltu ^{60}Co

W analogicznych reakcjach otrzymuje się z trwałych izotopów ${}^{35}_{17}Cl$ oraz ${}^{32}_{16}S$ cenne radioizotopy ${}^{35}_{16}S$ i ${}^{32}_{15}P$. Wszystkie wymienione radioizotopy emitują promieniowanie β^-, któremu w wypadku ^{60}Co towarzyszy wtórna emisja dwóch fotonów γ o energiach 1,17 i 1,33 MeV emitowanych przez wzbudzony ^{60}Ni.

Materiał tarczowy, w którym w kanałach reaktora powstały znaczne ilości radioizotopów, jest potężnym źródłem promieniowania i bezpośrednie zetknięcie się z nim byłoby zabójcze dla każdej istoty żywej. Dlatego wydobywanie tego materiału oraz jego dalsza

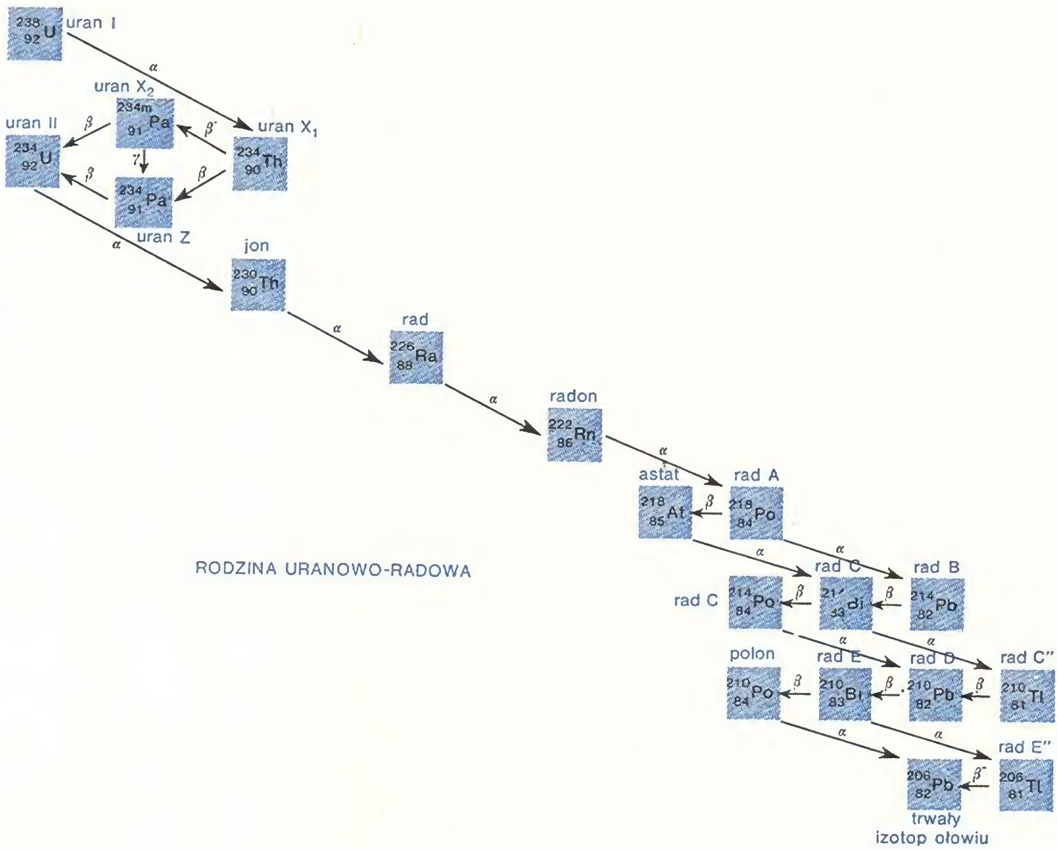
praca z radioizotopami

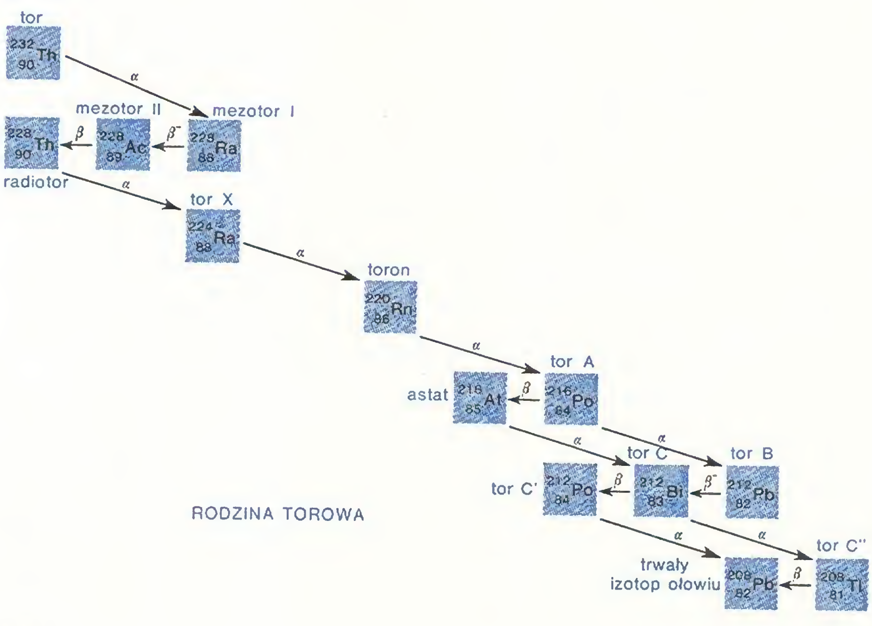
przeróbka mechaniczna (np. porcjowanie, pakowanie) lub chemiczna muszą być prowadzone w specjalnych komorach, zwanych gorącymi, izolowanych osłonami biologicznymi (ciężkie betony, ściany z cegieł ołowianych, wzierniki z grubego szkła ołowiowego itp.) za pomocą manipulatorów (il. 56, tabl. 15). Otrzymywane po obróbce próbki, kształtki itp. umieszcza się w specjalnych pojemnikach stanowiących właściwą biologiczną osłonę radioizotopu, która umożliwia bezpieczny transport (il. 54, tabl. 14).

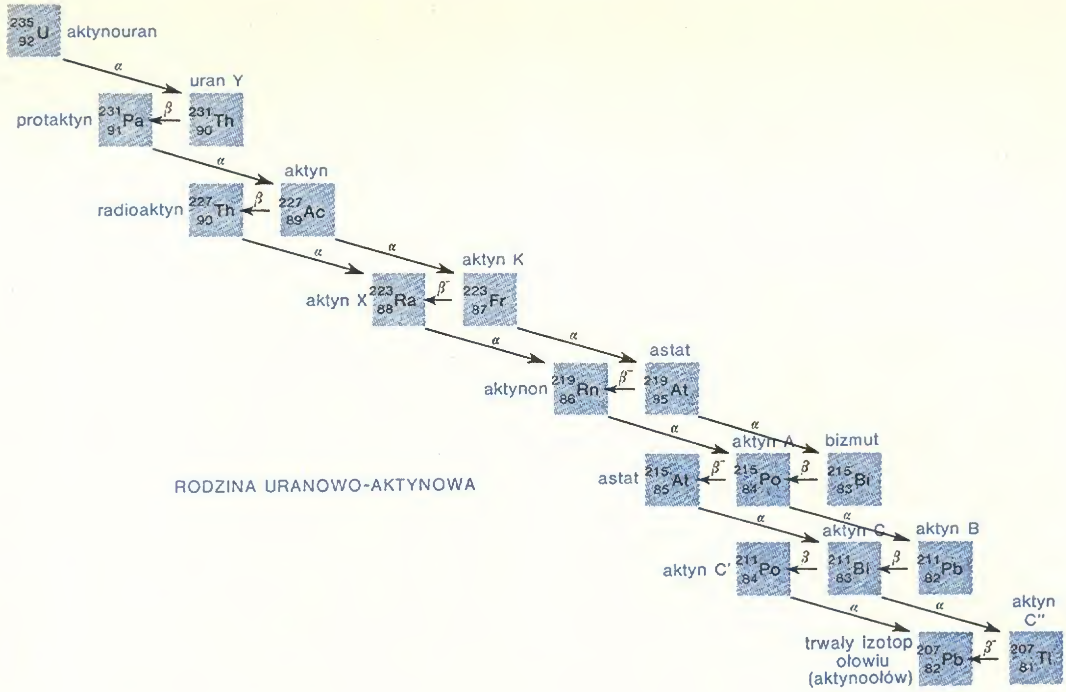
rozszczepienie w materiale paliwowym

Równorzędnym źródłem użytecznych radioizotopów jest sam proces łańcuchowego rozszczepienia przebiegający w materiale paliwowym reaktora. W jednym akcie rozszczepienia powstają dwa jądra, przy czym dla danych *A* i *Z* jądra macierzystego liczby masowe jąder produktów rozszczepienia wy-

rodziny promieniotwórcze







Rys. 1. Rodziny promieniotwórcze (historyczne nazwy radioizotopów podano w kolorze niebieskim)

kazują rozrzut w granicach od 70 do ok. 160 z maksimami wokół wartości 95 i 140. Najważniejszymi, długożyjącymi radioizotopami powstającymi w ten sposób są ^{90}Sr ($t_{1/2}$ ok. 19,9 lat), ^{137}Cs ($t_{1/2}$ ok. 331 lat), ^{95}Zr ($t_{1/2}$ ok. 65 dni), ^{91}Y ($t_{1/2}$ ok. 61 dni), ^{90}Tc ($t_{1/2}$ ok. 2,12 lat), ^{140}Ba ($t_{1/2}$ ok. 13,4 dni), ^{147}Pm ($t_{1/2}$ ok. 2,26 lat), ^{103}Ru ($t_{1/2}$ ok. 40 dni), ^{144}Cs ($t_{1/2}$ ok. 282 dni). Obok nich powstają znaczne ilości innych, krócej żyjących radioizotopów.

przeróbka wypalonego paliwa reaktorowego

Wiele powstałych izotopów silnie pochłania neutrony, co powoduje spowolnienie reakcji łańcuchowej, a z czasem jej zanik. Wyjmowanie paliwa wypalonego przeprowadza się jednak dopiero wtedy, gdy pochłanianie neutronów przez wytworzone izotopy uniemożliwia dalszą pracę reaktora, a wówczas w paliwie powstają już tak wielkie ilości radioizotopów, że wypalone paliwo jest ogromnym źródłem promieniowania. Paliwo takie przechowuje się więc przez pewien czas w specjalnych schronach, a po zniknięciu w nim najkrócej żyjących, ale najsilniej promieniujących radioizotopów może być poddane przeróbce chemicznej. Urządzenia do przeróbki paliwa są jednak znacznie większe i kosztowniejsze od przedstawionych na il. 15 (tabl. 56) urządzeń do przeróbki zaktywowanego w kanałach reaktora materiału tarczowego. Obecnie przeróbkę taką prowadzą jedynie nieliczne kraje. Głównym jej celem nie jest uzyskanie radioizotopów, ale odzyskanie zawartego w paliwie wypalonym materiału rozszczepialnego, w szczególności uzyskanie wytworzonego plutonu. Radioizotopy wydziela się niejako ubocznie w ilościach określonych zapotrzebowaniem na nie.

Niektóre radioizotopy otrzymywane z reakcji rozszczepienia, zwłaszcza radioizotopy krótkożyjące ($t_{1/2}$ rzędu kilku dni) można otrzymywać w zwykłych urządzeniach do przeróbki materiału tarczowego. Wprowadza się w tym celu materiał rozszczepialny do kanałów reaktora i wyjmuje go wcześniej niż paliwo wypalone. Na przykład ^{235}U wprowadzony do kanału reaktora ulega oczywiście reakcji rozszczepienia i gromadzą się w nim radioizotopy — produkty rozszczepienia. Uran w kanale reaktora nie jest jednak prętem paliwowym, czas jego przebywania w reaktorze nie jest uzależniony od technologii pracy reaktora i wyjmuje się go z kanału wtedy, gdy nagromadzą się w nim odpowiednie ilości potrzebnych radioizotopów. Aktywność takiego materiału rozszczepieniowego jest wtedy jeszcze znacznie niższa od aktywności paliwa wypalonego. Dlatego wydzielanie radioizotopów można tu prowadzić w konwencjonalnych komorach gorących. Co więcej, radioizotopy takie, podobnie jak i radioizotopy wytwarzane innymi sposobami, rozpadają się w czasie prowadzenia procesu ich wytwarzania, a więc w ten sposób osiąga się prawie stałe stężenie w czasie kilkakrotnie dłuższym od ich $t_{1/2}$. Stąd optymalny czas przebywania ^{235}U w komorze w celu wytworzenia izotopu ^{99}Mo ($t_{1/2}$ ok. 67 godzin) wynosi od kilkunastu do kilkadziesięciu dni, a nie wiele miesięcy — aż do wypalenia się paliwa. Ten sam izotop można otrzymywać aktywując neutronem trwały izotop ^{99}Mo, ale wówczas nie można go chemicznie oddzielić od materiału tarczowego (ten sam pierwiastek). Dlatego stosowany w medycznej scyntygrafii radioizotopowej ^{99}Mo otrzymuje się opisaną metodą.

wytwarzanie ^{99}Mo

Zastosowanie radioizotopów

Radioizotopy długożyjące ($t_{1/2}$ od godzin do tysięcy lat) wykorzystuje się w badaniach naukowych i technice jako atomy znaczone. Niektóre radioizotopy, szczególnie z $t_{1/2}$ rzędu lat, są wygodnymi, o różnej wydajności lub aktywności źródłami promieniowania α, β i γ. Wreszcie radioizotopy krótkożyjące ($t_{1/2}$ często rzędu sekund), ale łatwo powstające z określonych izotopów trwałych pod wpływem zderzeń z neutronami lub fotonami, mają zastosowanie w chemii analitycznej w metodzie zwanej analizą aktywacyjną.

atomy znaczone

Pierwsze z wymienionych zastosowań pozwala badać mechanizm i kinetykę wielu reakcji chemicznych, procesów biologicznych i procesów związanych z wędrówką materiałów (substancji, atomów, faz).

substancja znaczona Jeżeli np. jako jeden z substratów reakcji wprowadzi się substancję znaczoną, tzn. zawierającą w swym składzie izotopowym niewielką domieszkę określonego radioizotopu, to można śledzić, w jakim produkcie reakcji, a nawet w jakim miejscu w molekule wbudowuje się ten radioizotop. Domieszkę radioizotopu wprowadza się w takiej ilości, aby nie stanowiła ona zagrożenia ze względu na swoje promieniowanie, a przy tym, żeby można było przeprowadzać pomiary metodami radiometrycznymi.

Zastosowanie w medycynie

Podobnie w badaniach fizjologicznych (np. w medycynie) dodając związki znaczone do pokarmu, leku czy wprowadzając do organizmu inną drogą (np. przez iniekcje) możemy śledzić w organizmie żywym ich drogę, badać gromadzenie się, przetwarzanie i wydalanie. W medycynie, szczególnie w diagnostyce, są stosowane dziesiątki, a może nawet setki różnych metod wykorzystujących radioizotopy jako atomy znaczone.

scyntygrafia radioizotopowa Jedną z najnowszych metod z tej dziedziny jest scyntygrafia radioizotopowa, omówiona dokładniej w artykule „Fizyka medyczna". Wykorzystuje się tu głównie izomer technetu ^{99m}Tc o $t_{1/2}$ ok. 6 h, który emitując γ przechodzi w znacznie trwalszy ^{99}Tc. Promieniowanie γ tego procesu jest łatwo wykrywalne przy natężeniach jeszcze zupełnie nieszkodliwych. Izomer ^{99m}Tc otrzymujemy z molibdenu w reakcji $^{99}Mo \rightarrow \beta^- + {}^{99m}Tc$.

Do zakładu leczniczego dostarcza się tzw. generator technetu, czyli rurkę zawierającą radioizotop ^{99}Mo osadzony na podłożu. Tutaj powstały w generatorze izomer technetu ^{99m}Tc jest wymywany za pomocą standardowych zestawów odczynników i wprowadzany do organizmu badanej osoby, gdzie umiejscawia się na lub w odpowiednich narządach. Promieniowanie γ ^{99m}Tc pozwala badać te narządy, wykrywać w nich zmiany niedostrzegalne innymi metodami. Natężenie tego promieniowania jest niewielkie, a krótki $t_{1/2}$ sprawia, że maleje ono szybko w czasie. Dlatego metoda jest nieszkodliwa dla badanych. Sumaryczne dawki wprowadzonego do organizmu promieniowania są w tej metodzie mniejsze, niż w prześwietleniu promieniami rentgenowskimi. Powstały z ^{99m}Tc trwalszy radioizotop ^{99}Tc zawarty jest w takich ilościach, że jego promieniowanie jest już nie tylko nieszkodliwe, ale praktycznie niewykrywalne.

Zastosowanie w technice

W technice szczególnie użyteczne okazują się radioizotopy do śledzenia przebiegu procesów w niedostępnej z zewnątrz aparaturze, w reaktorach itp. Badanie radioizotopowe ustala rzeczywisty przebieg procesów w aparaturze, co pozwala na wykrycie i usunięcie nieprawidłowości, maksymalną wydajność i optymalizację procesów. W instalacjach, w których reagentami są ciecze lub gazy (np. przemysł rafineryjny, petrochemiczny, koksochemiczny itp.), wprowadza się rozpuszczalne lub lotne związki znaczone radioizotopami, a następnie radiometrycznie określa szybkości przepływu i czasy zatrzymania się reagentów w różnych częściach aparatury. Czujniki radiometryczne umieszcza się na zewnątrz aparatury bezpośrednio przy miejscach badanych lub wprowadza się je do wnętrza aparatury lub rurociągu.

bezpieczeństwo pracy Ilości radioizotopów i ich rodzaj ($t_{1/2}$) tak są dobierane, aby badania można było bezpiecznie przeprowadzić z wymaganą dokładnością i aby końcowe stężenie wprowadzonego radioizotopu, rozcienczonego materiałem nieaktywnym, było niższe od stężenia dopuszczalnego przepisami ochrony radiologicznej, albo aby wprowadzony radioizotop po użyciu mógł być oddzielony i zabezpieczony w pojemniku. Nad spełnieniem tych wymogów czuwają w Polsce specjalnie przeszkoleni inspektorzy ochrony przed promieniowaniem, a nadzór nad ich szkoleniem i wszelką działalnością związaną z radioizotopami prowadzi Centralne Laboratorium Ochrony Radiologicznej (CLOR).

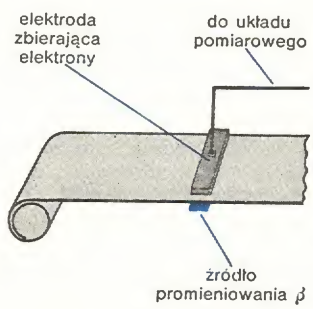
Badając procesy dotyczące sypkich ciał stałych radioizotop wprowadza się często w postaci grudek lub igiełek. Przykładem przemysłowego wykorzystania metody atomów znaczonych może być badanie w Polsce w Wizowie i w NRD w Coswig przyczyn **zastosowanie w cementowni** zaklejania się pieców cementowych. Po przeprowadzeniu badania okazało się, że w wyniku fluidyzacji ruch materiałów w niektórych partiach pieców obrotowych był szybszy od przewidywanego i do strefy wysokotemperaturowej dostawał się nierozłożony siarczan wapnia. Wmontowanie w piece progów spowalniających ruch materiału pozwoliło na usunięcie tych nieprawidłowości. Przykłady takie można mnożyć. Na pewno jednak można powiedzieć, że upowszechnienie stosowania radioizotopowych metod badania procesów technologicznych stanowi jeden z podstawowych, ciągle jeszcze niedostatecznie wykorzystywanych czynników usprawniania technologii i poprawy zarówno jakości produktów, jak i ekonomii ich wytwarzania.

wykrywanie wad urządzeń technicznych Innymi przykładami zastosowania radioizotopów jako atomów znaczonych mogą być: wykrywanie nieszczelności czynnych rurociągów (radioizotop przenika przez nieszczelności i gromadzi się w ich pobliżu), badanie rozchodzenia się ścieków w zbiornikach wodnych, badanie rozkładu zanieczyszczeń w atmosferze i wiele innych.

Zastosowanie źródeł promieniowania

Radioizotopy jako źródła promieniowania bywają wykorzystywane na wielką skalę w defektoskopii, czyli do wykrywania wad w odlewach, spawach i innych urządzeniach. Stosuje się tu niewielkie źródła przenikliwego promieniowania γ, które łatwo jest wprowadzić w rozmaite zakamarki badanych obiektów i wykonać odpowiednie prześwietlenia. Radioizotopowe źródełka można wprowadzać w miejsca, w których nie zmieściłaby się lampa rentgenowska i dlatego defektoskopia radioizotopowa jest znacznie wszechstronniejsza od rentgenowskiej, jakkolwiek zasady obu tych metod są takie same.

defektoskopia radioizotopowa Ilustracja 53 (tabl. 14) przedstawia jednoczesne prześwietlanie sześciu zaworów tym samym defektoskopem zawierającym radioizotop ^{60}Co, obok widzimy jedno z otrzymanych w ten sposób zdjęć. Aktywność radioizotopu jest tak dobrana, aby przebywanie w pobliżu źródła w czasie potrzebnym na jego ustawienie lub usunięcie powodowało wprowadzenie do ciała operatora dawki znacznie mniejszej od dopuszczalnej. Natomiast wymagane naświetlenie kliszy osiąga się odpowiednio dobierając czas ekspozycji, w trakcie której operator nie jest potrzebny.



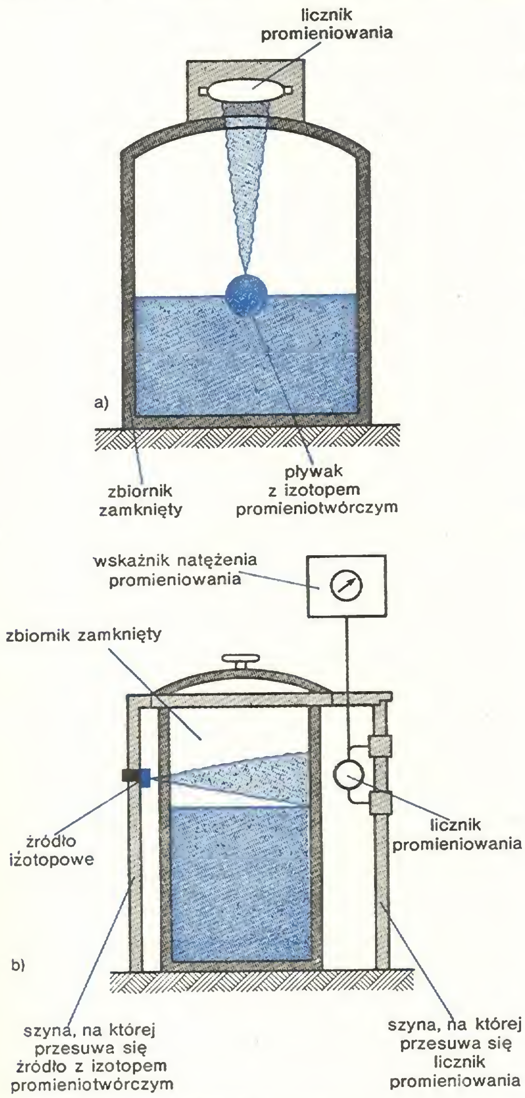
Rys. 2. Schemat izotopowego urządzenia do pomiaru grubości błon

Innym zastosowaniem radioizotopów jako źródeł promieniowania są rozmaite czujniki i wskaźniki; na rys. 2 widzimy schemat urządzenia do ciągłego po-

miaru grubości błony opartego na zasadzie pochłaniania promieniowania β. Z jednej strony błony znajduje się źródełko radioizotopowe, a z drugiej — detektor radiometryczny. Sygnał z urządzenia może automatycznie korygować grubość produkowanej błony. Urządzenia tego typu można stosować do regulacji grubości, np. w produkcji papieru, blach w walcowniach i in.

poziomomierz

Przykładami urządzeń o podobnej zasadzie działania są: pływakowy poziomomierz izotopowy oraz wskaźnik poziomu cieczy w zbiorniku zaopatrzonym

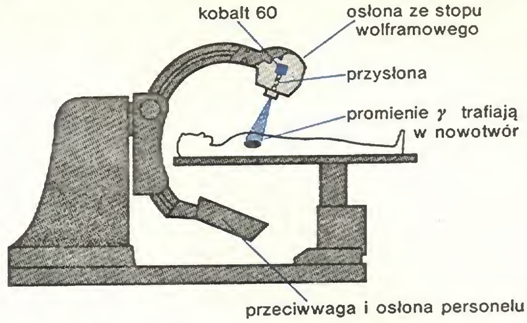


Rys. 3. Schematy poziomomierzy izotopowych: a) poziomomierz pływakowy; b) poziomomierz ze źródłem umieszczonym na zewnątrz zbiornika

w zewnętrzne źródło promieniowania (rys. 3). W poziomomierzach pływakowych wykorzystuje się fakt, że promieniowanie jest znacznie silniej pochłaniane przez ciała stałe i ciecze niż przez gaz, a w działaniu poziomomierza odgrywa głównie rolę zmniejszanie się natężenia promieniowania w miarę oddalania się detektora od źródła punktowego.

bomba kobaltowa

Znacznie większe źródła promieniowania są stosowane w medycynie do niszczenia tkanek nowotworowych. Są to tzw. bomby kobaltowe, czyli źródła zawierające (w odpowiedniej osłonie) radioizotop ^{60}Co w takiej ilości, że wypromieniowywana przezeń energia promieniowania γ ma moc ok. jednego lub kilku watów. Przez zastosowanie specjalnych osłon promieniowanie to (drobna jego część) skierowuje się na tkankę nowotworową. Czas napromieniania dobiera się tak, aby tkanka nowotworowa uległa



Rys. 4. Schemat kanadyjskiej medycznej bomby kobaltowej Theratron

zniszczeniu, a słabiej napromieniowane tkanki sąsiednie doznały tylko odwracalnych zmian (np. możliwe do zagojenia rany) (rys. 4 oraz il. 52, tabl. 14).

sterylizacja

Jeszcze silniejsze źródła promieniowania γ stosuje się w niektórych procesach technologicznych — np. w produkcji sprzętu medycznego jednorazowego zastosowania oraz sterylnych opatrunków. Stosowane tu radioizotopy to głównie kobalt ^{60}Co, a także na mniejszą skalę cez ^{137}Cs. Moc promieniowania γ emitowanego przez te źródła nieraz przekracza 10 kW.

Analiza aktywacyjna

Trzecią ważną dziedziną niejako pośredniego wykorzystania radioizotopu jest analiza aktywacyjna. Metoda ta, jak już wspomnieliśmy, polega na napromienieniu, najczęściej neutronami, a czasem fotonami o energii regulowanej w zakresie od 10 do 30 MeV, badanych próbek i analizie emitowanego po aktywacji promieniowania. W czasie aktywacji neutronowej z izotopów trwałych powstają w próbce możliwe do wykrycia i zidentyfikowania radioizotopy. W aktywacji fotonami niektóre trwałe izotopy z próbki ulegają wzbudzeniu i emitują promieniowanie γ. Obserwujemy wtedy rozproszenie promieniowania padającego, występujące rezonansowo dla energii odpowiadających energiom wzbudzenia i emisji γ izotopów trwałych zawartych w próbce. Aktywacja fotonami może również wywoływać rozpady wzbudzonych energetycznie jąder, co prowadzi do powstawania wykrywalnych radiometrycznie radioizotopów. Ilość każdego powstałego radioizotopu jest w danych warunkach aktywacji proporcjonalna do ilości macierzystego izotopu trwałego, analiza aktywacyjna pozwala zatem na wykrycie i oznaczenie ilości macierzystych izotopów trwałych w badanej próbce. W wypadku rozpraszania rezonansowego te same informacje uzyskuje się na podstawie analizy promieniowania pochłanianego i rozpraszanego. Czułość analizy aktywacyjnej przewyższa znacznie nie tylko czułość tradycyjnych metod chemicznych, ale i analizy spektralnej. Co więcej, analiza aktywacyjna pozwala na określenie zawartości wielu pierwiastków, których widmo spektralne leży w dalekim nadfiolecie i które z tego względu nie są wykrywalne zwykłymi metodami spektralnymi.

aktywacja neutronami i fotonami

czułość analizy aktywacyjnej

Aktywację neutronową nieraz prowadzi się w kanałach reaktora, ale wtedy nie można wykorzystać wielu przydatnych w analizie aktywacyjnej radioizotopów o krótkim $t_{1/2}$. Dlatego lepsze są inne źródła neutronów aktywujących próbki. Często stosuje się np. źródła radowo-berylowe, w których zachodzą reakcje: $^{226}Ra \rightarrow {}^{222}Rn + {}^{4}He$ i ${}^{4}_{2}He + {}^{9}_{4}Be \rightarrow ({}^{13}_{6}C) \rightarrow {}^{12}_{6}C + {}^{1}_{0}n$. Natężenie neutronów z takiego źródła jest jednak słabe, co ogranicza czułość metody. Naj-

większe możliwości w nowoczesnej analizie aktywacyjnej stwarza generator neutronów złożony z akceleratora przyspieszającego jądra deuteru D i wstrzeliwującego je w tarczę z tytanu lub cyrkonu nasyconego trytem T. Jądra deuteru otrzymuje się w procesie hydrolizy ciężkiej wody, które następnie w zmiennym polu elektrycznym wielkiej częstotliwości (w tzw. źródle jonów) ulegają dysocjacji. Otrzymywane jądra deuteru przyspieszane napięciem 150–200 kV trafiają w jądra trytu i zachodzi reakcja syntezy jądrowej ${}^{2}_{1}D+{}^{3}_{1}T \rightarrow {}^{4}_{2}He+{}^{1}_{0}n$. Powstałe neutrony mają energię 14 MeV.

Aktywacja fotonami jest metodą uzupełniającą, ponieważ jest dla pewnych pierwiastków lepsza, a dla innych gorsza od aktywacji neutronowej. Wzbudzona radioaktywność próbek badanych analizą aktywacyjną jest na tyle krótkotrwała i słaba, że można analizę aktywacyjną zaliczyć do grupy metod nieniszczących.

zastosowanie w kryminalistyce

Ilustracją czułości analizy aktywacyjnej może być przykład jej wykorzystania w kryminalistyce do określania, z jakiej odległości (w zakresie od 0 do 4 m) został oddany strzał do ofiary. Badanie polega na określeniu ilości i rozmieszczenia antymonu osadzonego na ciele (ubraniu) ofiary w pobliżu wlotu pocisku, a wydzielonego podczas wybuchu spłonki z zawartego w niej siarczku antymonowego. Obłok siarczku antymonowego w miarę oddalania się od wylotu lufy staje się większy i rzadszy. W wypadku strzału z odległości 2 m w pobliżu wlotu pocisku osadza się ok. 10^{-9} g antymonu. Po strzale z bliższej odległości ilość osadzonego antymonu jest większa i jest on osadzony bliżej wlotu. Problem ustalenia odległości strzału bywa nieraz najważniejszy w wypadku istnienia alternatywy samobójstwa lub morderstwa.

Datowanie promieniotwórcze

Oprócz wymienionych głównych dziedzin zastosowań radioizotopów istnieją inne, nie mieszczące się w powyższej klasyfikacji. Najważniejszą ich grupę stanowią metody datowania promieniotwórczego, wykorzystujące stałość okresu $t_{1/2}$. Najbardziej znaną metodą tego rodzaju jest określanie zawartości radioizotopu ^{14}C w szczątkach organicznych. Radioizotop ten pod wpływem promieniowania kosmicznego nieustannie powstaje w atmosferze z zawartego w niej azotu ^{14}N i nieustannie rozpada się z $t_{1/2} = 5740$ lat. Oba procesy prowadzą do wytworzenia się w atmosferycznym dwutlenku węgla CO_2 izotopu ^{14}C o stałym, ale bardzo małym natężeniu. W procesie fotosyntezy CO_2 izotop przyswajają organizmy żywe, a w skorupach korali, mięczaków i in. związany jest z wapniem Ca w $CaCO_3$. Po ustaniu procesów życiowych ustaje też przyswajanie ^{14}C z atmosferycznego CO_2 i w szczątkach organicznych oraz w wytworzonym wapieniu zawartość ^{14}C spada w czasie. Mierząc stężenie ^{14}C można więc określić, jak dawno zostały wytworzone badane szczątki organiczne lub wapienie. Czułość metody pozwala na datowanie próbek nie starszych niż 70 000 lat. Do oznaczania wieku starszych próbek geologicznych (rzędu milionów lat) można wykorzystać słabo promieniotwórczy, naturalny radioizotop potasu ^{40}K o $t_{1/2}$ = ok. $1{,}3 \cdot 10^9$ lat. Radioizotop ten emituje pozyton i przekształca się w trwały izotop argonu ^{40}Ar, wykrywalny w bardzo małych stężeniach metodami analizy aktywacyjnej. Wytworzony argon pozostaje uwięziony w skale zawierającej radioaktywny potas. Czas upływający od wytworzenia się skały można obliczyć oznaczając w niej stosunek stężeń ^{40}Ar do ^{40}K.

kontrola terminu ważności produktów

Wskaźniki radioizotopowe można wykorzystać również do niezawodnego określenia daty produkcji wyrobów przemysłowych, których czas używalności jest ściśle ograniczony. W tym celu wystarczy dodać do wyrobu dwa radioizotopy o różnych $t_{1/2}$ w określonym stosunku stężeń, a po pewnym czasie z pomiarów radiometrycznych określić stosunek ich stężeń. Inaczej mówiąc — wprowadzamy do wyrobu wewnętrzny zegar izotopowy. Dodane ilości radioizotopów muszą być oczywiście bezpiecznie małe. Zegar izotopowy jest niezależny od dalszych procesów technologicznych, jak np. porcjowanie, pakowanie.

Podane przykłady nie wyczerpują wszystkich możliwych zastosowań radioizotopów w różnych dziedzinach nauki, techniki. Wystarczy powiedzieć, że katalogi producentów związków znaczonych zawierają tysiące pozycji i wszystkie one znajdują odbiorców.

W Polsce radioizotopy i związki znaczone produkuje Ośrodek Produkcji i Dystrybucji Izotopów (OPiDI) w Świerku koło Otwocka. Ośrodek OPiDI zajmuje się nie tylko dystrybucją związków znaczonych i radioizotopów wyprodukowanych w kraju, ale i importowanych z zagranicy. Ze względów ekonomicznych i technicznych w Polsce nie prowadzi się przeróbki paliwa wypalonego, a radioizotopy produkuje się w kanałach reaktorów Instytutu Badań Jądrowych (IBJ) w Świerku. Z tych radioizotopów w OPiDI tworzy się (dozowanie, pakowanie) źródła wzorcowe, źródła dla defektoskopii i innych celów oraz wyrabia i pakuje związki znaczone. W OPiDI wyrabia się również wspomniane uprzednio generatory technetu dla medycyny.

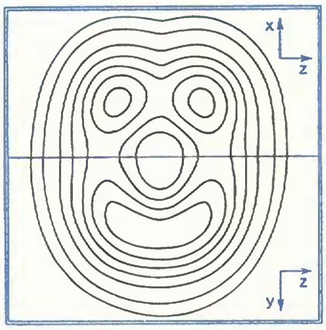
R. Szepke *Radiometria stosowana*, Warszawa 1967; W. Szymański *Wstęp do fizykochemii jądrowej*, Toruń 1974; O. Wolczek *Izotopy*, Warszawa 1965.

FIZYKA ATOMU, CZĄSTECZKI I CIAŁA STAŁEGO

Chemia kwantowa · Spektroskopia · Kierunki rozwoju optyki · Kriofizyka · Kryształy · Fizyka ciała stałego · Magnetyzm ·

CHEMIA KWANTOWA

Józef Stanisław Kwiatkowski



W trakcie wykonywania badań zarówno doświadczalnych, jak i teoretycznych otrzymuje się czasami nieoczekwane wyniki. Obok podany rysunek przedstawia obliczone przez fizyków jądrowych warstwice gęstości dla jądra ^{32}S. *Czyżby odkryto „szczęśliwe" jądro atomowe?*

Kiedy przed pięćdziesięciu laty powstała mechanika kwantowa, i kiedy z powodzeniem zaczęto stosować jej prawa do wyjaśniania elementarnych zjawisk zachodzących w mikroświecie atomów i cząsteczek, zaczęto rozumieć, że w teorii kwantów sformułowane zostały podstawowe prawa obejmujące znaczną część fizyki, całą chemię, a nawet niektóre zjawiska biologiczne. Przedmiotem chemii kwantowej jest zastosowanie mechaniki kwantowej do opisu budowy atomów i cząsteczek, oddziaływań między nimi i elementarnych procesów chemicznych. W realizacji tego napotykamy jednak poważne trudności, gdyż prawa mechaniki kwantowej wymagają stosowania równań, których nie można rozwiązać ściśle. Ta trudność matematyczna przez długi czas hamowała rozwój chemii kwantowej. Wprowadzano wiele drastycznych przybliżeń do teorii kwantowej, ograniczając się do jakościowego wyjaśniania własności cząsteczek.

Pojawienie się i rozwój elektronicznych maszyn cyfrowych (komputerów) nastąpił niezwykle szybki rozwój chemii kwantowej. Obecnie możemy obliczać i przewidywać wiele własności cząsteczek bez znajomości wartości doświadczalnych dla tych układów. Dzięki rozwojowi techniki obliczeniowej możemy uzyskiwać dla małych cząsteczek teoretyczne wartości fizyczne o dokładności porównywalnej z dokładnością najprecyzyjniejszych pomiarów lub przekraczającej dokładności tych pomiarów. Również dla większych cząsteczek w wielu wypadkach teoria może z powodzeniem współzawodniczyć z doświadczeniem w przewidywaniu ich własności.

chemia kwantowa jako narzędzie

Metody obliczeniowe chemii kwantowej stanowią jakby „teoretyczną technikę pomiarową" lub „przyrządy teoretyczne", których zdolność rozdzielcza (dokładność) zwiększa się z dnia na dzień. W chwili obecnej zdolność rozdzielcza „przyrządów teoretycznych" dla małych układów jest bardzo duża, natomiast dla układów dużych zdolność rozdzielcza jest jeszcze mała, chociaż i w tym wypadku metody obliczeniowe chemii kwantowej mają czasami przewagę nad pomiarami doświadczalnymi. Możemy bowiem obliczyć takie wartości, które dla cząstek są bardzo trudne do zmierzenia lub w ogóle nie są mierzalne (np. próbka nie jest rozpuszczalna lub związek chemiczny występuje w tak rzadkiej formie tautomerycznej, że nie można dokonać pomiaru).

dokładność obliczeń kwantowych

Coraz częściej więc chemicy zastanawiają się, czy jakąś konkretną wielkość fizyczną cząsteczki mierzyć, czy też lepiej ją obliczyć, rezygnując czasami z bardzo pracochłonnych i uciążliwych pomiarów na rzecz wartości teoretycznej, która jest wprawdzie mniej dokładna (powiedzmy, że dokładność jest rzędu 10–15%), ale którą otrzymuje się stosunkowo szybko. Coraz częściej w pracach chemicznych konfrontuje się wartości mierzone z wartościami obliczonymi lub używa się poglądowych metod chemii kwantowej do wyjaśnienia faktów doświadczalnych. Chemia kwantowa zajmuje się bowiem nie tylko obliczeniami z zastosowaniem komputerów, ale głównie teoretycznym wyjaśnianiem podstawowych zjawisk przyrody (w jaki sposób atomy łączą się tworząc cząsteczki, dlaczego określona cząsteczka ma takie a nie inne własności itp.), formułowaniem ogólnych praw dotyczących układów chemicznych oraz wyjaśnianiem własności konkretnych układów. Chemia kwantowa przekłada również abstrakcyjny język mechaniki kwantowej na język poglądowy, tworząc np. poglądowy model cząsteczki i jej oddziaływania z otoczeniem. Oczywiście model ten musi być zgodny z prawami mechaniki kwantowej. Od mechaniki kwantowej, jak od każdej innej teorii fizycznej, oczekujemy wyjaśniania znanych faktów doświadczalnych i przewidywania nowych.

chemia kwantowa jako teoria

Konfrontacja naszych wyobrażeń o świecie atomów i cząsteczek z faktami doświadczalnymi potwierdza jak dotąd słuszność naszych poglądów na budowę układów chemicznych. Należy zdawać sobie jednak sprawę, że zgodność wyników teoretycznych z doświadczeniem nie jest nigdy bezwzględnym potwierdzeniem teorii, gdyż każda teoria (a więc i mechanika kwantowa) ma ograniczony zakres stosowalności. Wcześniej czy później zostaną wykonane doświadczenia, których wyniki będą sprzeczne z teorią uznawaną obecnie za słuszną. I chociaż zostanie stworzona nowa teoria, to

stara teoria będzie nadal stosowana, tylko granice jej stosowalności zostaną zawężone i dokładniej sprecyzowane. Omawiana tutaj chemia kwantowa jest nierelatywistyczną mechaniką kwantową atomów i cząsteczek. Jest to więc metoda przybliżona, nie uwzględniająca wielu subtelnych efektów związanych z zależnością masy od prędkości. Wszystkie dotychczasowe wyniki świadczą jednak o tym, że nierelatywistyczna chemia kwantowa, z wyjątkiem bardzo nielicznych wypadków, zadowalająco opisuje układy chemiczne i zachodzące w nich procesy.

Fizyczne określenie układu chemicznego

Chemicznie interesującym układem może być atom, jon, cząsteczka, kompleks molekularny, cały kryształ lub zespół atomów, jonów, cząsteczek itd. poruszających się względem siebie lub biorących udział w reakcji chemicznej. Każdy z tych układów składa się z zespołu jąder atomowych i elektronów, które oddziałują ze sobą za pomocą sił elektrostatycznych; ruchy tych zespołów opisane są prawami mechaniki kwantowej.

funkcja falowa a stan układu

Dwoma podstawowymi pojęciami mechaniki kwantowej są funkcja falowa i operator. Funkcja falowa opisuje stan układu, tj. stan zespołu jąder atomowych i elektronów układu. Innymi słowy, funkcja falowa układu dla określonego stanu określa wszystkie własności cząstek w danym stanie. Funkcja falowa Ψ

Funkcja falowa, gęstość prawdopodobieństwa, prawdopodobieństwo

Wielkość	Zapis
Funkcja falowa stanu stacjonarnego n cząstek	$\Psi(x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_n, y_n, z_n)$ lub $\Psi(1, 2, \ldots, n)$
Gęstość prawdopodobieństwa w stanie opisanym funkcją Ψ	$\rho(1, 2, \ldots, n) = \Psi(1, 2, \ldots, n)\, \Psi^*(1, 2, \ldots, n)$
Prawdopodobieństwo znalezienia cząstki 1. w elemencie objętości $dv_1 = dx_1\, dy_1\, dz_1$ przy zupełnie dowolnych położeniach pozostałych cząstek	$\rho(1, 2, \ldots, n)\, dv_1$
Prawdopodobieństwo znalezienia cząstki 1. w dv_1, cząstki 2. w $dv_2, \ldots$, cząstki n-tej w dv_n	$\rho(1, 2, \ldots, n)\, dv_1 dv_2 \ldots dv_n$
Prawdopodobieństwo znalezienia układu n cząstek w całej przestrzeni	$\int \rho(1, 2, \ldots, n)\, dv_1 dv_2 \ldots dv_n = 1$

Zamiast współrzędnych x_1, y_1, z_1 cząstki 1. użyto symbolu 1. Podobnie oznaczono współrzędne cząstki 2., ..., n-tej

układu n cząstek zależy od współrzędnych wszystkich cząstek (tj. od $3n$ współrzędnych położeń cząstek i ich n współrzędnych spinowych, chociaż te ostatnie współrzędne w rozważaniach pominiemy) oraz, w wypadku stanów niestacjonarnych, dodatkowo od czasu (zob. tabela powyżej).

W dalszych rozważaniach będziemy ograniczali uwagę tylko do układów stacjonarnych, których energia nie zależy od czasu. Kwadrat funkcji falowej, a ściślej iloczyn funkcji falowej pomnożony przez jej wielkość sprzężoną $\Psi\Psi^*$ (funkcja falowa może być funkcją zespoloną) nazywamy gęstością prawdopodobieństwa ϱ. Funkcja falowa musi być unormowana, tzn. scałkowanie gęstości prawdopodobieństwa po całej przestrzeni daje 1, co fizycznie oznacza, że prawdopodobieństwo znalezienia w całej przestrzeni układu cząstek znajdującego się w stanie opisanym funkcją $\Psi(1, 2, \ldots, n)$ jest równe pewności.

gęstość prawdopodobieństwa

Drugim podstawowym pojęciem stosowanym w mechanice kwantowej jest pojęcie operatora, szczególnie operatora Hamiltona. Operator ten, zwany też hamiltonianem, jest operatorem całkowitej energii układu cząstek i można go przedstawić w postaci

hamiltonian

$$\hat{H} = \hat{T} + \hat{V},$$

gdzie $\hat{T}$ oznacza operator energii kinetycznej, a $\hat{V}$ jest operatorem energii potencjalnej. Dla każdej cząstki o masie m_i operator energii kinetycznej jest operatorem różniczkowym mającym postać sumy drugich pochodnych cząstkowych

$$\hat{T}_i = -\frac{\hbar^2}{2m_i}\left(\frac{\partial^2}{x_i^2} + \frac{\partial^2}{y_i^2} + \frac{\partial^2}{z_i^2}\right),$$

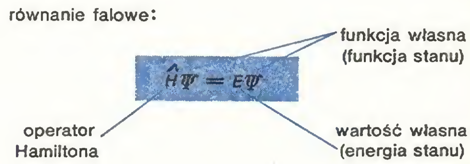
gdzie $\hbar$ jest stałą Plancka h podzieloną przez 2π. Operator energii kinetycznej całego układu n cząstek jest po prostu sumą operatorów energii kinetycznej poszczególnych cząstek, tj. $\hat{T} = \sum_{i=1}^{n} \hat{T}_i$. Natomiast operator energii potencjalnej jest sumą energii elektrostatycznego oddziaływania wszystkich cząstek ze sobą

$$\hat{V} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{z_i z_j e^2}{r_{ij}},$$

gdzie $z_i e$ i $z_j e$ są ładunkami cząstek odpowiednio i-tej oraz j-tej, a r_{ij} jest odległością pomiędzy tymi cząstkami.

Operator Hamiltona, jako sumę operatora $\hat{T}$ oraz $\hat{V}$, możemy bardzo łatwo podać dla dowolnego układu chemicznego. Innymi słowy, dla dowolnego zespołu jąder atomowych i elektronów możemy hamiltonian traktować jako znany. Posługując się tak określonym operatorem Hamiltona można sformułować podstawowe równanie mechaniki kwantowej (równanie Schrödingera) opisujące stany stacjonarne układu chemicznego:

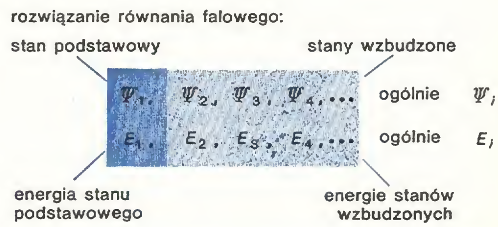
równanie Schrödingera



(1)

Równanie (1) jest równaniem różniczkowym (nazywanym w mechanice kwantowej zagadnieniem na wartości własne), które należy rozwiązać, aby znaleźć funkcję falową układu oraz odpowiadającą tej funkcji wartość energii. Okazuje się, że z rozwiązania równania (1) otrzymujemy wiele (cały układ) funkcji Ψ_i opisujących poszczególne stany układu oraz odpowiednie energie tych stanów E_i:

funkcje własne



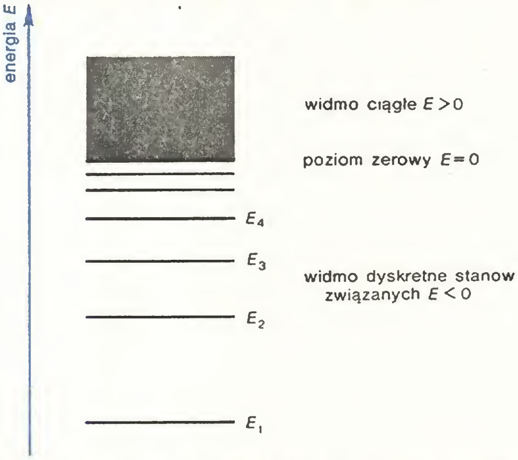
(2)

Ważną właściwością układu chemicznego, związaną z postacią energii potencjalnej w hamiltonianie, jest to, że układ w stanie stacjonarnym może przyjmować tylko określone dyskretne wartości energii. Najniższa wartość własna jest zawsze skończona (tutaj E_1), a odpowiedni stan jest nazywany stanem podstawowym (opisuje go funkcja Ψ_1), podczas gdy pozostałe dyskretne wartości energii (E_2, E_3 itd.) odnoszą się do stanów wzbudzonych (opisanych odpowiednio funkcjami Ψ_2, Ψ_3 itd.). Przy wyższych energiach występuje na ogół kontinuum, które odpowiada ciągłemu rozkładowi energii (rys. 1). Fizycznie oznacza to stany

wartości (energie) własne

energia zerowa układu

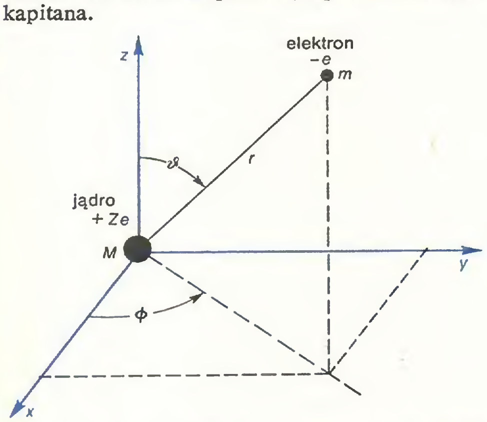
zjonizowane lub zdysocjowane. Energia zerowa układu jest z definicji energią stanu, w którym wszystkie jądra atomowe i elektrony są oddalone na nieskończoną odległość jedne od drugich. Poziomy energetyczne stanów związanych mają więc energie ujemne. Wartość liczbowa całkowitej energii układu w stanie podstawowym jest równa energii niezbędnej do całko-



Rys. 1. Schematyczne przedstawienie rozwiązania równania (1). Energie E_i dyskretnych stanów opisanych funkcjami Ψ_i są uporządkowane ze względu na wzrastającą ich wartość (stany związane mają ujemne wartości energii). Poziom o energii $E = 0$ odpowiada jonizacji, powyżej tego poziomu widmo energii jest ciągłe; wybór poziomu zerowego jest wyłącznie sprawą umowy

witego zjonizowania i zdysocjowania układu. Dla dużych układów ich energia może być rzędu tysięcy elektronowoltów. Z drugiej strony, mierzone w cząsteczkach organicznych energie przejść ze stanu podstawowego do stanu wzbudzonego są rzędu kilku elektronowoltów. Aby teoretycznie otrzymać energię przejścia, należy obliczyć energię odpowiedniego stanu wzbudzonego i odjąć od niej, również obliczoną, energię stanu podstawowego układu. Musimy więc najpierw obliczyć dwie bardzo duże liczby i przez odjęcie ich od siebie otrzymać małą liczbę odpowiadającą mierzonej energii przejścia. Przykład ten ukazuje, jak dokładne muszą być obliczenia mechaniki kwantowej, aby poprawnie interpretować wielkości fizykochemiczne układów. Nic dziwnego, że kilkanaście lat temu jeden z twórców chemii kwantowej, C.A. Coulson, przyrównał obliczenia chemii kwantowej do znajdowania masy kapitana statku przez zważenie statku wraz z kapitanem, a potem — statku bez kapitana.

atom jednoelektronowy



Rys. 2. Atom jednoelektronowy można poglądowo przedstawić w postaci układu składającego się z jądra o ładunku $+Ze$ i masie M oraz elektronu o ładunku $-e$ i masie m. Położenie elektronu względem jądra określa się wówczas jednoznacznie podając jego współrzędne kartezjańskie x, y, z lub współrzędne biegunowe r, ϑ, φ

Atomy jednoelektronowe

Równanie Schrödingera można rozwiązać w sposób ścisły (tzn. w postaci analitycznej, za pomocą znanych funkcji) wyłącznie dla atomów jednoelektronowych. W wypadku każdego innego układu chemicznego równanie (1) można rozwiązać jedynie w sposób przybliżony (może to jednak być rozwiązanie bardzo dokładne, przekraczające dokładnością pomiary doświadczalne!!!).

Funkcje falowe opisujące stany związane elektronu (o energiach $E<0$) w atomie jednoelektronowym mają we współrzędnych biegunowych (rys. 2) postać

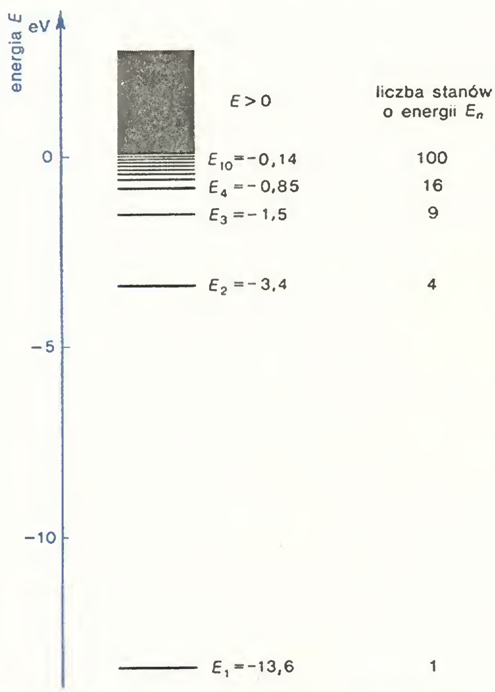
$$\Psi_{nlm} = \underbrace{R_{nl}(r)}_{\text{funkcja radialna}}\ \underbrace{Y_{lm}(\vartheta, \phi)}_{\text{funkcja kątowa}} \qquad (3)$$

liczby kwantowe
$$\begin{cases} n = 1, 2, 3, \ldots \\ l = 0, 1, 2, \ldots, n-1 \\ m = l, l-1, l-2, \ldots, -l-2, -l-1, -l \end{cases}$$

funkcja kątowa i funkcja radialna

w której funkcje radialne, zależne wyłącznie od odległości elektronu od jądra, wyrażone są za pomocą dobrze znanych w matematyce wielomianów Laguerre'a, natomiast funkcje kątowe wyrażone są za pomocą stowarzyszonych wielomianów Legendre'a. Funkcje falowe (3) oznaczone są trzema liczbami kwantowymi n, l, m, z których pierwsza może przyjmować dowolne, ale wyłącznie dodatnie wartości całkowite (nie może jednak być równa 0), natomiast liczby l i m mogą przyjmować ograniczone wartości. Fakt, że wspomniane liczby mogą przyjmować ściśle określone wartości, wynika z warunków matematycznych na istnienie rozwiązania równania (1) dla elektronu w atomie jednoelektronowym. Funkcje falowe

liczby kwantowe



Rys. 3. Energie elektronu w atomie wodoru w stanie podstawowym i w stanach wzbudzonych. Dla przejrzystości opuszczono wartości energii elektronu w stanach o n = 5, 6, 7, 8, 9. Wynoszą one odpowiednio −0,54, −0,38, −0,28, −0,21, −0,17 eV. Wzór (4) jest tylko częścią rozwiązania równania (1) dla atomu wodoru, opisującą najbardziej interesujące chemika i fizyka stany związane elektronu. Obszar widma ciągłego ($E > 0$) także wynika z rozwiązania równania (1)

opisujące stan elektronu w tym atomie „ponumerowane" są więc trzema wskaźnikami; nie jest to sprzeczne z tym, co powiedziano już wcześniej, że funkcje własne i energie własne otrzymane z rozwiązania równania (1) można ponumerować za pomocą kolejnych liczb całkowitych. Poszczególne trójki liczb n, l i m można przecież przyporządkować jednoznacznie kolejnym liczbom całkowitym.

Rozwiązując zagadnienie na wartości własne dla atomu jednoelektronowego, otrzymujemy oprócz funkcji własnych Ψ_{nlm} wartości własne odpowiadające danym funkcjom:

energie elektronu

$$E_n = -\frac{\mu Z^2 e^4}{2\hbar^2 n^2}, \tag{4}$$

gdzie masa zredukowana $\mu = mM/(m+M)$. Jak wynika ze wzoru, energia elektronu zależy wyłącznie od liczby kwantowej n (rys. 3). Tak więc różne funkcje własne Ψ_{nlm} różniące się liczbami l i m, ale mające takie samo n, mają identyczną wartość własną (liczba stanów o takiej samej energii E_n wynosi n^2). Inaczej można powiedzieć, że różne stany elektronu w atomie jednoelektronowym mogą mieć takie same energie. Takie wypadki nazywamy degeneracją stanów (w przypadku atomów jednoelektronowych mówimy o degeneracji typu l oraz m).

degeneracja stanów

Nie będziemy podawać ogólnej postaci analitycznej rozwiązania (3) dla elektronu w atomie. Podamy jedynie kilka funkcji Ψ_{nlm}, a właściwie pewne ich kombinacje. Okazuje się, że funkcje Y_{lm} występujące w funkcji Ψ_{nlm} dla $m \geqslant 1$ są funkcjami zespolonymi (a więc i funkcje Ψ_{nlm} są wówczas zespolone). Dla celów praktycznych w wielu wypadkach wygodniej jest posługiwać się funkcjami rzeczywistymi (można je wówczas łatwo przedstawić graficznie). Dlatego parę funkcji Ψ_{nlm} i $\Psi_{nlm'}$ różniących się znakiem liczby m (tzn. $m' = -m$) zastępujemy parą funkcji

Funkcje falowe elektronu atomu jednoelektronowego oraz powierzchnie graniczne

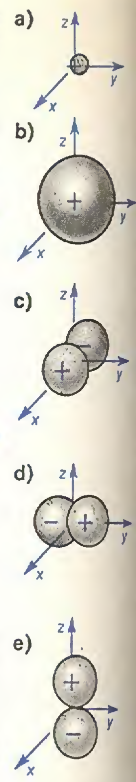
Funkcja falowa	Powierzchnia graniczna
$\Psi_{100} \equiv \Psi_{1s} \equiv 1s = N_{1s}e^{-Zr/a_0}$	a)
$\Psi_{200} \equiv \Psi_{2s} \equiv 2s = N_{2s}e^{-Zr/2a_0}(2-Zr/a_0)$	b)
Ψ_{211}: $\Psi_{2p_x} \equiv 2p_x = N_{2p}e^{-Zr/2a_0}x$	c)
Ψ_{21-1}: $\Psi_{2p_y} \equiv 2p_y = N_{2p}e^{-Zr/2a_0}y$	d)
$\Psi_{210} \equiv \Psi_{2p_z} \equiv 2p_z = N_{2p}e^{-Zr/2a_0}z$	e)

$a_0 = \hbar^2/me^2 = 0{,}0529$ nm (promień orbity Bohra),
$r = (x^2+y^2+z^2)^{1/2}$ (odległość elektronu od jądra),
Z — liczba porządkowa atomu (dla atomu wodoru $Z = 1$)

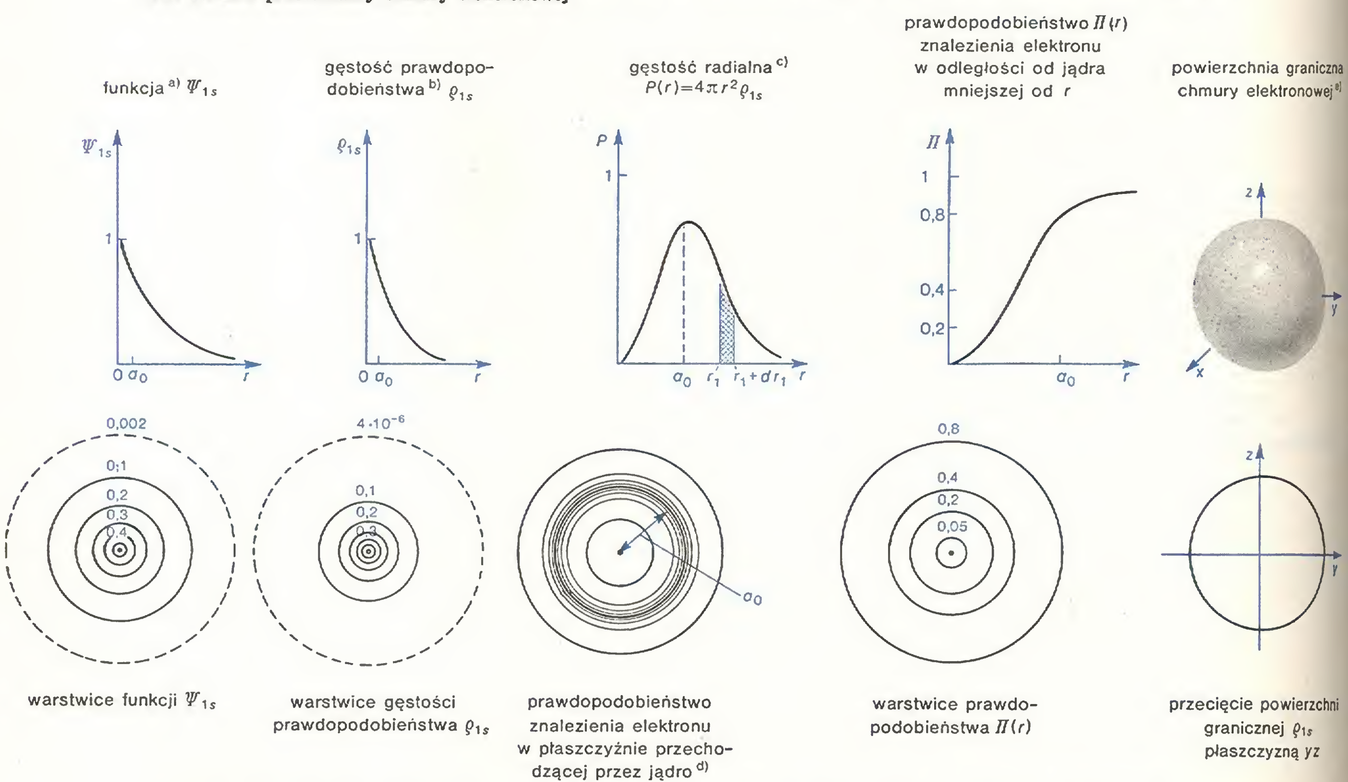
$$N_{1s} = \frac{1}{\sqrt{\pi}}\left(\frac{Z}{a_0}\right)^{3/2}, \quad N_{2s} = \frac{1}{4\sqrt{2\pi}}\left(\frac{Z}{a_0}\right)^{3/2},$$

$$N_{2p} = \frac{1}{4\sqrt{2\pi}}\left(\frac{Z}{a_0}\right)^{5/2}.$$

Funkcje falowe atomu wodoru oznacza się tradycyjnie w specjalny sposób. Każda funkcja, dla której $l = 0$ jest oznaczona symbolem s, funkcja, dla której $l = 1$ oznaczona jest symbolem p, dla $l = 2$ używa się symbolu d itp. Przed każdym z tych symboli umieszcza się wartość głównej liczby kwantowej. U dołu symboli p, d itd. zaznacza się własności kątowe funkcji. Dla funkcji typu ns, które nie zależą od kątów, nie umieszcza się żadnego indeksu. Funkcje falowe atomu wodoru można zapisywać w różny sposób, np. Ψ_{1s} lub po prostu $1s$.



Zależność funkcji falowej elektronu w atomie wodoru w stanie podstawowym i jej kwadratu od odległości od jądra oraz rozkład przestrzenny chmury elektronowej



[a]) Wartości Ψ_{1s} podane są w jednostkach $1/\sqrt{\pi a_0^3}$, [b]) Wartości ρ_{1s} podane są w jednostkach $1/\pi a_0^3$. [c]) Wielkość zakreskowanego pola przedstawia prawdopodobieństwo znalezienia elektronu w odległości między r_1 i r_1+dr_1 od jądra. [d]) Prawdopodobieństwo jest proporcjonalne do gęstości linii. Najbardziej prawdopodobnym miejscem znalezienia elektronu w przestrzeni jest powierzchnia kuli o promieniu a_0 (ściślej: w warstwie kulistej zawartej między kulami o promieniach $a_0-\varepsilon$ oraz $a_0+\varepsilon$, gdzie ε jest bardzo małe). [e]) Powierzchnia kuli, wewnątrz której znajduje się praktycznie cały ładunek. Atom wodoru w stanie podstawowym ma tylko 10% ładunku na zewnątrz kuli o promieniu $2{,}6a_0 = 0{,}14$ nm.

$\frac{1}{\sqrt{2}}(\Psi_{nlm}+\Psi_{nl-m})$ oraz $-\frac{1}{\sqrt{2}}(\Psi_{nlm}-\Psi_{nl-m})$.

Postępowanie takie jest uzasadnione matematycznie i podyktowane jest wyłącznie celami praktycznymi. W tabeli na stronie 252 przedstawiliśmy tylko kilka funkcji falowych. Czytelnik powinien pamiętać, że jest ich jednak nieskończenie wiele i oczywiście dla $n \geqslant 3$ postać funkcji jest bardziej skomplikowana aniżeli postać funkcji dla $n = 1$ i $n = 2$. Należy podkreślić jeszcze raz, że wspomniane funkcje (3) i energie (4) odpowiadające tym funkcjom są rozwiązaniami ścisłymi (1).

Atom wodoru

Elektron w stanie podstawowym atomu wodoru opisany jest funkcją falową $\Psi_{1s} = (\pi a_0^3)^{-1/2}e^{-r/a_0}$, natomiast gęstość prawdopodobieństwa dla tego stanu wynosi $\varrho_{1s} = |\Psi_{1s}|^2 = (\pi a_0^3)^{-1}e^{-2r/a_0}$. Z postaci tych funkcji widać, że zarówno Ψ_{1s} jak i ϱ_{1s} zależą funkcyjnie jedynie od odległości elektronu od jądra atomu wodoru. Obie funkcje mają kulistą symetrię względem początku układu.

Istnieje kilka sposobów przedstawiania właściwości przestrzennych tych funkcji. Zarówno funkcja Ψ_{1s}, jak i jej kwadrat mają największą wartość w początku układu ($r = 0$), w miarę oddalania się od tego punktu maleją wykładniczo. Elektron wyobrażamy sobie jako rozmyty ładunek rozłożony w przestrzeni z gęstością prawdopodobieństwa $e\varrho$ (scałkowanie po całej przestrzeni gęstości $e\varrho$ daje $-e$, tj. ładunek elektronu; ładunek elektronu będziemy oznaczali dalej przez e, ale pamiętajmy, że ładunek elektronu jest ujemny). Elektron jest więc „chmurą elektronową" o gęstości $e\varrho$. Chmura ta w stanie podstawowym atomu wodoru jest sferycznie symetryczna.

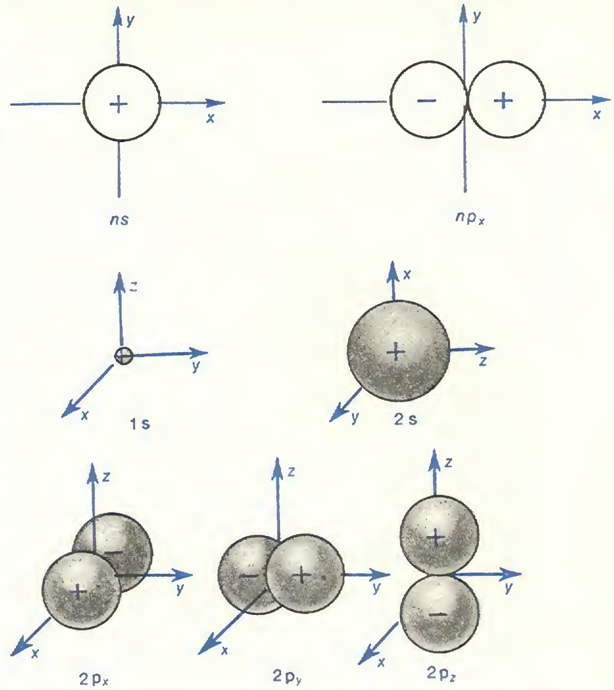
chmura elektronowa

Dokonując przekroju tej chmury dowolną płaszczyzną i łącząc punkty o jednakowej gęstości ładunku otrzymamy warstwice ładunku. W podobny sposób można przedstawić wartstwice funkcji falowej Ψ_{1s}. Interesujące jest prawdopodobieństwo znalezienia elektronu na pewnej odległości r od środka układu (ściślej w warstwie kulistej między kulami o promieniach r i $r+dr$). W tym celu należy obliczyć gęstość radialną $P(r)$ — w przypadku stanu podstawowego atomu wodoru $P(r) = \varrho_{1s}(r)4\pi r^2$. Funkcja ta ma dla $r = 0$ wartość 0, a na odległości a_0 ma maksimum. Mnożąc gęstość radialną $P(r)$ przez grubość warstwy kulistej dr otrzymujemy wspomniane prawdopodobieństwo znalezienia elektronu (ogólnie cząstki) w warstwie kulistej. Jeżeli wielkość $P(r)dr$ scałkujemy po r od wartości 0 do jakiejś wartości r', to otrzymamy prawdopodobieństwo znalezienia elektronu wewnątrz kuli o promieniu r': $\Pi(r') = \int_0^{r'} P(r)dr$. Ponieważ r' możemy wybierać w sposób dowolny, Π jest funkcją promienia. W przestrzeni funkcję $\Pi(r)$ dla danego r można przedstawić jako powierzchnię kuli o promieniu r. W miarę oddalania się od jądra prawdopodobieństwo znalezienia elektronu w kuli o promieniu r jest coraz większe. Dobierając promień tak, aby wewnątrz kuli znajdował się prawie cały ładunek (czyli $e\Pi(r) \approx e$), otrzymujemy kulę o powierzchni zwanej powierzchnią graniczną chmury elektronowej lub po prostu — powierzchnią graniczną, lub konturem granicznym.

powierzchnia graniczna

Dla celów porównawczych promień kuli dobiera się zwykle tak, aby powierzchnia graniczna obejmowała 90–95% ładunku. Tego typu powierzchnie graniczne wystarczą do zrozumienia wielu podstawowych zagadnień strukturalnych chemii bez szczegółowego rozpatrywania warstwic ładunku. W wielu wypadkach wystarczy nawet znajomość tylko powierzchni granicznej.

Powierzchnie graniczne (kontury) można oczywiście wykreślić nie tylko dla kwadratów funkcji falowych, ale również dla samych funkcji. Kształty konturów w obu wypadkach są takie same. Istotna różnica polega na tym, że funkcja falowa może w pewnych obszarach być dodatnia (dla stanu podstawowego atomu wodoru funkcja Ψ_{1s} przyjmuje zawsze wartości dodatnie), a w innych ujemna, podczas gdy gęstość prawdopodobieństwa jest zawsze wielkością dodatnią.



Rys. 4. Powierzchnie graniczne funkcji falowych elektronu w atomie wodoru dla $n = 1$ i $n = 2$. U dołu pokazano przecięcie powierzchni granicznej funkcji typu ns dowolną płaszczyzną przechodzącą przez jądro oraz przecięcie funkcji typu np_x płaszczyzną xy. Funkcje typu ns (1s, 2s, ...) są funkcjami kulistosymetrycznymi (nie zależą one od kątów). Funkcje typu np_x, np_y i np_z zależą od kątów i są antysymetryczne względem odbicia w płaszczyźnie odpowiednio yz, xz i xy. Znaki + oraz − oznaczają wartości dodatnie i ujemne funkcji falowych

Omawialiśmy dotychczas stan podstawowy atomu wodoru opisany funkcją falową Ψ_{1s}. W podobny sposób można przedstawić funkcje falowe opisujące poszczególne stany wzbudzone atomu wodoru.

stany wzbudzone atomu wodoru

Z różnego rodzaju przedstawień graficznych funkcji falowych najbardziej użyteczne są kontury funkcji falowych (lub ich przekroje płaszczyzną; rys. 4). Kontur każdej funkcji typu s, tzn. Ψ_{1s}, jak również Ψ_{2s}, Ψ_{3s} itd., ma kształt kulisty, a kule te są tym większe, im większa jest wartość głównej liczby kwantowej n. Jest to związane z faktem, że w wyższych stanach „chmura elektronowa" jest bardziej rozmyta w przestrzeni, należy więc przestrzeń otoczyć kulą o większym promieniu, aby wewnątrz tej kuli znalazł się prawie cały ładunek elektronowy. Najbardziej prawdopodobna odległość elektronu od jądra w stanie Ψ_{2s} atomu wodoru wynosi $(3+\sqrt{5})a_0$, czyli jest ponad pięć razy większa aniżeli analogiczna odległość w stanie podstawowym Ψ_{1s}. Funkcje falowe dla $n = 2$ i $l = 1$ zależą od kątów (innymi słowy są one ukierunkowane).

Na przykład funkcja Ψ_{2p_x} jest proporcjonalna do x, i znika dla $x = 0$. Dla dodatnich wartości x funkcja ta jest dodatnia, natomiast dla ujemnych x przyjmuje wartości ujemne. Te własności kątowe funkcji oznaczone są na powierzchni granicznej znakami + i −. Łatwo prześledzić w analogiczny sposób własności przestrzenne funkcji Ψ_{2p_y} i Ψ_{2p_z}. Powierzchnie graniczne dla funkcji falowych, dla których $n \geqslant 3$ i $l \geqslant 2$, są bardziej skomplikowane. Nie będziemy ich jednak tutaj przedstawiać.

Spin elektronu

Badania doświadczalne nad wpływem pola magnetycznego na własności atomów srebra wykazały, że elektron ma pewien własny moment pędu nazwany spinem. Nie będziemy szczegółowo uzasadniać istnienia tej wielkości (wymagałoby to odwołania się do relatywistycznej mechaniki kwantowej); dla naszych celów wystarczy wiedzieć, że spin elektronu jest pewnym dodatkowym momentem pędu nie związanym z ruchem orbitalnym elektronu wokół jądra. Czasami tłumaczy się istnienie spinu jako wynik obrotu elektronu dookoła własnej osi, ale nie należy tego brać zbyt dosłownie, gdyż spin nie ma swojego odpowiednika w mechanice klasycznej.

spin

Do pełnego opisu stanu elektronu nie wystarczy podać funkcję falową Ψ_{nlm} zależną od trzech współrzędnych przestrzennych r, ϑ, φ. Poprawna funkcja falowa powinna zawierać również informacje o spinie elektronu. W tym celu wystarczy funkcję Ψ_{nlm} pomnożyć przez funkcję opisującą stan spinu elektronu (jest to słuszne tylko w nierelatywistycznej mechanice kwantowej, którą się właśnie zajmujemy). Spin jednego elektronu jest jednoznacznie określony przez podanie tzw. rzutu spinu na wyróżnioną oś; rzut ten może przyjmować tylko dwie wartości $m_s = +{}^1/_2\hbar$ lub $m_s = -{}^1/_2\hbar$ (albo 1/2 lub −1/2 w jednostkach $\hbar$). Mówimy, że „elektron ma spin 1/2 lub −1/2", i tę sytuację charakteryzujemy za pomocą strzałek ↑ lub ↓. Umówimy się dalej, że funkcję spinową opisującą stan, w którym $m_s = 1/2$, oznaczać będziemy przez α, natomiast przez β oznaczymy funkcję spinową elektronu, dla którego $m_s = -1/2$. Tak więc poprawna funkcja falowa opisująca stan elektronu w atomie jednoelektronowym ma postać.

$$\lambda_{nlmm_s} = \begin{cases} \Psi_{nlm}\alpha & \text{dla } \uparrow \\ \Psi_{nlm}\beta & \text{dla } \downarrow \end{cases}. \tag{5}$$

Pojęcie spinu w zasadzie nie jest potrzebne dla zrozumienia własności stanów elektronu i ich energii w przypadku atomu jednoelektronowego. Jest ono natomiast konieczne przy opisie zespołu elektronów. Zanim zaczniemy omawiać bardziej złożone układy, wprowadzimy jeszcze pojęcie multipletowości stanu. Jest to wielkość określona wyłącznie własnościami spinów poszczególnych elektronów układu. Jak już wspominaliśmy, każdy elektron ma spin 1/2 lub −1/2. Elektrony można więc podzielić na dwie grupy — elektrony o spinach 1/2 i elektrony o spinach −1/2. Multipletowością stanu (lub jego krotnością) nazywamy wielkość $M = 2\left|\sum m_s\right| + 1$, gdzie suma przebiega po wszystkich elektronach układu. Tak więc stan elektronu w atomie jednoelektronowym ma multipletowość równą 2 (stan dubletowy), gdyż $M = 2 \cdot 1/2 + 1 = 2$. Dla dwóch elektronów możemy mieć dwie multipletowości. Jeżeli elektrony mają taki sam spin, wówczas $M = 3$ (stan trypletowy), a gdy spiny są przeciwne, wówczas mamy stan singletowy, gdyż $M = 1$. Bardzo łatwo jest obliczyć multipletowość dowolnego stanu układu elektronów, jeżeli tylko wiemy, jakie są spiny poszczególnych elektronów.

multipletowość (krotność) stanu

stan dubletowy

stan trypletowy

stan singletowy

Metody przybliżone mechaniki kwantowej

Jak już poprzednio wspomnieliśmy, problem ruchu elektronu w polu działania jądra atomowego o ładunku $+Ze$ (inaczej — zagadnienie ruchu dwóch cząstek) ma ścisłe rozwiązanie w mechanice kwantowej — równania (3) i (4). Jest to problem analogiczny do problemu ruchu dwóch ciał w mechanice klasycznej. Na przykład równania Newtona, które opisują ruch Ziemi wokół Słońca z pominięciem wpływu innych ciał Układu Słonecznego, mają rozwiązania ścisłe. Dobrze wiadomo jednak, że ruch Ziemi w Układzie Słonecznym zależy nie tylko od oddziaływania pomiędzy Słońcem a Ziemią, lecz także od wpływu wszystkich planet i innych ciał Układu Słonecznego i chociaż potrafimy sformułować równania Newtona opisujące ruch wszystkich ciał, nie potrafimy rozwiązać tych równań w sposób ścisły.

Podobna sytuacja występuje w mechanice kwantowej, gdzie nawet dla atomu helu (trzy cząstki: jądro + +dwa elektrony), dla którego znamy równanie Schrödingera (rys. 5), nie potrafimy tego równania rozwiązać w sposób ścisły.

Rozwiązanie numeryczne równania Schrödingera

Czytelnik interesujący się rozwojem i możliwościami komputerów zaproponuje niewątpliwie numeryczne rozwiązanie równania (1) dla atomu helu. I rzeczywiście tak można uczynić. W tym celu należy ustalić w przestrzeni położenie np. elektronu drugiego i rozwiązać numerycznie równanie (1), opisujące ruch elektronu pierwszego znajdującego się w polu działania jądra oraz „unieruchomionego" elektronu drugiego. Pole działające na elektron pierwszy nie jest już polem centralnym (jak to było w wypadku elektronu w atomie jednoelektronowym) i rozwiązanie równania (1) nie ma postaci analitycznej; jest ono przedstawione w postaci zbioru wartości funkcji falowej dla poszczególnych położeń elektronu pierwszego w przestrzeni (rozwiązanie numeryczne). Rozwiązanie będzie tym bardziej dokładne, im więcej zostanie podanych wartości funkcj ifalowej dla różnych położeń elektronu pierwszego.

Przyjmijmy, że zadowoli nas podanie wartości funkcji falowej w 1000 punktach przestrzeni. Ponieważ atom helu ma dwa elektrony, to aby mieć rozwiązanie numeryczne opisujące ruch obu elektronów w atomie, należy drugi elektron umieścić w 1000 punktach przestrzeni i dla każdego „ustalonego" położenia rozwiązywać równanie ruchu dla pierwszego elektronu w 1000 punktach. Przedstawienie takiej funkcji falowej w postaci książkowej wymagałoby wydania tomu o 1000 stronach, z 1000 wartości funkcji na każdej stronie. Obliczona w ten sposób funkcja falowa dla atomu litu (3 elektrony) wymagałaby całego księgozbioru (1000 książek 1000-stronicowych, a na każdej stronie 1000 wartości funkcji!!!). Proszę wyobrazić sobie podobne przedstawienie funkcji falowej dla cząsteczki benzenu zawierającej 12 jąder i 42 elektrony. Nic dziwnego, że ten sposób przedstawienia stanów układu chemicznego jest nie tylko niepopularny, ale jest zupełnie bezużyteczny. Na szczęście mechanika kwantowa dysponuje metodami przybliżonymi, które pozwalają opracować metody obliczeniowe, za pomocą których w wielu wypadkach można uzyskać wyniki zgodne z danymi doświadczalnymi.

Rachunek zaburzeń

Pierwszą z metod przybliżonych mechaniki kwantowej jest rachunek zaburzeń. Omówimy schematycznie tylko jedną z wersji tej metody, tzw. rachunek zaburzeń Rayleigha–Schrödingera. Rozważmy układ chemiczny opisany równaniem falowym (1) $\hat{H}\Psi = E\Psi$. Dowolny stan układu opisany jest funkcją falową Ψ_i o energii E_i (por. wzory (1) i (2)). Oczywiście nie znamy rozwiązania równania $\hat{H}\Psi = E\Psi$, ale szukamy tego rozwiązania w szczególny sposób. Przyjmijmy mianowicie, że hamiltonian układu $\hat{H}$ można zapisać w postaci sumy operatora $\hat{H}^{(0)}$ i operatora $\hat{H}'$, przy czym załóżmy, że znamy rozwiązanie równania falowego w przypadku operatora $\hat{H}^{(0)}$, tzn. znamy rozwiązanie równania $\hat{H}^{(0)}\varphi^{(0)} = \mathcal{E}^{(0)}\varphi^{(0)}$ (znamy więc

rachunek zaburzeń Rayleigha-Schrödingera

operator zaburzeniowy

zbiór funkcji własnych $\varphi_k^{(0)}$ i energii $\mathcal{E}_k^{(0)}$). Podziału hamiltonianu $\hat{H}$ na dwie części należy dokonać w taki sposób (istnieje kilka sposobów tego podziału), aby operator $\hat{H}'$ (operator zaburzeniowy) był mały w stosunku do operatora $\hat{H}^{(0)}$. Wówczas szukana funkcja falowa Ψ_i i energia E_i układu dają się wyrazić jako sumy odpowiednich wielkości niezaburzonego układu, tzn. $\varphi_i^{(0)}$ i $\mathcal{E}_i^{(0)}$ oraz poprawek zaburzeniowych różnego rzędu:

$$\begin{aligned} \Psi_i &= \varphi_i^{(0)} + \delta\Psi_i^{(1)} + \delta\Psi_i^{(2)} + \delta\Psi_i^{(3)} + \dots \\ E_i &= \mathcal{E}^{(0)} + \delta E_i^{(1)} + \delta E_i^{(2)} + \delta E_i^{(3)} + \dots \end{aligned} \tag{6}$$

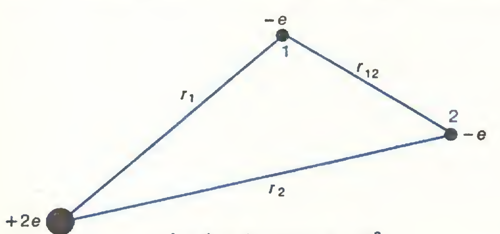
gdzie

$$\delta\Psi_i^{(1)},\ \delta\Psi_i^{(2)},\ \dots,\ \delta E_i^{(1)},\ \delta E_i^{(2)},\ \dots$$

są poprawkami 1., 2., ... rzędu odpowiednio do funkcji falowej i energii układu w i-tym stanie. Kolejne poprawki $\delta\Psi_i$ i δE_i wyrażone są przez całki, których wartość zależy jedynie od hamiltonianu zaburzeniowego $\hat{H}'$ oraz od funkcji własnych $\varphi_k^{(0)}$ i energii własnych $\mathcal{E}_k^{(0)}$ układu niezaburzonego. Stosunkowo prostą postać ma poprawka zaburzeniowa 1. rzędu do energii. Wyraża się ona prostą całką $\delta E_i^{(1)} = \int \varphi_i^{(0)*} \hat{H}' \varphi_i^{(0)} dv$, a więc poprawka ta zależy jedynie od $\hat{H}'$ oraz od jednej funkcji falowej $\varphi_i^{(0)}$ układu niezaburzonego. Niestety, istotną wadą rachunku zaburzonego w podanym tu sformułowaniu jest to, że w przypadku poprawek 1. rzędu do funkcji falowej oraz poprawek wyższych rzędów do energii i funkcji falowych, musimy posługiwać się sumami nieskończonymi (musimy znać wszystkie funkcje własne i wszystkie energie własne układu niezaburzeniowego). Niemniej, wspomniany rachunek zaburzeniowy umożliwia w prosty sposób obliczenie poprawki 1. rzędu do energii, i w wielu wypadkach energia układu $E_i \approx \mathcal{E}_i^{(0)} + \delta E_i^{(1)}$ stosunkowo dobrze zgadza się z danymi doświadczalnymi.

stan podstawowy atomu helu

Zilustrujemy powyższe rozważania na przykładzie stanu podstawowego atomu helu. W atomie tym (rys. 5) operatorem $\hat{H}^{(0)}$ może być suma energii kinetycznej obu elektronów ($\hat{T}_1 + \hat{T}_2$) i energii potencjalnej



$$\hat{H} = \hat{T}_1 + \hat{T}_2 + \hat{V}_1 + \hat{V}_2 + \frac{e^2}{r_{12}}$$

Rys. 5. Schematyczny obraz atomu helu; pod spodem hamiltonian opisujący ruchy elektronów w tym atomie, $\hat{T}_1$ i $\hat{T}_2$ operatory energii kinetycznej elektronu *1* i *2*, $\hat{V}_1$ i $\hat{V}_2$ operatory elektrostatycznego oddziaływania pomiędzy jądrem i elektronem *1* oraz jądrem i elektronem *2*, e^2/r_{12} oddziaływanie między elektronami. Gdyby usunąć jeden z elektronów, tzn. gdyby zjonizować atom helu, to ruch pozostałego elektronu opisany byłby w sposób ścisły funkcjami (3) o energiach ze wzoru (4) przy $Z = 2$

obu elektronów w polu jądra atomu helu ($\hat{V}_1 + \hat{V}_2$). Operatorem zaburzeniowym jest wówczas energia potencjalna oddziaływania elektrostatycznego pomiędzy elektronami e^2/r_{12}. Ponieważ w operatorze $\hat{H}^{(0)}$ nie ma oddziaływania pomiędzy elektronami, równanie falowe $\hat{H}^{(0)}\varphi^{(0)} = \mathcal{E}^{(0)}\varphi^{(0)}$ rozbija się na dwa równania, z których każde opisuje elektron *1* lub elektron *2* w polu działania jądra o ładunku $+2e$. A więc każde z tych równań opisuje przypadek atomu jednoelektronowego dla $Z = 2$. Rozwiązanie tych równań ma postać ścisłą: funkcje falowe mają postać (3), natomiast energie można obliczyć ze wzoru (4) (należy oczywiście za Z podstawić 2). Układ dwóch elektronów atomu helu, w wypadku gdy pomiędzy elektronami pominięto oddziaływanie, opisany jest funkcją falową, która jest iloczynem funkcji falowych opisujących oddzielnie każdy z elektronów w jonie atomu helu, natomiast energia układu tych elektronów jest sumą energii obu elektronów w poszczególnych stanach jonu. Dla stanu podstawowego energia obu elektronów (energia niezaburzonego układu) wynosi $\mathcal{E}_1^{(0)} = 2E_1 = 2Z^2(-13{,}6\text{ eV}) = -108{,}8$ eV (lub w tzw. jednostkach atomowych energii $-4{,}0$ j.at.).

jednostki atomowe

Jednostki atomowe wprowadza się w celu uniezależnienia wielkości charakteryzujących układ (energia, odległości między cząstkami układu itp.) od dokładności pomiaru stałych uniwersalnych, jak ładunek i masa elektronu, stała Plancka. Przyjmując, że $m = 1$, $e = 1$ oraz $\hbar = 1$, energia elektronu w stanie podstawowym wodoru wynosi $E_1 = -0{,}5$ j.at. $\approx -13{,}6$ eV. Tak więc 1 j.at. energii $\approx 27{,}2$ eV. Za jednostkę atomową odległości przyjmuje się a_0 (1 bohr) $= 0{,}0529$ nm.

Doświadczalna wartość energii obu elektronów w stanie podstawowym atomu helu wynosi $-78{,}98$ eV ($-2{,}90$ j.at.). Jak widać, jeżeli pominiemy oddziaływanie pomiędzy elektronami, opis własności obu elektronów w atomie helu nie jest poprawny. Obliczona energia układu niezaburzonego różni się o przeszło 30% od energii doświadczalnej. Jak już wspominaliśmy poprzednio można łatwo obliczyć poprawkę 1 rzędu do energii. Funkcją falową układu niezaburzonego jest w przypadku stanu podstawowego helu iloczyn dwóch funkcji Ψ_{100} ze wzoru (3) dla $Z = 2$ (tabela str. 252). Obliczona poprawka 1 rzędu do energii stanu podstawowego jest równa $\delta E_1^{(1)} = 34$ eV, a więc energia całkowita atomu helu obliczona z dokładnością do poprawki 1 rzędu wynosi $E_1 = \mathcal{E}_1^{(0)} + \delta E_1^{(1)} = -74{,}8$ eV ($-2{,}75$ j.at.), i zgadza się stosunkowo dobrze z wartością doświadczalną (obliczona wartość energii stanowi 95% energii doświadczalnej).

Rachunek zaburzeń ograniczony do poprawki 1 rzędu do energii, nie daje bardzo dokładnych wyników. Nic zresztą dziwnego, gdyż energia zaburzeń wcale nie jest małą i wynosi przeszło 30% energii rzędu zerowego. Natomiast uwzględnienie poprawki 2 rzędu do energii daje już energię całkowitą stanu podstawowego atomu helu z dokładnością do 0,5% wartości doświadczalnej. Istnieje wiele sformułowań rachunku zaburzeń, i w różnych wersjach był on stosowany do obliczenia m.in. energii stanu podstawowego atomu helu. W jednej z tych wersji poprawka 18 rzędu do energii wynosi $5 \cdot 10^{-9}$ eV. Pomimo trudności z obliczaniem poprawek wyższych rzędów rachunek zaburzeń ma duże znaczenie w mechanice kwantowej, a szczególnie przy opisie oddziaływań międzycząsteczkowych lub przy opisie wpływu pól zewnętrznych działających na rozważany układ.

Zasada wariacyjna

Drugą z metod przybliżonych mechaniki kwantowej jest metoda oparta na zasadzie wariacyjnej. Jest ona w wielu wypadkach bardziej dogodna od rachunku zaburzeń, gdyż można ją stosować do dowolnego układu, dla którego szukamy rozwiązania (2) równania falowego (1). Opis tej metody ograniczymy do stanu podstawowego układu E_1, a potem zilustrujemy przykładowymi obliczeniami dla stanu podstawowego atomu helu.

funkcja próbna

Wybierzmy dowolną funkcję próbną Φ, która w sposób przybliżony opisuje stan podstawowy rozważanego układu. Z zasady wariacyjnej wynika, że obliczona energia stanu podstawowego układu opisanego tą próbną funkcją nie jest nigdy mniejsza niż dokładna energia E_1 z dokładnego rozwiązania równania (1), tzn.

$$\mathcal{E} = \int \Phi^* \hat{H} \Phi dv \geqslant E_1. \tag{7}$$

Próbne funkcje falowe Φ można wybierać różnymi sposobami. Jeżeli wybierzemy kilka funkcji próbnych $\Phi_1, \Phi_2, \Phi_3, \dots$, to z wszystkich obliczonych energii $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \dots$ najlepsze przybliżenie do energii dokładnej E_1 będzie mieć energia $\mathcal{E}_i$ o najmniejszej wartości.

Na ogół jednak postępujemy trochę inaczej. Uzależniamy wybór funkcji próbnej Φ opisującej układ n elektronów od parametrów $c_1, c_2, \dots, c_k$, tzn. funkcja falowa powinna mieć postać $\Phi(1, 2, \dots, n;\ c_1, \dots, c_k)$. Mamy więc nie jedną funkcję Φ, ale cały zbiór funkcji, ponieważ dla każdego zbioru parametrów mamy inną funkcję. Energia układu opisanego próbną funkcją zależy oczywiście od tych parametrów, tzn. $\mathcal{E} = \mathcal{E}(c_1, c_2, \dots, c_k)$. Z zasady wariacyjnej wynika po pierwsze to, że należy tak dobrać parametry $c_1, \dots, c_k$, aby energia $\mathcal{E}$ była najmniejsza, ponieważ ta najmniejsza wartość energii $\mathcal{E}_{\min}(c_1, \dots, c_k)$ będzie najbliższa wartości dokładnej, a po drugie to, że przy wprowadzeniu dodatkowych parametrów $c_{k+1}, \dots, c_{k+l}$ otrzymujemy $\mathcal{E}_{\min}(c_1, c_2, \dots, c_{k+l}) < \mathcal{E}_{\min}(c_1, \dots, c_k)$ i jednocześnie funkcję falową $\Phi(1, 2, \dots, n;\ c_1, \dots, c_{k+l})$ o lepszym przybliżeniu do funkcji falowej opisującej dokładnie stan układu. A więc zwiększenie ilości tzw. parametrów wariacyjnych w funkcji falowej pozwala obliczyć energię całkowitą układu z coraz większą dokładnością (należy przy tym pamiętać, że wartość $\mathcal{E}_{\min}(c_1, c_2, \dots)$ nigdy nie może być mniejsza od wartości dokładnej E_1!!!). Jeżeli więc użyjemy bardzo wielu parametrów w próbnej funkcji falowej, uzyskać możemy wartość energii układu bardzo zbliżoną do energii dokładnej E_1 (czasami takie obliczenia są bardzo kłopotliwe z powodu dużej liczby niezbędnych parametrów).

parametry wariacyjne

Funkcję próbną Φ można w różny sposób uzależniać od parametrów c_p. Szczególnie prosta sytuacja jest wówczas, gdy funkcja próbna Φ jest przedstawiona w postaci liniowej kombinacji znanych funkcji χ_p

$$\Phi = \sum_{p=1}^{m} c_p \chi_p. \tag{8}$$

Mówimy wówczas, że funkcja Φ zależy liniowo od parametrów c_p. Zbiór funkcji χ_p może być zbiorem dowolnych znanych funkcji, ale przy jego wyborze kierujemy się oczywiście intuicją fizyczną lub chemiczną. Zasada wariacyjna w przypadku funkcji Φ postaci (8) daje prosty sposób (metoda Ritza), na znajdowanie współczynników c_p — należy mianowicie rozwiązać układ równań liniowych m-tego stopnia na szukane współczynniki

metoda Ritza

$$\sum_{p=1}^{m} c_p(H_{pq} - \mathcal{E} S_{pq}) = 0, \text{ dla } q = 1, 2, \dots, m, \tag{9}$$

gdzie $\quad H_{pq} = \int \chi_p^* \hat{H} \chi_q dv, \quad S_{pq} = \int \chi_p^* \chi_q dv.$

W równaniach tych H_{pq} i S_{pq} są całkami, które możemy obliczyć, gdyż zależą one od $\hat{H}$ oraz od znanych funkcji χ_p. Niewiadomymi wielkościami są współczynniki c_p, jak również $\mathcal{E}$, które daje się jednak wyznaczyć z warunku istnienia rozwiązania układu równań (9). Mianowicie układ ten ma rozwiązanie tylko wówczas, gdy znika wyznacznik zbudowany z wyrazów przy niewiadomych współczynnikach c_p (pomijamy tutaj trywialne (zerowe) rozwiązanie równań (9), gdy wszystkie c_p są równe 0), tzn. gdy

$$\det|H_{pq} - \mathcal{E} S_{pq}| = 0.$$

Ostatnie wyrażenie jest równaniem stopnia m na niewiadome $\mathcal{E}$. Rozwiązanie ma m pierwiastków $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_m$. Numerujemy je zgodnie ze wzrastającą energią; tak więc $\mathcal{E}_1$ będące najmniejszą liczbą ze wszystkich $\mathcal{E}_i$ jest energią stanu podstawowego układu opisanego próbną funkcją Φ postaci (8). Aby podać szczegółową postać tej funkcji, należy, wstawiając $\mathcal{E}_1$ do układu (9), znaleźć układ współczynników c_p odpowiadający danej energii $\mathcal{E}_1$. Oczywiście w ten sam sposób można obliczyć funkcje próbne Φ odpowiadające energiom $\mathcal{E}_2, \dots, \mathcal{E}_m$. Jak widać, każdemu $\mathcal{E}_i$ $(i = 1, \dots, m)$ odpowiada inna funkcja próbna Φ. Dlatego funkcję (8) oznaczamy dodatkowo indeksem i, tzn. zapisujemy ją w postaci $\Phi_i = \sum_{p=1}^{m} c_{pi}\chi_p$ (c_{pi} jest p-tym współczynnikiem do i-tej funkcji próbnej).

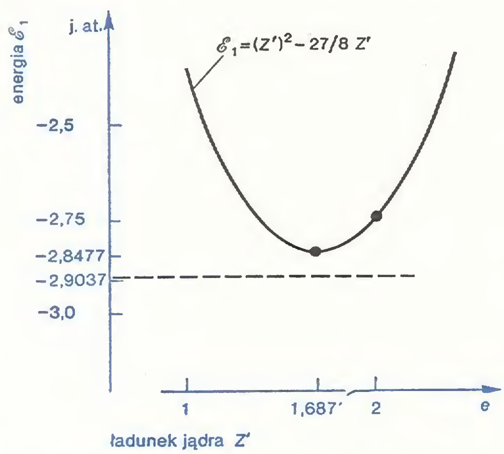
Rozpatrzmy teraz stan podstawowy atomu helu. W przypadku jonu He^+ wiadomo, że jeden elektron opisany jest w sposób ścisły funkcją falową $\Psi_{1s} = N_{1s}e^{-Zr/a_0}$, gdzie $Z = 2$. W przypadku atomu helu, gdy mamy w atomie dwa elektrony, możemy przyjąć, że w przybliżeniu własności każdego z elektronów opisane są funkcją falową Φ_{1s} podobną do funkcji Ψ_{1s}, ale nie mamy już powodu, aby przyjmować, że ładunek jądra wynosi $+2e$. Jeden z elektronów, powiedzmy elektron 1, „widzi" bowiem jądro atomu helu nie o ładunku $+2e$, ale o ładunku $Z'e$ gdyż elektron 2 ekranuje ładunek jądra. Analogicznie, elektron 2 „widzi" jądro atomu o ładunku $Z'e$, takim samym jak w przypadku rozważanego elektronu 1. Intuicyjnie należy oczekiwać, że Z' będzie zawarte pomiędzy 1 a 2, tzn. należy oczekiwać ładunku pośredniego między ładunkiem atomu wodoru a ładunkiem jonu He^+. Możemy więc przyjąć, że funkcje opisujące każdy z elektronów w atomie helu mają postać $\varphi_{1s} = N_{1s}e^{-Z'r/a_0}$. Funkcja falowa opisująca układ tych dwóch elektronów zależy więc bezpośrednio od współrzędnych obu elektronów oraz od jednego parametru wariacyjnego Z', tzn.

stan podstawowy atomu helu

$$\Phi(1, 2;\ Z') = \varphi_{1s}(1)\varphi_{1s}(2) = |N_{1s}|^2 e^{-Z'(r_1+r_2)/a_0}.$$

Energia układu opisanego taką funkcją falową wynosi $\mathcal{E}_1 = (Z')^2 - 27/8Z'$. Przyjmując więc różne ładunki Z', otrzymujemy różne funkcje próbne Φ $(1, 2;\ Z')$ i różne energie układu odpowiadające tym funkcjom. Najlepszą funkcją Φ $(1, 2;\ Z')$ jest funkcja dla takiego Z', dla którego $\mathcal{E}_1(Z')$ ma wartość minimalną (rys. 6),

obliczanie energii



Rys. 6. Energia $\mathcal{E}_1$ układu dwóch elektronów w stanie podstawowym atomu helu obliczona przy założeniu, że każdy z elektronów jest opisany funkcją $\varphi_{1s} = 1/\sqrt{\pi}\,(Z'/a_0)^{3/2}\,e^{-Z'r/a_0}$ (Z' ładunek jądra). Linia przerywana odpowiada dokładnej energii atomu helu

tzn. dla $Z' = 1{,}6875$ (jak widać Z' jest rzeczywiście zawarte pomiędzy 1 a 2). Przy tej wartości Z' energia $\mathcal{E}_1 = -2{,}8477$ j.at. stanowi 98,07% energii dokładnej, a więc różnica między wartością obliczoną a energią dokładną jest bardzo mała. Różnicę tę można jeszcze bardziej zmniejszyć, modyfikując funkcję próbną. Przyjmijmy np. że funkcja falowa opisująca jeden elektron w atomie helu ma postać $\varphi_{1s} = N'_{1s}r^{n'-1}\cdot e^{-Z'r/a_0}$, gdzie n' jest bliskie jedności, tak że czynnik $r^{n'-1}$ nieznacznie tylko zmienia poprzednio użytą funkcję φ_{1s} (zarówno n', jak i Z' traktujemy jako parametry wariacyjne). Funkcja układu dwóch elektronów, będąca iloczynem powyższych funkcji jednoelektronowych, ma wówczas postać Φ $(1, 2;\ n', Z')$. Okazuje się, że energia układu $\mathcal{E}_1$ dla tej funkcji jest minimalna wówczas, gdy $Z' = 1{,}61162$ oraz $n' = 0{,}955$, i wynosi $-2{,}8542$ j.at. Widać więc wyraźnie, że zwiększenie liczby parametrów w próbnej funkcji polepsza zgodność energii obliczonej z energią dokładną (obliczona energia $\mathcal{E}_1$ w drugim wypadku sta-

nowi już 98,30% energii dokładnej). Funkcję próbną można jeszcze bardziej uogólnić, mnożąc $\Phi\,(1,2;\ Z')$ np. przez wielomian typu $(1+c_1r_1^p+c_2r_2^q+c_3r_{12}^s)$, gdzie c_1, c_2, c_3 oraz p, q i s są parametrami. Energia $\mathcal{E}_1$ obliczona przy użyciu takiej funkcji stanowi jeszcze lepsze przybliżenie do energii dokładnej. W 1958 r. C. L. Pekeris użył w obliczeniach dla atomu helu funkcji próbnej zawierającej kilkaset parametrów. Obliczony przez niego potencjał jonizacyjny atomu helu, wynoszący 198310,67 cm^{-1}, nie tylko zgadza się z wartością doświadczalną 198310,8 $\pm$ 0,15 cm^{-1}, ale jest od niej bardziej dokładny!

potencjał jonizacyjny atomu helu

Porównywaliśmy wyniki obliczeń $\mathcal{E}_1$ za pomocą funkcji próbnej Φ z energią dokładną E_1. Dla atomu helu nie znamy funkcji falowej Ψ_1, a więc i nie znamy E_1. Co oznacza więc wspomniana energia dokładna wynosząca −2,9037 j.at.? Czy jest nią wartość energii otrzymana z pomiarów doświadczalnych? I tak, i nie. Mierząc wartość energii otrzymujemy rzeczywistą wielkość energii dla układu, który podlega prawom relatywistycznej mechaniki kwantowej. Ponieważ zajmujemy się tutaj mechaniką kwantową nierelatywistyczną, mierzoną wartość doświadczalną należy więc „przecechować" do wielkości, z którą możemy porównywać wyniki otrzymane z rozwiązania nierelatywistycznego równania (1). W tym celu należy od zmierzonej wielkości doświadczalnej odjąć szereg poprawek, np. poprawki relatywistyczne i promieniste. Zmierzona energia stanu podstawowego atomu helu wynosi −2,9033 j.at., a wspomniane poprawki są małe, rzędu 0,0004 j.at. Tak więc cytowana już wcześniej wielkość −2,9037 j.at. jest „przecechowaną" energią doświadczalną. Warto dodać, że ponieważ obliczona przez Pekerisa wielkość energii stanu podstawowego atomu helu (−2,90372 j.at.) jest bardziej dokładna od wielkości mierzonej, z wynikiem teoretycznym Pekerisa porównywane są wszystkie inne wyniki przybliżonych obliczeń dla atomu helu.

Przybliżenie jednoelektronowe

Przybliżeniem oddającym ogromne usługi przy opisie własności układu elektronów w atomie lub w cząsteczce i będącym podstawą opisu struktur chemicznych jest tzw. przybliżenie jednoelektronowe. Zgodnie z nim, ruch każdego elektronu układu rozpatruje się w polu jądra (lub jąder) i w uśrednionym polu pochodzącym od pozostałych elektronów układu. Innymi słowy stan każdego elektronu układu opisany jest jednoelektronową funkcją falową, zależną od współrzędnych określonego elektronu oraz od jego spinu. Funkcje takie nazywamy spinorbitalami i oznaczamy je symbolem λ_i (czytelnik nie powinien utożsamiać spinorbitali λ_i z funkcją falową λ_{nlmm_s} (5), która opisuje w sposób ścisły stan elektronu w atomie jednoelektronowym). W wypadku pominięcia efektów relatywistycznych, spinorbital jest iloczynem orbitalu atomowego (AO — *atomic orbital*) lub orbitalu molekularnego (MO — *molecular orbital*) przez funkcję spinową α lub β (wzór 10).

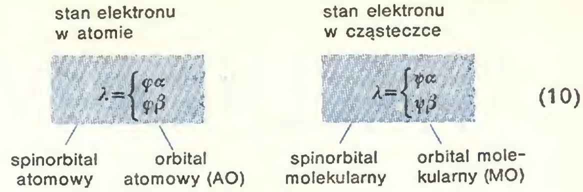
spinorbital

Orbitalem atomowym lub molekularnym nazywać będziemy tutaj funkcję zależną od współrzędnych jednego elektronu, opisującą ten elektron w atomie lub cząsteczce. Funkcję falową opisującą stan wszystkich n elektronów układu można zapisać w postaci wyznacznika:

AO — orbital atomowy
MO — orbital molekularny

$$\Phi\,(1,2,\ldots,n)=\frac{1}{\sqrt{n!}}\begin{vmatrix}\lambda_1(1) & \lambda_1(2) & \ldots & \lambda_1(n)\\ \lambda_2(1) & \lambda_2(2) & \ldots & \lambda_2(n)\\ \cdot & & & \\ \cdot & & & \\ \lambda_n(1) & \lambda_n(2) & \ldots & \lambda_n(n)\end{vmatrix} \tag{11}$$

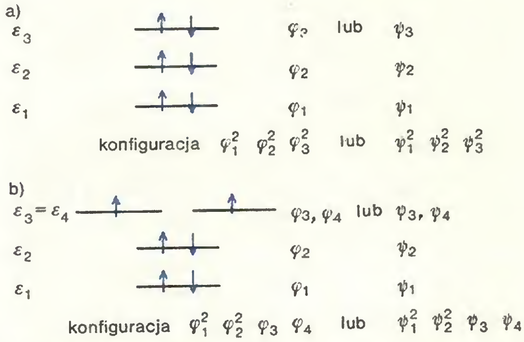
Przedstawienie funkcji falowej Φ w postaci (11) podyktowane jest ogólnymi własnościami układu n elektronów, takimi jak to, że: 1) elektrony są nierozróżnialne, 2) dowolne dwa elektrony układu nie mogą być opisane takim samym spinorbitalem λ_i — zasada Pauliego (ale ten sam AO lub MO może opisywać dwa elektrony, ponieważ w myśl (10) z każdego orbitalu



można utworzyć dwa spinorbitale) oraz 3) funkcja opisująca układ n elektronów musi być antysymetryczna ze względu na zamianę współrzędnych dwóch elektronów, tzn. np. $\Phi\,(1,2,\ldots,n) = -\Phi\,(2,1,\ldots,n)$.

W przybliżeniu jednoelektronowym stan podstawowy układu (rys. 7) można opisać wprowadzając konfigurację elektronową, która przedstawia uporządkowany ze względu na wzrastającą energię ciąg

konfiguracja elektronowa



Rys. 7. Konfiguracja opisująca w przybliżeniu stan podstawowy układu 6 elektronów w atomie (AO — φ_i) lub w cząsteczce (MO — ψ_i). Liczba elektronów na danym orbitalu podana jest u góry symbolu orbitalu (φ_i — jeden elektron na orbitalu φ_i, φ_i^2 — dwa elektrony na orbitalu φ_i); a) orbitale są niezdegenerowane; b) dwa orbitale są zdegenerowane

orbitali z podaniem „obsadzenia" poszczególnych orbitali przez elektrony. Zgodnie z zasadą Pauliego na każdym z orbitali „umieszczamy" po dwa elektrony o przeciwnych spinach, rozpoczynając „obsadzanie" od najniższego energetycznie orbitalu. Może się zdarzyć, że jakieś dwa orbitale mają taką samą energię orbitalną (degeneracja orbitali). Spośród kilku możliwości rozmieszczenia dwóch elektronów na zdegenerowanych orbitalach (pamiętajmy, że energia układu dwóch elektronów nie równa się sumie energii orbitalnych) wybieramy najbardziej korzystną energetycznie — jest nią sytuacja, gdy każdy z elektronów jest na różnych orbitalach, ale oba elektrony mają takie same spiny (reguła Hunda).

degeneracja orbitali

reguła Hunda

Równanie Hartree'ego–Focka

Aby otrzymać przynajmniej jakościowy opis konfiguracji elektronów w atomie lub cząsteczce i podać względne energetyczne rozmieszczenie orbitali, należy znać postać tych orbitali. Jeśli chodzi o atomy sytuacja jest stosunkowo prosta, gdyż należy intuicyjnie spodziewać się, że orbitale atomowe mogą mieć zbliżony charakter do funkcji falowych Ψ_{1s}, Ψ_{2s}, Ψ_{2p} itd. opisujących w sposób ścisły elektron w atome jednoelektronowym. Orbitale atomowe φ_i mogą np. różnić się ładunkiem efektywnym $Z'e$ (jak to było w przypadku atomu helu) w stosunku do ładunku Ze atomu jednoelektronowego (tabela str. 252). Ale czy takie orbitale atomowe są najlepszymi jednoelektronowymi funkcjami opisującymi elektron w atomie? Orbitale atomowe można przecież wybierać różnymi sposobami. Odpowiedź na pytanie wynika z za-

sady wariacyjnej. Szukamy próbnej funkcji falowej $\Phi(1, 2, ..., n)$ postaci (11) — należy więc tak zmieniać postać AO lub MO, aby energia konfiguracji była jak najmniejsza. Zasada wariacyjna pozwala otrzymać

równania HF

równania (tzw. równania Hartree'ego-Focka lub w skrócie równania HF), z rozwiązania których otrzymujemy orbitale atomowe φ_i lub molekularne ψ_i oraz odpowiadające im energie orbitalne ε_i. W przypadku gdy mamy parzystą liczbę elektronów oraz tzw. zamkniętopowłokową konfigurację stanu podstawowego (po dwa elektrony na każdym orbitalu), równania HF mają postać

zamkniętopowłokowa konfiguracja stanu podstawowego

$$\hat{F}\varphi_i = \varepsilon_i\varphi_i$$

lub (12)

$$\hat{F}\psi_i = \varepsilon_i\psi_i \text{ (dla } i = 1, ..., n/2).$$

Jest to postać równań bardzo zbliżona do ogólnego równania Schrödingera (1). Zasadnicza różnica polega na znacznym uproszczeniu równań HF. Podczas gdy równanie (1) było równaniem na znajdowanie jednej funkcji falowej dla całego układu, równania HF są układem $n/2$ równań na orbitale jednoelektronowe opisujące w sposób przybliżony każdy z elektronów oddzielnie.

operator Hartree'ego-Focka

Operator Hartree'ego-Focka $\hat{F}$ występujący w (12) zależy od energii kinetycznej jednego elektronu, od jego oddziaływania z jądrem atomu (lub jądrami cząsteczki) oraz od uśrednionego potencjału oddziaływania ze wszystkimi pozostałymi elektronami. Uśredniony potencjał jest niestety określony przez wszystkie orbitale, których szukamy. Oznacza to, że operator $\hat{F}$ zależy od współrzędnych określonego elektronu, np. 1 oraz parametrycznie od orbitali φ_i lub ψ_i, tzn. $\hat{F} = \hat{F}(1; \varphi_1, \varphi_2, ..., \varphi_{n/2})$ lub $\hat{F} = \hat{F}(1; \psi_1, ..., \psi_{n/2})$. Znalezienie rozwiązania układu równań (12) nie jest więc proste, rozwiązuje się je metodą iteracyjną.

W skrócie postępowanie iteracyjne jest następujące.

metoda iteracyjna

1. Zakładamy początkowe orbitale $\varphi_1^{(0)}, \varphi_2^{(0)}, ..., \varphi_{n/2}^{(0)}$; definiujemy $\hat{F}^{(0)}(1; \varphi_1^{(0)}, ..., \varphi_{n/2}^{(0)})$; rozwiązujemy układ równań $\hat{F}^{(0)}\varphi_i = \varepsilon_i\varphi_i$ dla $i = 1, ..., n/2$ znajdując orbitale $\varphi_1^{(1)}, \varphi_2^{(1)}, ..., \varphi_{n/2}^{(1)}$.

2. Definiujemy $\hat{F}^{(1)}(1; \varphi_1^{(1)}, \varphi_2^{(1)}, ... \varphi_{n/2}^{(1)})$; rozwiązujemy układ równań $\hat{F}^{(1)}\varphi_i = \varepsilon_i\varphi_i$ dla $i = 1, ..., n/2$ znajdując orbitale $\varphi_1^{(2)}, \varphi_2^{(2)}, ..., \varphi_{n/2}^{(2)}$.

3. Definiujemy $\hat{F}^{(2)}(1; \varphi_1^{(2)}, \varphi_2^{(2)}, ..., \varphi_{n/2}^{(2)})$, itd.

Znalezione w m-tym cyklu obliczeń orbitale $\varphi_1^{(m)}, ..., \varphi_{n/2}^{(m)}$ używamy do określenia operatora $\hat{F}^{(m)}$, i rozwiązujemy układ równań $\hat{F}^{(m)}\varphi_i = \varepsilon_i\varphi_i$ dla $i = 1, ..., n/2$ znajdując orbitale $\varphi_1^{(m+1)}, ..., \varphi_{n/2}^{(m+1)}$. Jeżeli ten układ orbitali nie różni się od orbitali $\varphi_i^{(m)}$ z poprzedniego cyklu, obliczenia przerywamy, gdyż dalsze postępowanie iteracyjne nie zmieni już postaci orbitali. Oczywiście tak samo rozwiązuje się równania HF w przypadku cząsteczki.

Orbitale pola samouzgodnionego SCF

Orbitale otrzymane w wyniku rozwiązania iteracyjnego równań (12) nazywa się orbitalami pola samouzgodnionego (SCF — *Self Consistent Field*), gdyż orbitale określają „samouzgodniony" potencjał uśredniony (pole) działający na poszczególne elektrony układu.

orbitale jednocentrowe orbitale wielocentrowe

Cząsteczka jest o wiele bardziej złożonym układem niż atom. W atomie położenia elektronów określone są względem jednego centrum (jądra atomu) — orbitale atomowe są więc jednocentrowe. W cząsteczce elektrony są własnością całej cząsteczki, a orbitale molekularne są wielocentrowe. Nie znamy jednak ich ogólnej postaci, dlatego też rozwiązywanie równań HF dla cząsteczek nastręcza duże trudności, czasami wręcz nie do pokonania. Wobec tego rozwiązujemy równania (12) w uproszczonej postaci (tzw. równania Hartree'ego-Focka-Roothaana lub po prostu równania HFR). Orbitale molekularne przybliża się przez liniową kombinację znanych funkcji jednoelektronowych

równania HFR

$$\psi_i = \sum_{p=1}^{m} c_{pi}\varphi_p. \tag{13}$$

Układ funkcji φ_p nazywa się bazą. Bardzo często jako bazę wybiera się orbitale atomowe φ atomów wchodzących w skład cząsteczki. Mówimy wówczas, że orbitale molekularne przybliżone są przez liniową kombinację orbitali atomowych (LCAO — *Linear Combination of AO*). Rozwiązanie równań HF z przybliżeniem (13) — równań HFR, daje wówczas orbitale molekularne, które nazywa się orbitalami SCF LCAO-MO. Warto dodać, że stosując coraz większą bazę gdy w wyrażeniu (13) $m \to \infty$ otrzymujemy rozwiązania odpowiadające równaniom HF (12). Użycie bazy nieskończonej jest oczywiście niemal niemożliwe (prowadzi to do rozwiązywania wyznacznika nieskończonego rzędu, por. (8) i (9)). Obliczenia wykazują jednak, że użycie dużej skończonej bazy dla małych układów (np. dla atomów) daje zbliżone wyniki do rozwiązania równań HF (12).

LCAO

orbitale SCF LCAO-MO

Atomy wieloelektronowe

Wspominaliśmy już o tym, że użycie przybliżenia jednoelektronowego dla układu n elektronów wyjaśnia wiele własności atomów lub cząsteczek. Jak opisujemy w tym przybliżeniu atomy wieloelektronowe? Na rys. 8 podano schematyczne przedstawienie konfiguracji stanów podstawowych atomów wodoru, litu, itd. aż do neonu. Podobne konfiguracje można zbudować dla pozostałych pierwiastków układu okresowego. Zanim omówimy krótko charakterystyczne własności konfiguracji elektronowych atomów, powiemy kilka słów o orbitalach atomowych potrzebnych do określania konfiguracji.

Orbitale atomowe

Orbitale atomowe mogą mieć różną postać — mogą to być np. funkcje „podobne" do funkcji (3) atomu wodoru. Tego typu orbitale nazywane są orbitalami wodoropodobnymi. Bardzo często używane są orbitale typu Slatera w postaci

orbitale Slatera

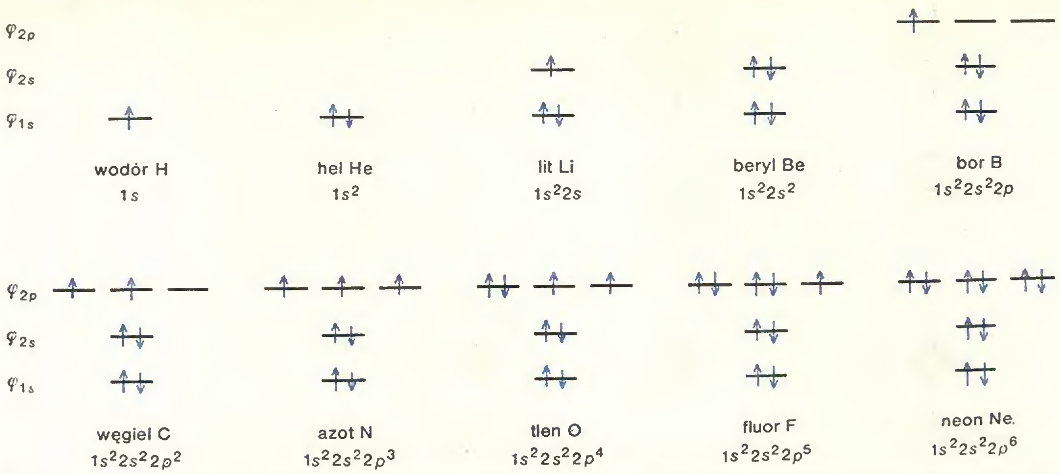
$$\varphi_{n'lm} = N_{n'}r^{n'-1}e^{-Z'r/n'a_0}Y_{lm}(\vartheta, \varphi), \tag{14}$$

gdzie $N_{n'}$ jest stałą normalizacyjną, n' jest tzw. efektywną liczbą kwantową (według reguł Slatera $n' = n$ dla $n = 1, 2$ i 3, ale n' nie musi być w ogólności liczbą całkowitą), natomiast $Z'e$ jest efektywnym ładunkiem jądra. Dla orbitali poszczególnych atomów łatwo znajduje się ładunki według reguł, które podał Slater (reguły te pominiemy, ponieważ nie będziemy z nich dalej korzystać). Część kątowa Y_{lm} orbitali Slatera jest taka sama, jak w funkcjach (3).

orbitale atomowe HF

Orbitale Slatera są tylko przybliżonymi orbitalami atomowymi, ale mają bardzo wygodną postać analityczną. Najlepszymi jednak orbitalami są funkcje otrzymane z rozwiązania równań HF (orbitale atomowe HF, φ^{HF}), które z kolei nie mają zbyt dogodnej postaci, wyniki bowiem są podane w postaci zbioru liczb (tabel), a nie w postaci funkcji analitycznych. Wcześniej już wspominaliśmy o zniechęcającym sposobie przedstawiania numerycznego równania (1) dla funkcji wieloelektronowej układu. Dla atomu helu np., taka funkcja wymagałaby zapisania całego tomu, dla atomu litu potrzebny już byłby cały księgozbiór. Jeśli chodzi o orbitale HF sytuacja jest o wiele prostsza. Do opisu atomu helu potrzebny jest jeden orbital; numeryczne przedstawienie tej funkcji wymaga-

więc np. tabeli z tysiącem wartości funkcji. Dla litu musimy znać dwa orbitale, a więc należy podać dwie podobne tabele, itd. Operowanie takimi tabelami



Rys. 8. Konfiguracje elektronowe pierwszych 10 atomów układu okresowego pierwiastków. Dla atomu wodoru orbital $1s$ jest ściślym rozwiązaniem równania (1). Dla pozostałych atomów orbitale $1s, 2s$ oraz $2p$ są przybliżonymi funkcjami jednoelektronowymi. W każdej konfiguracji uwzględniono względne rozmieszczenie energetyczne orbitali, czyli $\varepsilon_{1s} < \varepsilon_{2s} < \varepsilon_{2p_x} = \varepsilon_{2p_y} = \varepsilon_{2p_z}$, ale nie uwzględniono zróżnicowania energetycznego orbitali między różnymi konfiguracjami (rys. 10)

(funkcjami) jest również kłopotliwe. Ale postęp w technice obliczeniowej jest bardzo duży i nie tabelaryzuje się już funkcji HF dla atomów. Szybciej bowiem można taką funkcję obliczyć mając program obliczeniowy na komputer, aniżeli przepisać ją z tablicy bezbłędnie. Bardzo często funkcje HF zapisane są na taśmach magnetycznych, tak że wykorzystanie ich w obliczeniach jest proste.

rozwinięcie na orbitale Slatera

Istnieje jeszcze inny sposób postępowania. Za pomocą liniowej kombinacji funkcji typu Slatera (14) można przybliżyć orbitale HF z dowolnym stopniem dokładności. Orbital HF ma wówczas postać

$$\varphi_i^{HF} = \sum_{j=1}^{m} c_{ij}\varphi_j,$$

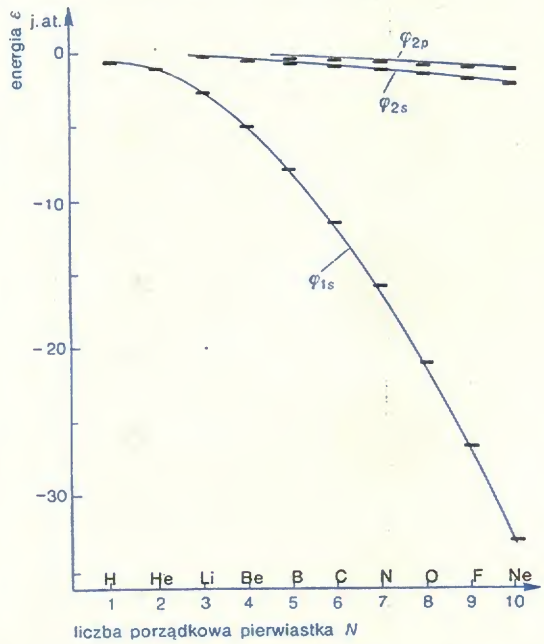
gdzie φ_j oznacza orbital Slatera (14) z parametrem Z' traktowanym jako nieznany. E. Clementi pokazał, że orbitale HF dla atomów od litu do neonu są bardzo dobrze przybliżone przez układ 9 funkcji Slatera. Energie całkowite atomów obliczone w tym przybliżeniu zgadzają się z energiami HF z dokładnością do 0,0001 j.at. Przybliżenie orbitali HF przez liniową kombinację orbitali atomowych (nie muszą to być koniecznie orbitale Slatera) jest korzystne, gdyż orbitale HF są wówczas scharakteryzowane przez podanie małego układu liczb (współczynniki c_{ij} oraz np. wartości parametrów Z' dla poszczególnych orbitali Slatera).

Orbitale atomowe mają pewną cechę wspólną. Kontury tych orbitali, jak również kontury gęstości są bardzo podobne do odpowiednich konturów atomu jednoelektronowego. Dlatego też poszczególne orbitale oznacza się symbolami φ_{1s}, φ_{2s}, φ_{2p} itd., a w przypadku orbitali Hartree'ego–Focka dodaje się dodatkowo indeks HF u góry orbitali φ_{1s}^{HF}, φ_{2s}^{HF} itd. Bardzo często skraca się cały zapis, oznaczając orbitale jedynie symbolami $1s$, $2s$, $2p$ ($2p_x$, $2p_y$, $2p_z$) itd. Tak więc orbitale atomowe użyte do konstrukcji konfiguracji elektronowych atomów (rys. 8) są orbitalami przybliżonymi albo orbitalami HF.

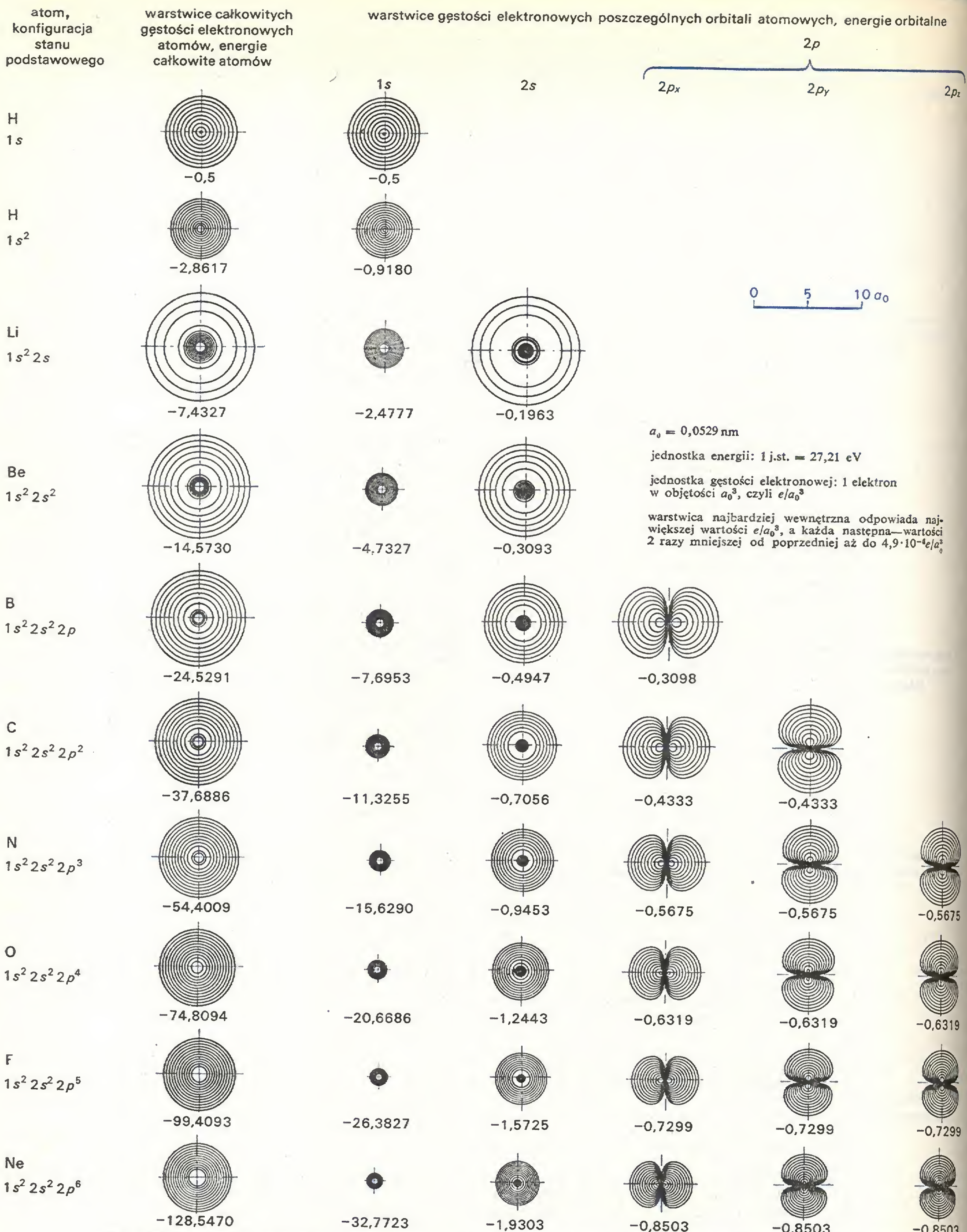
Na rys. 9 pokazano warstwice gęstości poszczególnych orbitali HF dla atomów od wodoru do neonu (w przypadku wodoru orbitale HF są ściślymi funkcjami (3)) oraz warstwice całkowitych gęstości elektronowych dla atomów. Widać wyraźnie, że zmieniają się rozmiary określonych orbitali przy przejściu od atomu do atomu, ale zasadniczo charakter przestrzenny jest taki sam, jak dla odpowiednich funkcji atomu jednoelektronowego (tabela str. 252 i rys. 4). Ze względu na to, że w atomach wieloelektronowych na każdy elektron atomu nie działa już potencjał centralny, jak to było w przypadku atomu jednoelektronowego, energia poszczególnych orbitali zależy od liczby kwantowej n oraz l (tzn. np. $\varepsilon_{2s} \neq \varepsilon_{2d}$, $\varepsilon_{3s} \neq \neq \varepsilon_{3p} \neq \varepsilon_{3d}$). Dla określonego atomu sekwencja energii orbitalnych jest następująca:

zależność energii od n, l i Z

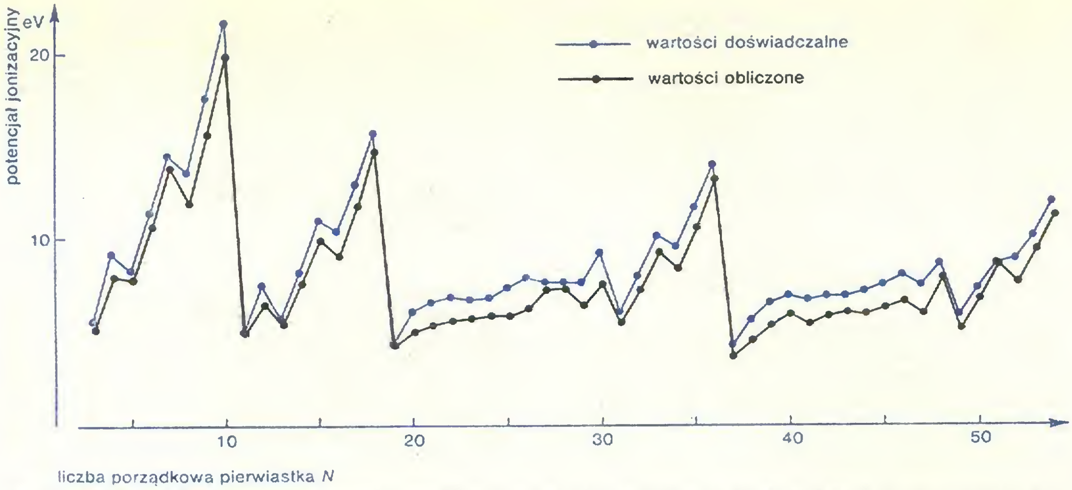
$\varepsilon_{1s} < \varepsilon_{2s} < \varepsilon_{2p}(\varepsilon_{2p_x} = \varepsilon_{2p_y} = \varepsilon_{2p_z}) < \varepsilon_{3s} < \varepsilon_{3p} < \varepsilon_{4s} \leqslant \leqslant \varepsilon_{3d}$ itd. Jeszcze jedną charakterystyczną cechą konfiguracji elektronowych jest to, że energie poszczególnych orbitali ulegają obniżeniu, gdy liczba elektronów w atomie zwiększa się. Widać to szczególnie wyraźnie na rys. 10. Bardzo znacznemu obniżeniu ulegają energie orbitalne ε_{1s}. Dla atomu wodoru ε_{1s}(H) = −0,5 j.at. (−13,6 eV), podczas gdy dla węgla ε_{1s}(C) = −11,3255 (−308,2), a dla azotu i tlenu energie ε_{1s} wynoszą odpowiednio −15,6290 (−425,3)



Rys. 10. Energie orbitali $1s$, $2s$ i $2p$ pierwszych 10 pierwiastków układu okresowego obliczone metodą HF (energia $\varepsilon_{1s} = -0{,}5$ j.at. = −13,6 eV dla atomu wodoru wynika ze ścisłego rozwiązania równania (1). Orbitale typu $2p$ każdego atomu są potrójnie zdegenerowane ($\varepsilon_{2p_x} = \varepsilon_{2p_y} = \varepsilon_{2p_z}$)



Rys. 9. Warstwice gęstości elektronowych (powiększenie ok. pół miliarda razy) pierwszych 10 atomów układu okresowego otrzymane za pomocą komputera z rozwiązania równań Hartree'ego–Focka. Dla każdego atcmu przedstawiono warstwice całkowitej gęstości oraz warstwice gęstości poszczególnych orbitali. Gęstości obliczano wprost z każdego orbitalu HF (przy obliczaniu całkowitej gęstości dokonywano uśrednienia sferycznego orbitali 2p, ponieważ elektrony mogą zajmować różne kombinacje orbitali 2p z równym prawdopodobieństwem — w atomie nie ma wyróżnionego układu współrzędnych). Rys. 9 oraz 31–34 i 37 wg A.C. Wahl, Scien. American 222, 54 (1970)



Rys. 11. Porównanie doświadczalnych wartości potencjałów jonizacyjnych z wartościami obliczonymi metodą HF dla atomów o liczbach porządkowych od 3 do 54 (dane liczbowe wg S. Fraga i in., Canad. J. Phys., 51, 2063 (1973) oraz C. Roetti, E. Clementi, J. Chem. Phys. 61, 2062 (1974)

oraz −20,6686 (−562,4); pierwsze wartości są w j.at., a w nawiasach — w eV. Z drugiej strony energie orbitalne ε_{2s} oraz ε_{2p} ulegają znacznie słabszemu obniżeniu, np. ε_{2s}(C) = −0,7056 (−19,2), ε_{2p}(C) = = −0,4333 (−11,8), ε_{2s}(N) = −0,9453 (−25,7), ε_{2p} (N) = −0,5675 (−15,4), ε_{2s} (O) = −1,2443 (−33,9), ε_{2p}(O) = −0,6319 (−17,2) i są porównywalne z energią orbitalną ε_{1s} dla wodoru. Mówimy, że elektrony znajdujące się na orbitalach 1*s* (skrótowo: elektrony 1*s*) w atomach węgla, azotu czy tlenu są energetycznie znacznie głębiej położone aniżeli elektrony 2*s* czy 2*p* tych atomów. Okaże się to szczególnie ważne przy omawianiu budowy cząsteczek.

Orbitale HF dają najlepszy obraz atomu w przybliżeniu jednoelektronowym. Na poprzednich stronach porównywaliśmy wyniki obliczeń dla atomu helu. Użyte orbitale były orbitalami Slatera. Według reguł Slatera dla orbitalu φ_{1s} atomu wieloelektronowego należy przyjąć $Z' = 1,70$. Obliczona energia atomu helu przy użyciu takiej funkcji Slatera jest rzędu 98% energii dokładnej (rys. 6). Używając orbitali HF (rys. 9) obliczona energia wynosi $\mathcal{E}^{HF}$(He) = −2,8617 j.at., co stanowi 98,55% energii dokładnej. Jest to najlepszy wynik uzyskany w przybliżeniu jednoelektronowym — lepszego już nie można uzyskać w tym modelu, gdyż równania HF wynikają z zasady wariacyjnej. Dla atomu litu energia HF atomu wynosi $\mathcal{E}^{HF}$(Li) = −7,4327 j.at. (stanowi to 99% energii dokładnej litu −7,47807 j.at.), a dla atomu berylu $\mathcal{E}^{HF}$(Be) = −14,5730 j.at. zgadza się w 99,3% z energią dokładną. Jak wskazują te liczby, obliczenia Hartree'ego–Focka dla atomów dają dobre wyniki dla energii całkowitych atomów. Inne wielkości obliczone tą metodą też zgadzają się z danymi doświadczalnymi. Z rys. 11 wynika, że chociaż w niektórych wypadkach obliczone potencjały jonizacyjne atomów są niższe od odpowiednich wartości doświadczalnych, to jednak teoretyczne wyniki doskonale przewidują zmiany wartości potencjałów przy przejściu od atomu do atomu.

Korelacja elektronów

Metoda Hartree'ego–Focka ma charakter przybliżony. Podstawą tej metody jest model elektronów (ogólnie cząstek) niezależnych, w którym każdy elektron porusza się niezależnie od pozostałych. Innymi słowy prawdopodobieństwo znalezienia elektronu w określonym obszarze przestrzeni nie zależy od tego, gdzie przebywa drugi elektron. Intuicyjnie wiemy jednak, że z powodu odpychania się kulombowskiego elektronów przebywanie np. dwóch elektronów w atomie helu po przeciwnych stronach jądra powinno być uprzywilejowane. Ruchy elektronów nie są niezależne od siebie, są one skorelowane. Obliczając energie metodą Hartree'ego–Focka popełniamy błąd, gdyż nie uwzględniamy korelacji elektronów. Ten błąd nosi nazwę energii korelacji (E^{kor}) i jest różnicą

energia korelacji

$$E^{kor} = E^{dokł} - \mathcal{E}^{HF},$$

gdzie $E^{dokł}$ oznacza dokładną wartość energii dla danego układu otrzymaną z rozwiązania równania (1) (lub wartość doświadczalną energii atomu po odjęciu poprawek relatywistycznych itp.).

Energia korelacji jest mała. Dla atomu helu wynosi ona −0,042 j.at. (ok. 1,5% energii dokładnej), a dla litu −0,045 j.at. (1% energii dokładnej). W atomie berylu energia korelacji równa się −0,094 j.at., ale stanowi ona tylko 0,7% energii całkowitej. Energia korelacji dla innych atomów jest również mała i stanowi mniej niż 1% energii dokładnej. Bezwzględna wartość energii korelacji w atomach oczywiście rośnie wraz ze wzrostem liczby elektronów w atomie. Wzrost ten nie jest jednak równomierny. Na przykład energia korelacji helu i litu oraz neonu i sodu jest praktycznie taka sama, a energia korelacji berylu jest około dwukrotnie wyższa od energii korelacji litu. Porównanie z konfiguracjami tych atomów (rys. 8) wskazuje, że najważniejszą rolę odgrywa korelacja par elektronów $1s^2$ lub $2s^2$, natomiast korelacja między parą $1s^2$ a elektronem 2*s* lub między parą $1s^2$ i parą $2s^2$ (np. w berylu) nie ma istotnego znaczenia. Gdy przechodzimy od berylu o konfiguracji $1s^22s^2$ do innych atomów zapełniając orbitale 2*p*, energia korelacji wzrasta. Wzrost ten związany jest ze zbliżonymi energiami orbitali 2*s* i 2*p* oraz z efektami degeneracji orbitali 2*p*. Wyjaśnienie tych spraw wymagałoby jednak szerszej dyskusji.

energia korelacji a konfiguracja

Zwróćmy jeszcze uwagę na fakt, że energie korelacji dla jonu Li^+ oraz atomu Li są praktycznie takie same, a stanie się jasne, dlaczego obliczony potencjał jonizacyjny litu za pomocą metody Hartree'ego–Focka (rys. 11) jest w bardzo dobrej zgodności z doświadczeniem. Zgodnie z definicją potencjału powinniśmy odejmować od energii dokładnej Li^+ energię dokładną Li. Ale $E^{dokł}(Li^+) = \mathcal{E}^{HF}(Li^+) + E^{kor}(Li^+)$ oraz $E^{dokł}(Li) = \mathcal{E}^{HF}(Li) + E^{kor}(Li)$, a ponieważ energie korelacji Li^+ i Li są takie same, więc $E^{dokł}(Li^+) - E^{dokł}(Li) = \mathcal{E}^{HF}(Li^+) - \mathcal{E}^{HF}(Li)$. Tak więc, odejmując dwie przybliżone liczby (energie HF) otrzymaliśmy potencjał jonizacyjny w bardzo dobrej zgodności z doświadczeniem, ponieważ przy odejmowaniu skasowały się błędy (energie korelacji). Podobna sytuacja występuje dla sodu (liczba porządkowa 11) i jego jonu oraz dla potasu (liczba porządkowa 19)

i jego jonu, natomiast dla innych atomów już nie, ponieważ energia korelacji atomu nie jest na ogół równa energii korelacji jego jonu.

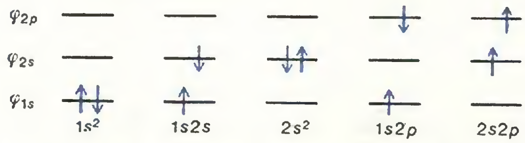
W jaki sposób można uzyskać lepsze wyniki aniżeli uzyskiwane metodą Hartree'ego–Focka? Należy wyjść poza przybliżenie jednoelektronowe i uwzględnić korelacje elektronów. Można to zrobić kilkoma sposobami — omówimy tutaj tylko jeden, tzw. metodę oddziaływania konfiguracji (metoda CI — *Configuration Interaction*).

metoda CI

Metoda oddziaływania konfiguracji

Omawialiśmy poprzednio konfiguracje podstawowe układu elektronów. Ale można oczywiście dla każdego układu elektronów przyporządkować orbitale, które nie są obsadzone przez elektrony konfiguracji podstawowej; pozwala to zbudować tzw. konfiguracje wzbudzone układu (rys. 12). Każda z konfiguracji opisana jest funkcją mającą postać wyznacznika (lub

konfiguracje wzbudzone



Rys. 12. Konfiguracja podstawowa i kilka konfiguracji wzbudzonych atomu helu. Pierwsze cztery konfiguracje są singletami, ostatnia konfiguracja jest trypletem

sumy wyznaczników) postaci (11). Jeżeli znamy postać orbitali dla konfiguracji podstawowej (np. mogą to być orbitale wyznaczone metodą HF) oraz dla konfiguracji wzbudzonych, wówczas możemy szukać funkcji próbnej opisującej układ elektronów w postaci (por. wzór 8):

$$\Phi^{\mathrm{pr}} = \sum_{p=1}^{m} A_p \Phi_p,$$

gdzie Φ_p jest funkcją opisującą różne konfiguracje. Okazuje się, że dla $m = \infty$ za pomocą powyższego rozwinięcia można wyrazić funkcję dokładną danego układu. Oczywiście w praktyce bierze się skończoną sumę (skończoną liczbę konfiguracji), a współczynniki A_p oblicza się metodą Ritza. Uzyskana funkcja falowa Φ^{pr} jest tym lepsza od funkcji Hartree'ego–Focka, im więcej uwzględnia się konfiguracji wzbudzonych. W tabeli podano przykładowo wyniki dla helu i berylu uzyskane metodą HF oraz metodą CI.

Energie atomów helu i berylu obliczone metodą Hartree'ego-Focka i metodą oddziaływania konfiguracji

Liczba konfiguracji	Energia, j.at.	Błąd, j.at.
HEL		
1 (HF)	−2,8617	−0,0421
20 (CI)	−2,9028	−0,0009
35 (CI)	−2,9032	−0,0005
wartość dokładna	−2,9037	—
BERYL		
1 (HF)	−14,573	−0,094
2 (CI)	−14,617	−0,050
55 (CI)	−14,661	−0,006
wartość dokładna	−14,667	—

porównanie wyników metod HF i CI

Metoda oddziaływania konfiguracji wyraźnie podkreśla przybliżony charakter pojęcia konfiguracji elektronowej atomu. Na przykład dla atomu berylu przy użyciu funkcji falowej będącej kombinacją liniową dwóch konfiguracji $1s^2 2s^2$ oraz $1s^2 2p^2$ otrzymuje się energię stanu podstawowego atomu znacznie lepszą niż metodą HF. Z drugiej strony, ze współczynników rozwinięcia wynika, że konfiguracja $1s^2 2s^2$ występuje ze współczynnikiem $A_1 = 0{,}95$, natomiast konfiguracja $1s^2 2p^2$ ze współczynnikiem $A_2 = 0{,}31$. Współczynniki te oczywiście zmieniają się w przypadku użycia większej liczby konfiguracji. Jak więc widać, model jednoelektronowy (tzn. przypisanie elektronów określonym orbitalom) jest przybliżonym, ale bardzo użytecznym, opisem rzeczywistej sytuacji.

Podstawy spektroskopii molekularnej

W równaniu Schrödingera (1) opisującym dowolny układ chemiczny, np. cząsteczkę, elektrony i jądra układu są traktowane w ten sam sposób. Jednakże masa najlżejszego jądra (protonu) jest przeszło 1800 razy większa niż masa elektronu. Dlatego też jądra w cząsteczce poruszają się znacznie wolniej niż elektrony; fakt ten jest przyczyną wielu istotnych własności cząsteczek.

Rozdzielenie ruchu elektronów i jąder

Różnice pomiędzy ruchem jąder i ruchem elektronów cząsteczki powodują, że jej stany energetyczne można opisać w pewien przybliżony sposób. Można pokazać mianowicie, że równanie Schrödingera (1) opisujące dowolną cząsteczkę można zastąpić układem dwóch równań

$$\hat{H}^{\mathrm{el}} \Psi^{\mathrm{el}} = E^{\mathrm{el}} \Psi^{\mathrm{el}} \qquad (16a)$$

$$(\hat{H}^{\mathrm{j}} + E^{\mathrm{el}}) \Psi^{\mathrm{j}} = E \Psi^{\mathrm{j}}. \qquad (16a')$$

Pierwsze równanie, zwane elektronowym równaniem Schrödingera, opisuje ruch elektronów cząsteczki w polu nieruchomych jąder. Hamiltonian elektronowy $\hat{H}^{\mathrm{el}}$ składa się z energii kinetycznej wszystkich elektronów, energii oddziaływania elektronów z nieruchomymi jądrami, energii oddziaływania elektrostatycznego pomiędzy elektronami oraz z wyrazu opisującego elektrostatyczne oddziaływanie pomiędzy jądrami (stały wyraz w przypadku nieruchomych jąder). Równanie elektronowe rozwiązuje się dla różnych ustalonych położeń jąder cząsteczki. Funkcje elektronowe Ψ^{el} oraz energie E^{el} zależą więc parametrycznie od położeń jąder.

elektronowe równanie Schrödingera

Drugie równanie opisuje ruch jąder, na które działa potencjał E^{el} pochodzący od elektronów. Hamiltonian $\hat{H}^{\mathrm{j}}$ jest operatorem energii kinetycznej wszystkich jąder cząsteczki.

Z postaci równań (16) i (16a') wynika, że nie są to równania niezależne od siebie. Wprawdzie pierwsze równanie można rozwiązać nie znając rozwiązania drugiego równania, ale nie odwrotnie. Rozwiązania równań (16) związane są z rozwiązaniem równania (1) w prosty sposób:

$$\Psi(\vec{r}, \vec{R}) = \Psi^{\mathrm{el}}(\vec{r}; \vec{R})\, \Psi^{\mathrm{j}}(\vec{R}).$$

$$E = E^{\mathrm{el}}(\vec{R}) + E^{\mathrm{j}}. \qquad (17)$$

Przez podanie $\vec{r}$ i $\vec{R}$ jako argumentów funkcji falowych zaznaczono tutaj bezpośrednią ich zależność od zbioru wektorów opisujących położenia wszystkich elektronów $(\vec{r})$ i wszystkich jąder $(\vec{R})$; występowanie $\vec{R}$, jako argumentu w E^{el} oznacza, że energia elektronowa cząsteczki zależy parametrycznie od położeń jąder. Przyjęcie funkcji falowej i energii w postaci (17) nosi nazwę przybliżenia Borna–Oppenheimera (przybliżenie BO). Jest to najczęściej stosowane przybliżenie przy opisie własności cząsteczek. Niekiedy (ale bardzo rzadko) stosuje się przybliżenie adiabatyczne — do potencjału E^{el} $(\vec{R})$ w (17) i w równaniu (16a') dodaje się poprawkę uwzględniającą drobny wpływ ruchu jąder na ruch elektronów. Poprawkę tę można obliczyć, jeśli znamy funkcję falową opi-

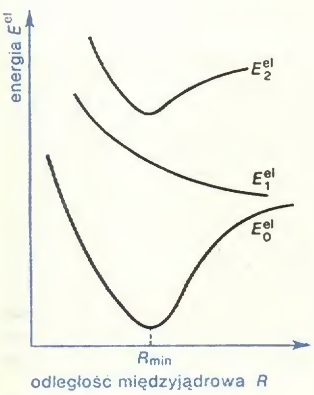
przybliżenie Borna-Oppenheimera (BO)

przybliżenie adiabatyczne

sującą układ elektronów cząsteczki w przybliżeniu BO. Poprawkę adiabatyczną oblicza się jednak tylko w nielicznych przypadkach, gdyż jest ona bardzo mała. Na przykład dla cząsteczki wodoru poprawka adiabatyczna do energii dysocjacji cząsteczki jest rzędu 0,01% tej energii. Jeszcze mniejsze znaczenie mają efekty nieadiabatyczne. Należałoby je uwzględnić np. w obliczeniach, w których jest potrzebna wysoka dokładność wyników.

cząsteczka dwuatomowa

Zajmijmy się teraz równaniem elektronowym (16a). Dla uproszczenia za podstawę rozważań weźmiemy cząsteczkę dwuatomową, choć uogólnienie rozważań na cząsteczki wieloatomowe nie sprawia żadnego kłopotu. Jak wspomnieliśmy już poprzednio, aby opisać ruch elektronów w cząsteczce należy rozwiązać równanie elektronowe dla ustalonych położeń jąder cząsteczki. Niestety, jak wiadomo, nie można uzyskać ścisłego rozwiązania w postaci analitycznej dla układu większego niż jednoelektronowy. Rozwiązanie równania elektronowego znajdujemy więc stosując metody przybliżone chemii kwantowej (przeważnie metodę wariacyjną). Równanie to ma wiele rozwiązań — jest nim zbiór funkcji falowych $\Psi_n^{el}(\vec{r}; \vec{R})$ oraz zbiór odpowiadających im energii $E_n^{el}(\vec{R})$. Indeks n numeruje poszczególne stany cząsteczek (w ogólności n jest tylko indeksem i nie ma nic wspólnego z główną liczbą kwantową dla atomów jednoelektronowych). Umówimy się, że stan podstawowy oznaczać będziemy indeksem $n = 0$, a stany wzbudzone numerować będziemy kolejnymi liczbami



Rys. 13. Schematyczne przedstawienie zależności energii E_n^{el} od odległości międzyjądrowej R dla kilku stanów elektronowych cząsteczki dwuatomowej (krzywe energii potencjalnej). E_0^{el} — energia stanu podstawowego cząsteczki (cząsteczka stabilna), R_{min} odległość międzyjądrowa, przy której występuje minimum energii elektronowej stanu podstawowego; E_1^{el}, E_2^{el} — energie elektronowe odpowiednio niestabilnego i stabilnego stanu wzbudzonego cząsteczki

całkowitymi. Typowe wykresy $E_n^{el}(\vec{R})$ dla stanu podstawowego oraz dla kilku stanów wzbudzonych cząsteczki dwuatomowej są przedstawione na rys. 13. Wykresy te nazywane są krzywymi energii potencjalnej. Należy pamiętać, że mamy nieskończoną liczbę stanów elektronowych cząsteczki, ale w praktyce ograniczamy się tylko do rozważania stanu podstawowego i kilku najniżej leżących stanów wzbudzonych. Kształt krzywej energii potencjalnej ukazującej zmianę energii E_n^{el} jako funkcji odległości międzyjądrowej jest jedną z najbardziej charakterystycznych właściwości struktury cząsteczki. Stabilność cząsteczki jest uwarunkowana faktem występowania względnego minimum funkcji E_n^{el}. Każdej wartości E_n^{el} dla określonej odległości międzyjądrowej odpowiada funkcja falowa Ψ_n^{el}, która jest przedstawiona w postaci numerycznej. Uogólnienie pojęcia krzywych energii potencjalnej podanych na rys. 13 w stosunku do cząsteczki wieloatomowej jest bardzo proste. Dla cząsteczki dwuatomowej funkcje E_n^{el} są zależne

krzywe energii potencjalnej

od jednego parametru — odległości międzyjądrowej R. Dla cząsteczki wieloatomowej należy uwzględnić wszystkie parametry wyznaczające względne położenia jąder cząsteczki (zbiór tych parametrów oznaczamy nadal symbolem $\vec{R}$). Energie $E_n^{el}(\vec{R})$ cząsteczki wieloatomowej są funkcjami wieloparametrowymi — są to powierzchnie energii potencjalnej w przestrzeni wielowymiarowej. Przedstawienie graficzne tych krzywych dla cząsteczek wieloatomowych jest niemożliwe. Nie przeszkadza to zupełnie w dyskusji stanów elektronowych cząsteczek złożonych posługiwać się krzywymi przedstawionymi na rys. 13.

cząsteczka wieloatomowa

Oscylacje i rotacje cząsteczek

W przypadku cząsteczek dwuatomowych — w sposób ścisły, a dla cząsteczek wieloatomowych — z bardzo dobrym przybliżeniem, równanie (16a′) opisujące ruchy jąder cząsteczki rozbija się na dwa równania postaci

$$\hat{H}^{rot}(\vartheta, \varphi)\Psi_J^{rot}(\vartheta, \varphi) = E_J^{rot}\Psi_J^{rot}(\vartheta, \varphi), \qquad (16b)$$

$$(\hat{H}^{osc}(R) + E_n^{el}(R) + E_J^{rot})\Psi_{nvJ}^{osc}(R) = E_{nvJ}\Psi_{nvJ}^{osc}(R), \qquad (16c)$$

gdzie $\hat{H}^j = \hat{H}^{rot} + \hat{H}^{osc}$, $\Psi^j = \Psi^{rot}\Psi^{osc}$

oraz $E^j = E^{rot} + E^{osc}$.

W równaniach powyższych kąty ϑ oraz φ charakteryzują położenie cząsteczki podczas obrotu. Pierwsze z powyższych równań (zapisane tutaj w postaci symbolicznej) opisuje ruch obrotowy (rotację) cząsteczki dwuatomowej i jest bardzo dobrze znane matematykom. Rozwiązaniem tego równania są funkcje kątowe takiej samej postaci, jak dla atomu wodoru, por. wzór (3). Ze względu na tradycję, w przypadku ruchu rotacyjnego zamiast używania indeksu l jak w (3), wprowadza się liczbę kwantową J, określającą stany rotacyjne cząsteczki. Oczywiście, tak jak liczba l, liczba rotacyjna J może przyjmować wartości $J = 0, 1, 2, \ldots$. Energie poszczególnych stanów rotacyjnych Ψ_J^{rot} wynoszą $E_J^{rot} = BJ(J+1)$, gdzie B jest tzw. stałą rotacyjną.

liczba rotacyjna J

Równanie (16c) opisuje ruch związany ze zmianą współrzędnej R — opisuje więc ono oscylacje jąder cząsteczki zachodzące w polu sił o potencjale równym sumie potencjału elektronowego E_n^{el} dla różnych stanów oraz członu rotacyjnego zależącego od rotacyjnej liczby kwantowej J. Dla danych n oraz J równanie (16c) ma wiele rozwiązań, które rozróżniamy wskaźnikiem v. Dlatego też w (16c) funkcje falowe Ψ^{osc} oraz energie E zostały oznaczone trzema wskaźnikami n, v, J oznaczającymi odpowiednio stan elektronowy, oscylacyjny i rotacyjny.

stan elektronowy, oscylacyjny i rotacyjny

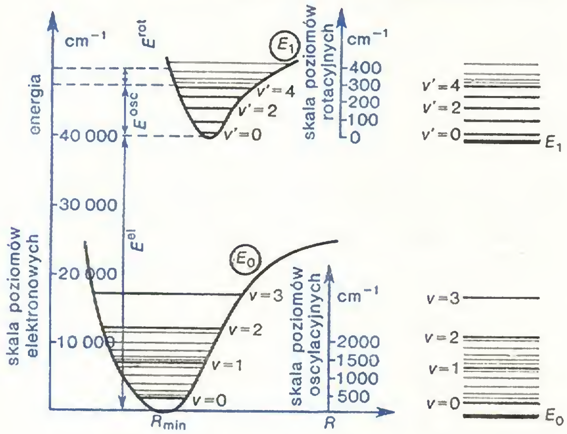
Równania (16a), (16b) oraz (16c) mają zasadnicze znaczenie dla spektroskopii molekularnej. Rozwiązując układ tych równań znajdujemy układ poziomów energetycznych cząsteczki i odpowiadające im funkcje falowe; możemy więc znaleźć różnice pomiędzy energiami poziomów energetycznych (różnice te są mierzone w doświadczeniu), jak również potrafimy obliczyć prawdopodobieństwa przejść, czy też określić reguły wyboru dla określonych typów przejść.

Z równań (16a)–(16c) wynika, że w przybliżeniu funkcja falowa cząsteczki $\Psi = \Psi^{el}\Psi^{osc}\Psi^{rot}$, natomiast jej energia wynosi $E = E^{el} + E^{osc} + E^{rot}$. Określając stan energetyczny cząsteczki musimy więc oprócz podania stanu elektronowego określić jej stan oscylacyjny oraz stan rotacyjny. Na rys. 14 przedstawiono schemat diagramu poziomów energetycznych cząsteczki dwuatomowej (rysunek ten można traktować oczywiście jako uogólniony diagram poziomów energetycznych dla cząsteczki wieloatomowej).

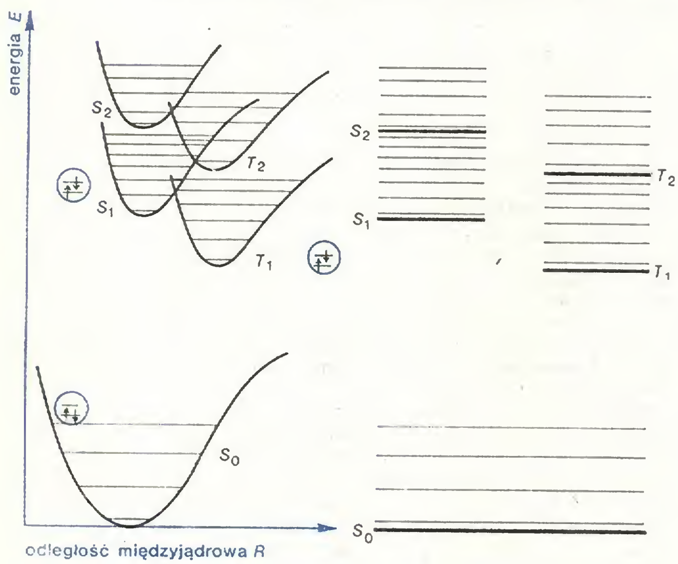
Spektroskopia molekularna zajmuje się badaniem oddziaływań pomiędzy materią (cząsteczki) a promieniowaniem elektromagnetycznym. Używając pomiarowych technik absorpcyjnych lub emisyjnych wy-

różnice energii między stanami

znaczamy różnice energii pomiędzy różnymi stanami cząsteczki. Jak widać z rys. 14 różnice energii pomiędzy stanami elektronowymi są dużo większe aniżeli różnice pomiędzy stanami oscylacyjnymi określonego stanu elektronowego, a te z kolei są większe niż różnice pomiędzy poziomami rotacyjnymi. Typowe różnice wartości energii wynoszą: dla przejść rotacyjnych od 0,001 do 0,01 eV, dla przejść oscylacyjnych od 0,05 do 0,2 eV, dla przejść elektronowych od 2 do



Rys. 14. Poziomy energetyczne cząsteczki dwuatomowej. Krzywe energii potencjalnej wykreślone są dla elektronowego stanu podstawowego E_0 oraz dla pierwszego stanu wzbudzonego E_1. Przedstawiono kilka stanów oscylacyjnych oraz kilka stanów rotacyjnych dla poziomów oscylacyjnych $v = 0$, $v = 1$ oraz dla $v' = 4$. Dla trzech typów przejść użyto różnych skal energetycznych. Na prawo pokazano ten sam układ poziomów energetycznych, ale bez krzywych energii potencjalnej — poziomy E_0 oraz E_1 odpowiadają położeniu minimum odpowiednich krzywych



Rys. 15. Dwa różne sposoby przedstawiania stanów energetycznych złożonej cząsteczki. W przybliżeniu nierelatywistycznym każdy stan ma określoną multipletowość zależną od wzajemnego ukierunkowania spinów elektronów cząsteczki (zob. rys. 12). Dla cząsteczek o parzystej liczbie elektronów elektronowy stan podstawowy jest stanem singletowym (S_0). Na schemacie zaznaczono po dwa wzbudzone stany singletowe (S_1, S_2) i trypletowe (T_1, T_2). Symboli S_n i T_n używa się zamiast funkcji falowych Ψ^{el} określających poszczególne stany elektronowe cząsteczki. Dla trzech stanów S_0, S_1 i T_1 zaznaczono w kółkach kierunki spinów dwóch elektronów położonych najwyżej energetycznie w konfiguracji elektronowej opisującej dany stan cząsteczki — linie poziome oznaczają orbitale molekularne. Połowa pozostałych elektronów w każdym ze stanów ma spiny skierowane przeciwnie do spinów drugiej połowy elektronów. Z obliczeń kwantowych wynika, że dwa stany opisane konfiguracjami różniącymi się tylko kierunkiem spinów (np. stany S_1 i T_1) mają różną energię. W wypadku stanów singletowych i trypletowych zawsze ten ostatni stan ma niższą energię aniżeli odpowiadający mu stan singletowy. Każdemu stanowi elektronowemu odpowiada układ stanów oscylacyjnych

podział spektroskopii

5 eV. Jak widać, różne typy przejść obserwuje się w różnych częściach widma elektromagnetycznego. Rozróżniamy dlatego m.in. spektroskopię mikrofalową, podczerwieni oraz światła widzialnego i nadfioletu. Podział ten oczywiście nie wyczerpuje wszystkich możliwości technicznych. Istnieje dziedzina zwana spektroskopią ramanowską (badanie widm oscylacyjnych w obszarze widzialnym i nadfioletu). Ściśle ze spektroskopią molekularną wiążą się spektroskopia magnetycznego rezonansu jądrowego oraz spektroskopia elektronowego rezonansu spinowego (→ Spektroskopia rezonansów magnetycznych).

Dla dużych cząsteczek, kiedy zahamowana jest ich rotacja (szczególnie gdy cząsteczki znajdują się w roztworze lub w ośrodku sztywnym) na schemacie poziomów energetycznych podawane są tylko stany elektronowe i stany oscylacyjne (rys. 15). Aby określić stan takiej złożonej cząsteczki, należy oprócz podania funkcji falowej określającej stan elektronowy (rozwiązanie równania 16a) podać funkcję falową Ψ^{osc} opisującą określony stan oscylacyjny cząsteczki. Innymi słowy, w przypadku pominięcia rotacji, należy znać rozwiązanie równań (16a) i (16c). Aby podać rozwiązanie tego ostatniego równania, należy przede wszystkim znać energię $E_n^{el}(R)$. Dla cząsteczki dwuatomowej należy podać przebieg krzywej energii potencjalnej (rys. 13). Postać tej funkcji, jak już wspomnieliśmy poprzednio, jest nieznana. Stosuje się jednak wiele uproszczeń opisujących przebieg takich krzywych.

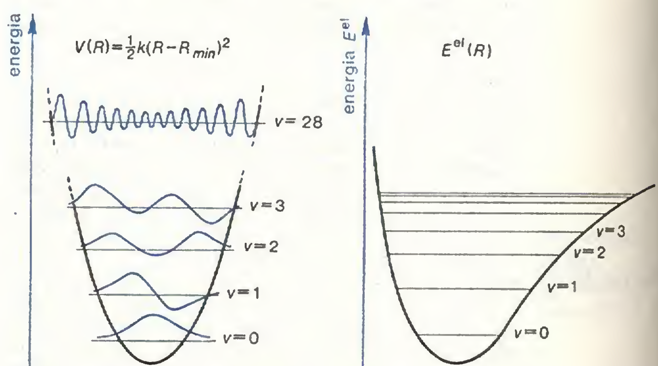
Oscylator harmoniczny

cząsteczka dwuatomowa

W przypadku cząsteczki dwuatomowej najprostszym przybliżeniem funkcji energii $E_n^{el}(R)$ jest parabola postaci $V(R) = {}^1/_2 k(R-R_{min})^2$, gdzie k jest tzw. stałą siłową charakteryzującą siły powstające w cząsteczce przy przesunięciu jąder z ich położenia równowagi R_{min}. Założenie to jest równoważne przyjęciu, że jądra drgają ruchem harmonicznym (jest to przypadek tzw. oscylatora harmonicznego). Rozwiązanie równania (16c) dla tego przypadku jest bardzo dobrze znane. Funkcje własne Ψ_v^{osc}, charakteryzujące poszczególne stany oscylacyjne cząsteczki, mają postać analityczną (nie będziemy podawali tutaj tych funkcji, ale przebieg ich dla kilku stanów oscylacyjnych pokazany jest na rys. 16), a wartości własne E_v^{osc} wynoszą

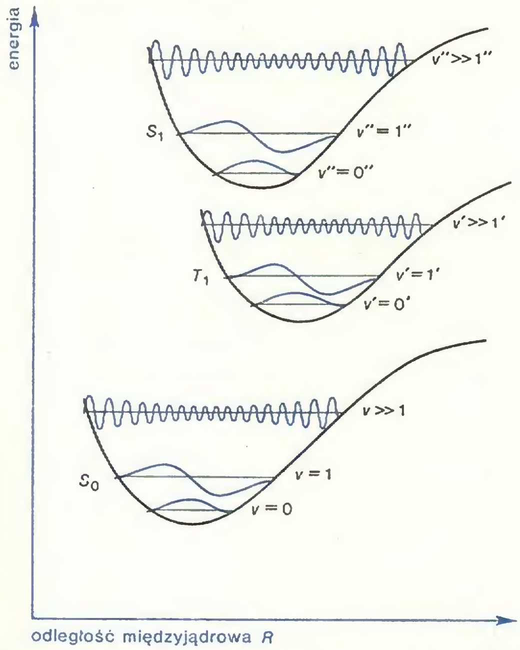
$$E_v^{osc} = h\nu(v+{}^1/_2), \qquad (18)$$

gdzie ν jest częstością oscylacji, natomiast v jest oscylacyjną liczbą kwantową ($v = 0, 1, 2, \ldots$). Model



Rys. 16. Porównanie energii cząsteczki dwuatomowej i oscylatora harmonicznego: a) poziomy energetyczne i przebieg odpowiednich funkcji falowych oscylatora harmonicznego, b) krzywa energii potencjalnej cząsteczki dwuatomowej

oscylatora harmonicznego opisuje w dobrym przybliżeniu tylko najniższe stany oscylacyjne cząsteczki. Związane jest to z faktem, że parabola $V(R)$ jest dobrym przybliżeniem dla krzywej $E^{el}(R)$ tylko w pobliżu jej minimum. W rzeczywistości drgania oscylacyjne w cząsteczce są anharmoniczne. Anharmoniczność ta powoduje, że odstępy pomiędzy poziomami oscylacyjnymi cząsteczki dwuatomowej maleją ze wzrostem pobudzenia, podczas gdy w oscylatorze harmonicznym są one stałe (rys. 16).



Rys. 17. Schematyczne przedstawienie trzech stanów energetycznych złożonej cząsteczki (zob. rys. 15) z zaznaczonymi kształtami funkcji oscylacyjnych (funkcje oscylacyjne zaznaczono tylko dla dwóch najniższych stanów oscylacyjnych oraz dla stanu oscylacyjnego o dużej wartości liczby kwantowej v). Natężenie różnych przejść między stanem podstawowym S_0 i stanem wzbudzonym (np. S_1) jest m.in. proporcjonalne do kwadratu całki nakrywania odpowiednich stanów oscylacyjnych. Na przykład, dla sytuacji przedstawionej na rysunku przejście

$v = 0 \rightarrow v'' = 1''$ jest silniejsze niż przejście

$v = 0 \rightarrow v'' = 0''$, ponieważ

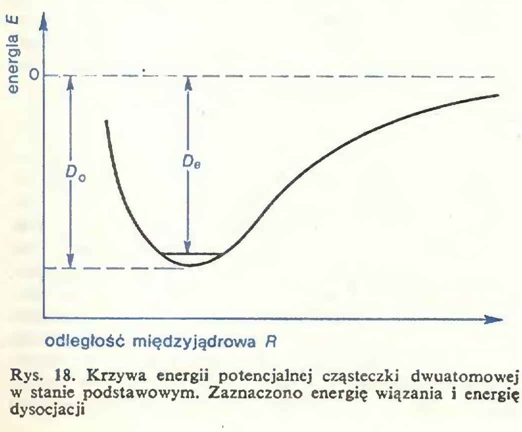
$$\int \Psi_0^{osc} \ \Psi_{1''}^{osc} \, dv > \int \Psi_0^{osc} \ \Psi_{0''}^{osc} \, dv$$

Ważną konsekwencją wzoru (18) jest to, że dla $v = 0$ energia oscylacyjna $E_0^{osc} = {}^1/_2\hbar v$ i cząsteczka nie może nigdy mieć energii całkowitej odpowiadającej minimalnej wartości $E^{el}(R)$ dla $R = R_{min}$.

cząsteczka wieloatomowa

Rozważania dotyczące cząsteczki dwuatomowej można rozszerzyć na cząsteczki wieloatomowe (rys. 17). Cząsteczka złożona z N jąder posiada $3N-6$ drgań podstawowych (lub $3N-5$ drgań w przypadku cząsteczki liniowej) i w przybliżeniu każde z tych drgań można opisać równaniami oscylatora harmonicznego. Tak więc całkowita energia oscylacyjna cząsteczki jest sumą energii poszczególnych oscylatorów harmonicznych, tzn. $E^{osc} = \sum_{i=1}^{3N-6} \hbar v_i(v_i + {}^1/_2)$.

Model oscylatora harmonicznego można użyć do przybliżonego opisu oscylacji stanu podstawowego cząsteczki, jak również do opisu oscylacji jej stanów wzbudzonych. Pozwala to wyjaśnić wiele jakościowych właściwości widm elektronowych złożonych układów (np. różnice w natężeniu pasm oscylacyjnych widma elektronowego).



Rys. 18. Krzywa energii potencjalnej cząsteczki dwuatomowej w stanie podstawowym. Zaznaczono energię wiązania i energię dysocjacji

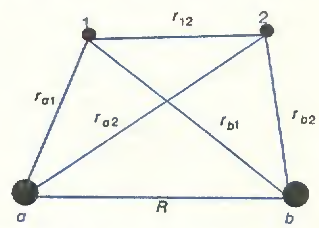
Cząsteczka H_2

Zanim zaczniemy omawiać zastosowania metod chemii kwantowej do opisu właściwości cząsteczek, poświęcimy kilka słów cząsteczce wodoru. Cząsteczka ta zajmuje szczególne miejsce w mechanice kwantowej atomów i cząsteczek. Jest ona najprostszą cząsteczką zawierającą dwa elektrony, nic więc dziwnego, że mimo iż nie można było dla niej znaleźć ścisłego rozwiązania równania Schrödingera, od początku powstania chemii kwantowej usiłowano obliczyć w sposób przybliżony jej energię wiązania (rys. 18). Zagadnieniu temu poświęcono wiele prac, uzyskując teoretyczne wielkości, które mniej lub bardziej zgadzały się z doświadczalną wartością energii wiązania wynoszącą 4,75 eV. Już w 1933 r. H.M. James i A.S. Coolidge wykonali obliczenia dla cząsteczki wodoru używając funkcji falowej typu (8), gdzie funkcje χ_p zawierały wielomiany postaci (zob. oznaczenia na rys. 19):

energia wiązania

$$e^{-\alpha(\xi_1+\xi_2)} \xi_1^s \eta_1^t \xi_2^u \eta_2^v r_{12} .$$

Jest to szereg potęgowy pomnożony przez czynnik wykładniczy zależny od nieznanego parametru α.



Rys. 19. Oznaczenia położeń elektronów *1* i *2* oraz jąder *a* i *b* w cząsteczce wodoru. Zamiast współrzędnych kartezjańskich r_a lub r_b można użyć współrzędnych eliptycznych

$$\xi = \frac{r_a + r_b}{R} \quad \text{oraz} \quad \eta = \frac{r_a - r_b}{R}$$

Oczywiście w wyrażeniu (8) bierze się pod uwagę skończoną liczbę wyrazów, a współczynniki c_p oraz parametr α wyznacza się wariacyjnie. Im więcej wyrazów będzie występowało w rozwinięciu funkcji (8), tym uzyskane wyniki będą dokładniejsze.

wynik Jamesa i Coolidge'a

James i Coolidge stosując 13-członowe rozwinięcie funkcji falowej (8) uzyskali dla energii wiązania wartość 4,72 eV, zgadzającą się dobrze z wynikiem doświadczalnym.

Interesowano się jednak wykonaniem bardziej dokładnych obliczeń, które umożliwiłyby zbadanie stosowalności praw mechaniki kwantowej dla układów cząsteczkowych (m.in. można byłoby sprawdzić stopień dokładności szeregu przybliżeń stosowanych w chemii kwantowej).

wynik Kołosa i Roothaana

Obliczenia tego typu stały się dopiero możliwe po pojawieniu się w pracowniach naukowych komputerów. W 1959 r. W. Kołos i C.C.J. Roothaan, rozszerzając obliczenia Jamesa i Coolidge'a i stosując 40-członowe rozwinięcie, uzyskali energię wiązania $D_e = 38\,286{,}9$ cm^{-1}, a w kilka miesięcy później

pomiar Herzberga i Monfilsa

Herzberg i Monfils zmierzyli dokładniej wartość tej energii uzyskując $D_e = 38\,287{,}0 \pm 0{,}8\ cm^{-1}$. Pomiary energii wiązania, jak również jej obliczona wartość były tak dokładne, że podanie wartości D_e w „dużych jednostkach" jakimi są jednostki atomowe energii (1 j.at. energii $\approx 2{,}19 \cdot 10^5\ cm^{-1}$) czy elektronowolty (1 eV $\approx 8 \cdot 10^3\ cm^{-1}$) nie jest celowe.

Jak widać, wartość teoretyczna i doświadczalna prawie idealnie zgadzały się ze sobą. Była to jednak zgodność pozorna. Należało oczekiwać, że energię wiązania można obliczyć jeszcze dokładniej przy użyciu lepszej funkcji próbnej (zawierającej więcej parametrów wariacyjnych), aniżeli ta, którą użyto w 1959 r. Nieznane były różnego rodzaju poprawki, np. poprawki relawistyczne wynikające z zależności masy od prędkości, czy też poprawki promieniste związane z oddziaływaniem pomiędzy naładowanymi cząstkami. Nie znano dokładnie poprawek adiabatycznych na ruch jąder, ani poprawek wynikających z efektów nieadiabatycznych. Zastrzeżenie budziło też porównywanie teoretycznej i doświadczalnej energii wiązania, gdyż ta ostatnia wielkość nie jest bezpośrednio mierzalna. Należało raczej obliczyć bardzo dokładnie krzywą energii potencjalnej, a następnie należało obliczyć energię najniższego poziomu oscylacyjnego, a tym samym energię dysocjacji. Dopiero tak obliczoną wartość D_0 można porównać z bezpośrednio mierzoną wartością energii dysocjacji cząsteczki.

obliczenia Kołosa i Wolniewicza

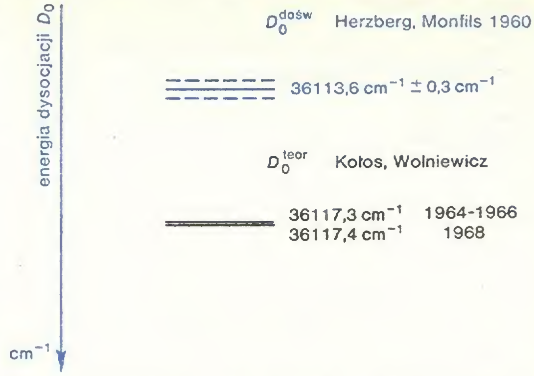
Obliczenia takie przeprowadzili W. Kołos (Uniwersytet Warszawski) oraz L. Wolniewicz (Uniwersytet Mikołaja Kopernika w Toruniu) w czasie pobytu na University of Chicago. Po opracowaniu specjalnego programu obliczeniowego, korzystając z 80-członowego rozwinięcia funkcji falowej (8) uzyskali wyniki, które przedstawiono w tabeli. Jak widać, przybliżenie adiabatyczne daje wartość energii dysocjacji D_0 nie różniącą się wiele od D_0 obliczonej w przybliżeniu BO. Różnica ta wynosi tylko 5,9 cm^{-1}. Ale wynik doświadczalny był znany wówczas z dokładnością

Energia dysocjacji cząsteczki wodoru

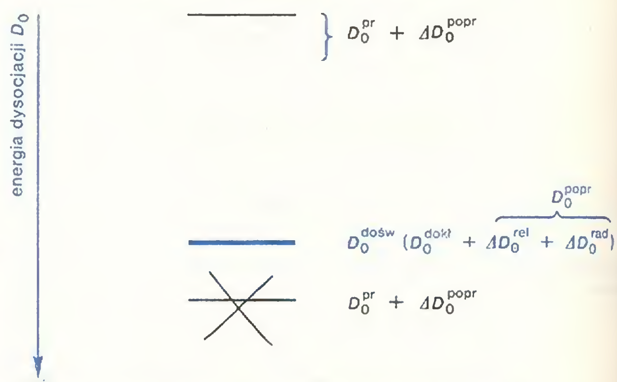
Metoda obliczenia	Otrzymana wartość, cm^{-1}
Przybliżenie Borna–Oppenheimera	$D_0 = 36112{,}2$
Przybliżenie adiabatyczne	$D_0 = 36118{,}1$
Poprawki relatywistyczne	$\Delta D_0^{rel} = -0{,}5$
Poprawki promieniste	$\Delta D_0^{rad} = -0{,}2$
Całkowita wartość teoretyczna	36117,3

do $\pm 0{,}3\ cm^{-1}$, dlatego bardzo ważne było obliczenie D_0 w przybliżeniu adiabatycznym. Poprawki relatywistyczne i promieniste chociaż bardzo małe, musiały być również uwzględnione w tak dokładnych obliczeniach. Zupełnie niedawno okazało się, że nie uwzględnione początkowo przez Kołosa i Wolniewicza efekty nieadiabatyczne są zupełnie do pominięcia.

Ostateczna teoretyczna wartość energii dysocjacji $D_0 = 36\,117{,}3\ cm^{-1}$ zaskoczyła teoretyków (rys. 20). Była ona większa niż wartość doświadczalna! Było to nie tylko dziwne, ale również niepokojące. Zgodnie bowiem z zasadą wariacyjną, jeśli korzystając z szeregu funkcji próbnych Φ^{pr} obliczymy odpowiadające tym funkcjom energie dysocjacji D_0^{pr}, to najlepszym przybliżeniem do doświadczalnej energii dysocjacji będzie ta wartość D_0^{pr}, która jest największa. Ale nigdy obliczona wartość D_0 (po uwzględnieniu poprawek) nie powinna być większa od wartości doświadczalnej (rys. 21). Tymczasem wartość teoretyczna była większa od doświadczalnej o 3,7 cm^{-1}, a więc wynosiła znacznie więcej niż błąd pomiaru doświadczalnego (rys. 20). Co więcej, należało oczekiwać, że gdyby obliczyć D_0 z jeszcze lepszą funkcją próbną lub gdyby uwzględnić efekty nieadiabatyczne (później okazało się, że te ostatnie poprawki są nieistotne), to wówczas energia dysocjacji jeszcze zwiększyłaby się, a tym samym powiększyłaby się rozbieżność pomiędzy doświadczeniem a teorią. Było to bardzo niepokojące,



Rys. 20. Porównanie energii dysocjacji D_0 cząsteczki wodoru wyznaczonej doświadczalnie oraz obliczonej teoretycznie



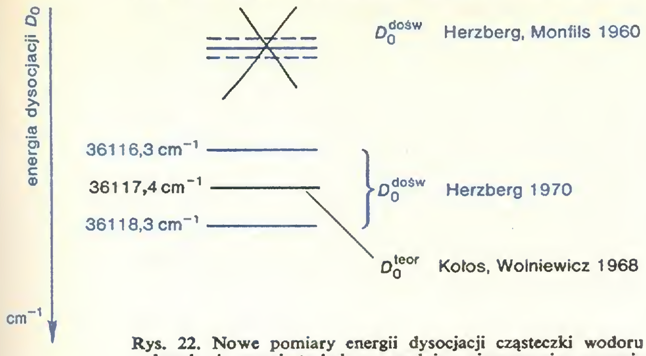
Rys. 21. Wartość energii dysocjacji obliczana metodą wariacyjną przy użyciu przybliżonej (próbnej) funkcji falowej nie może być większa od wartości otrzymanej z dokładnego rozwiązania równania Schrödingera

bo gdyby tak było w rzeczywistości, to metod chemii kwantowej nie można byłoby stosować do układów cząsteczkowych, ściślej mówiąc, zasada wariacyjna nie mogłaby być używana do znajdowania kresu dolnego obliczonej energii cząsteczek. Traciłyby wówczas sens jedne z podstawowych równań, opisujące w przybliżony sposób cząsteczki — równania Hartree'ego–Focka. Nie miałaby sensu metoda oddziaływania konfiguracyjnego.

Istniała jednak pewna szansa wytłumaczenia występującej rozbieżności, np. wyniki obliczeń mogły być błędne lub mogły być niepoprawne wyniki pomiarów lub ich interpretacja. Doświadczalną wartość energii dysocjacji uważano jednak wówczas za bardzo dokładną, natomiast sceptycznie odnoszono się do wyniku obliczonego. Rozbieżność spowodowaną nieuwzględnianiem w obliczeniach nie znanego dotychczas efektu można było śmiało wykluczyć, gdyż obliczenia kwantowe np. dla atomu helu dały wyniki zgodne z doświadczalnymi.

Kołos i Wolniewicz udali się powtórnie do USA i powtórzyli obliczenia energii dysocjacji dla cząsteczki wodoru. Obliczenia te zostały wykonane za pomocą nowego programu obliczeniowego i z zastosowaniem lepszej funkcji falowej (100-członowe rozwinięcie). Obliczenia prowadzono w tzw. podwójnej precyzji eliminującej ewentualne błędy wynikające z zaokrąglenia. Poprzednie wyniki zostały w pełni potwierdzone — stosując lepszą funkcję falową uzyskali oni wynik teoretyczny lepszy od poprzedniego o 0,1 cm^{-1}. Tym razem zaniepokoił się Herzberg o swoje poprzednie pomiary — powtórzył je w innych warunkach i z większą dokładnością. Wynik nowych

pomiarów był nieoczekiwany. Okazało się bowiem, że poprzednie wyniki były fałszywe. Herzberg stwierdził tym razem zachodzenie na siebie linii widmo-



Rys. 22. Nowe pomiary energii dysocjacji cząsteczki wodoru wykazały, że energia ta była poprzednio zmierzona niepoprawnie

nowy pomiar Herzberga

wych i niestety nie udało mu się dokładnie wyznaczyć nowej wartości energii dysocjacji. Mógł on jedynie stwierdzić, że energia dysocjacji jest zawarta pomiędzy 36 118,3 a 36 116,3 cm^{-1} (rys. 22). Ostateczny wynik teoretyczny otrzymany przez Kołosa i Wolniewicza prawie dokładnie odpowiadał wartości pośredniej zakresu danych doświadczalnych.

Wyniki obliczeń wykonane przez Kołosa i Wolniewicza dla cząsteczki H_2 mają ogromne znaczenie dla podstaw chemii kwantowej. Wykonane zostały dla bardzo małej cząsteczki, ale wykazały, że wyniki teoretyczne mogą współzawodniczyć z powodzeniem w dokładności z wynikami doświadczalnymi, a co więcej, że dzięki nim można nawet korygować błędy w wynikach pomiarów. Na ogół, kiedy porównuje się jakieś wartości obliczone z odpowiednimi danymi zmierzonymi, mówi się, że „wyniki teoretyczne zgadzają się z danymi doświadczalnymi", natomiast w przypadku cząsteczki wodoru należało powiedzieć, że „wynik doświadczalny zasadniczo zgadzał się z wartością teoretyczną".

Warto podkreślić, że oprócz energii dysocjacji obliczono dla H_2 wartości wielu innych wielkości mających znaczenie dla spektroskopii (m.in. wielkości charakteryzujące oddziaływania pola elektromagnetycznego z cząsteczką, jak momenty elektryczne czy polaryzowalność). Okazuje się, że obliczenia te mają również praktyczne znaczenie. Na przykład obliczony tzw. moment kwadrupolowy cząsteczki H_2 został niedawno zastosowany przez astronomów do wyznaczania zawartości wodoru na Jowiszu. Obliczono również potencjał jonizacyjny dla cząsteczki H_2 — wartość teoretyczna 124 416,8 cm^{-1} zgadza się bardzo dobrze z wartościami doświadczalnymi zmierzonymi przez Herzberga (124 418,4$\pm$0,4 cm^{-1}) lub przez Takezawę (124 417$\pm$2 cm^{-1}). Z innych wielkości, które obliczono, warto wymienić energie poziomów oscylacyjnych stanu podstawowego cząsteczki H_2, jak również krzywe energii potencjalnej dla kilku stanów wzbudzonych cząsteczki. Podobne obliczenia wykonano dla cząsteczek D_2 i HD.

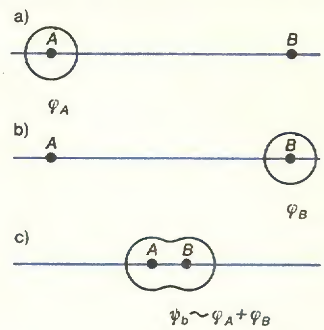
Cząsteczki dwuatomowe

Aby wyjaśnić wiele istotnych właściwości cząsteczek, nie musimy wykonywać tak skomplikowanych obliczeń, jak dla cząsteczki wodoru. Stan elektronów w cząsteczce decyduje w przeważającej mierze o jej właściwościach fizykochemicznych. Aby określić ten stan, należy oczywiście rozwiązać równanie elektronowe, które jest bardzo podobne do równania elektronowego dla atomu. Oba równania różnią się jedynie energią potencjalną — w atomie występuje oddziaływanie z jednym jądrem, natomiast w cząsteczce ruch elektronów zachodzi w polu co najmniej dwóch jąder.

Najprostszy opis stanu elektronów w cząsteczce otrzymujemy w modelu jednoelektronowym (por. Przybliżenie jednoelektronowe), a najlepszymi orbitalami molekularnymi są orbitale uzyskane za pomocą rozwiązania równań Hartree'ego-Focka. Niestety, dla cząsteczek wieloatomowych numeryczne rozwiązywanie tych równań jest niemożliwe i zadowalamy się przybliżeniami analitycznymi, które szczególnie dla dużych cząsteczek są mało dokładne. Niemniej, wyznaczone w ten sposób orbitale molekularne umożliwiają poglądowe wyjaśnienie właściwości cząsteczek i powiązanie ich z właściwościami atomów. Zadawalające jakościowe wyniki można uzyskać również w przypadku posługiwania się orbitalami molekularnymi nie wyznaczonymi metodą SCF.

jon H_2^+

Rozpatrzmy najprostszą cząsteczkę, jaką jest jon H_2^+. Przy dużych odległościach między jądrami układ ten składa się w zasadzie z atomu wodoru oraz protonu (sytuacja a lub b na rys. 23). Jeśli jednak odległość między jądrami jest mała, wówczas nie można



Rys. 23. Cząsteczka H_2^+ przy dużej (a, b) i małej (c) odległości międzyjądrowej. Zaznaczono przekroje konturów orbitali atomowych φ_A oraz φ_B typu 1s oraz wiążącego orbitalu molekularnego ψ_b. Kropkami zaznaczono położenie protonów

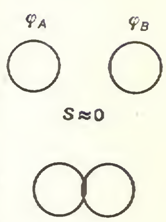
elektronu przypisać ani jądru A ani też jądru B — elektron jest wspólną własnością obu jąder. Elektron w cząsteczce H_2^+ przy dużej odległości międzyjądrowej R jest opisany albo orbitalem atomowym φ_A albo φ_B (oba typu 1s). W przypadku skończonej odległości międzyjądrowej stan podstawowy elektronu w H_2^+ opisujemy orbitalem molekularnym ψ i naturalne jest przybliżenie MO w postaci $\psi = c_1\varphi_A + c_2\varphi_B$. Współczynniki c_1 oraz c_2 można łatwo wyznaczyć z warunku normalizacji orbitalu ψ oraz z własności symetrii układu. Otrzymujemy wówczas dwa orbitale molekularne

$$\psi_b = \frac{1}{\sqrt{2(1+S)}}(\varphi_A+\varphi_B)$$
$$\psi_a = \frac{1}{\sqrt{2(1-S)}}(\varphi_A-\varphi_B), \tag{19}$$

gdzie $S = \int \varphi_A \varphi_B dv$ jest całką nakrywania przyjmującą wartości pośrednie pomiędzy 0 oraz 1 (rys. 24). Znając orbitale molekularne możemy obliczyć przyporządkowane im energie orbitalne. W przybliżeniu energie te wynoszą

$$\varepsilon_b = \varepsilon_H + \beta, \quad \varepsilon_a = \varepsilon_H - \beta, \tag{20}$$

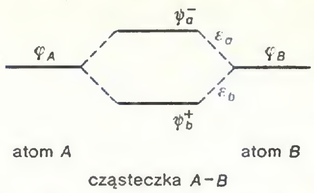
gdzie ε_H jest energią elektronu atomu wodoru w stanie opisanym orbitalem φ_A (lub φ_B), natomiast β jest tzw. całką rezonansową. Nie będziemy tutaj szczegółowo omawiać tej całki, zaznaczymy jedynie, że jej



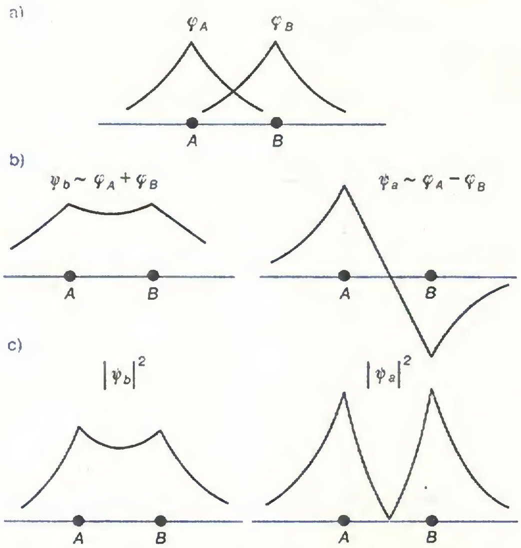
Rys. 24. Nakrywanie się przekrojów konturów orbitali atomowych typu 1s przy różnych odległościach międzyjądrowych R

całka rezonansowa

wartości nie są większe od zera (tzn. $\beta \leqslant 0$), oraz że $\beta \to 0$, gdy $R \to \infty$. Jak więc widać energia orbitalna ε_b jest niższa od energii orbitalnej ε_a. Ze wzorów



Rys. 25. Poziomy energetyczne w rozdzielonych atomach i w dwuatomowej cząsteczce homojądrowej. Znaki + oraz − przy wiążącym i antywiążącym orbitalu oznaczają, że orbitale te są przybliżone odpowiednio przez sumę lub różnicę orbitali atomowych φ_A oraz φ_B



Rys. 26. Wykresy wartości funkcji falowych wzdłuż osi międzyjądrowej: a) orbitale atomowe φ_A i φ_B typu 1*s*, b) wiążący i antywiążący orbital molekularny, c) kwadraty orbitali molekularnych

(20) wynika również, że energia ε_b ma niższą wartość od energii orbitalu atomowego φ_A (lub φ_B), natomiast energia ε_a ma wartość wyższą od tych energii (rys. 25).

Rysunek 26 przedstawia schemat wykresu obu orbitali molekularnych oraz wykres kwadratów tych funkcji. Na wykresie $|\psi_b|^2$ widać wyraźnie, że największa gęstość prawdopodobieństwa $\varrho_b = e|\psi_b|^2$ znalezienia elektronu w cząsteczce jest przy jądrach. Charakterystyczny przy tym jest fakt, że gęstość prawdopodobieństwa znalezienia elektronu pomiędzy jądrami jest duża — oznacza to, że istnieje wiązanie elektronowe pomiędzy jądrami. Orbital molekularny opisujący to wiązanie nazywamy dlatego orbitalem wiążącym (indeks *b* przy funkcji ψ_b jest skrótem wyrazu *bonding*). Wiążący orbital molekularny nie ma symetrii sferycznej, jak orbitale atomowe typu 1*s*, lecz ma symetrię cylindryczną wokół osi międzyjądrowej. O każdym orbitalu molekularnym tego typu będziemy mówili, że ma symetrię σ lub po prostu, że jest to MO typu σ. Wiązanie elektronowe, które jest opisane takim orbitalem, nazywać będziemy wiązaniem σ, a elektrony opisane takim orbitalem nazywać będziemy elektronami σ.

orbital wiążący

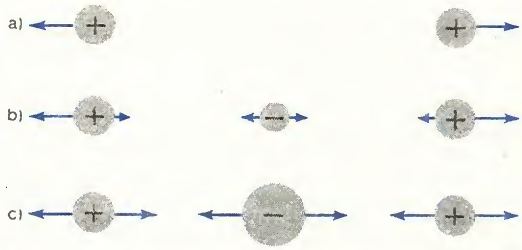
wiązanie σ

elektrony σ

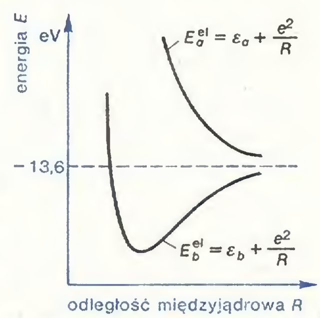
Wykres $|\psi_a|^2$ jest zupełnie inny od wykresu kwadratu funkcji ψ_b. Jeżeli elektron jest opisany orbitalem molekularnym ψ_a, wówczas gęstość prawdopodobieństwa $\varrho_a = e|\psi_a|^2$ znalezienia elektronu jest największa przy jądrach i jest zdecydowanie mała w obszarze między jądrami cząsteczki ($\varrho_a = 0$ dokładnie w środku odległości międzyjądrowej). Łatwo też zauważyć, że gęstość prawdopodobieństwa znalezienia elektronu poza obszarem międzyjądrowym jest dość duża dla niewielkich odległości od jądra. Te własności orbitalu molekularnego ψ_a uzasadniają nazywanie go orbitalem antywiążącym (indeks *a* przy funkcji ψ_a jest skrótem wyrazu *antibonding*). Zauważmy też, że orbital ψ_a ma podobnie jak orbital ψ_b symetrię cylindryczną; ma więc on symetrię σ. Aby jednak wyróżnić jego antywiążący charakter, oznaczać będziemy orbital ψ_a symbolem σ^*.

orbital antywiążący

Jaki wpływ na energię cząsteczki H_2^+ mają omówione różnice w rozkładzie ładunku elektronowego? Możemy to wyjaśnić za pomocą prostego modelu elektrostatycznego (rys. 27). Układ składający się z dwóch ładunków dodatnich (protony cząsteczki H_2^+) jest układem nietrwałym (sytuacja a). Wyobraźmy sobie, że między ładunkami dodatnimi pojawił się mały ładunek ujemny (sytuacja b). Układ ten jest nadal nietrwały, gdyż mały ładunek ujemny nie jest w stanie skompensować odpychających sił pomiędzy ładunkami dodatnimi. Dopiero gdy ładunek ujemny jest wystarczająco duży (sytuacja c), układ jest trwały w wyniku kompensacji sił odpychających przez siły przyciągające występujące pomiędzy ładunkami. Ten prosty model tłumaczy trwałość cząsteczki H_2^+ wówczas, gdy elektron jest opisany orbitalem ψ_b (orbitalowi temu odpowiada wzrost gęstości ładunku elektronowego między jądrami), oraz nietrwałość cząsteczki, gdy elektron jest opisany orbitalem ψ_a (orbitalowi temu odpowiada mały ładunek elektronowy pomiędzy jądrami). Zrozumiały się staje teraz przebieg krzywych potencjalnych $E^{el}(R)$ dla cząsteczki H_2^+ przedstawiony na rys. 28.



Rys. 27. Model elektrostatyczny wyjaśniający zależność wiążących i antywiążących własności orbitalu od rozkładu ładunku



Rys. 28. Krzywe energii potencjalnej dla dwóch najniższych stanów elektronowych (wiążącego i antywiążącego) jonu H_2^+; ε_H jest energią elektronu w stanie podstawowym atomu wodoru

Cząsteczki dwuatomowe homojądrowe

Orbitale molekularne ψ_b oraz ψ_a dla cząsteczki H_2^+ są orbitalami przybliżonymi utworzonymi z orbitali atomowych typu 1*s*. Dla cząsteczki H_2^+, jak również dla innych homojądrowych cząsteczek dwuatomowych, można utworzyć przybliżone orbitale molekularne z orbitali atomowych odpowiadających stanom wzbudzonym atomu wodoru. Na przykład można łatwo utworzyć orbitale molekularne z orbitali typu 2*s*:

$$\psi_b \sim 2s_A + 2s_B, \qquad \psi_a \sim 2s_A - 2s_B.$$

Zarówno niższy energetycznie orbital wiążący, jak i orbital antywiążący mają symetrię σ. W przypadku tworzenia orbitali molekularnych z orbitali atomo-

orbitale typu σ **orbitale typu π**

wych typu $2p$, ze względu na ukierunkowanie tych ostatnich, mamy dwa rodzaje MO (rys. 29) — orbitale o symetrii σ i orbitale o symetrii π. Te ostatnie orbitale są antysymetryczne względem odbicia w płaszczyźnie przechodzącej przez oś wiązania. Wprowadzona symbolika orbitali molekularnych (por. rys. 29, 30) wymaga krótkiego wyjaśnienia. Homojądrowe cząsteczki dwuatomowe mają tzw. środek symetrii, którym jest punkt leżący w środku odległości między jądrami cząsteczki. Orbitale molekularne, które są symetryczne względem odbicia w środku symetrii oznacza się wskaźnikiem g (niem. — *gerade*), natomiast MO będące antysymetrycznymi względem tego odbicia mają wskaźnik u (niem. — *ungerade*).

Znając przybliżoną postać orbitali molekularnych możemy ustalić elektronowe konfiguracje dla homojądrowych cząsteczek dwuatomowych. Należy oczywiście każdemu MO przyporządkować określoną energię. Niestety, nie można tego zagadnienia rozwiązać w sposób ogólny — dla każdej cząsteczki należy wykonać oddzielnie obliczenia numeryczne. Tak się jednak składa, że względna kolejność poziomów energetycznych na ogół nie zależy od rodzaju cząsteczki (rys. 30). Warto zwrócić uwagę na fakt, że orbitale $\pi(2p_x)$ i $\pi^*(2p_x)$ różnią się od odpowiednich orbitali $\pi(2p_z)$ i $\pi^*(2p_z)$ jedynie obrotem o 90° wokół osi y. Ponieważ żaden z kierunków prostopadłych do osi cząsteczki nie jest wyróżniony, orbitale $\pi(2p_x)$ i $\pi(2p_z)$ mają taką samą energię orbitalną (podobna sytuacja jest dla orbitali $\pi^*(2p_x)$ i $\pi^*(2p_z)$); występują więc poziomy podwójnie zdegenerowane.

Ustalenie konfiguracji elektronowej określonej cząsteczki jest bardzo proste (rys. 7). Poszczególne orbitale molekularne obsadzamy kolejno elektronami zaczynając od orbitali najniższych energetycznie i pamiętając, że na każdym orbitalu mogą być najwyżej dwa elektrony o przeciwnie skierowanych spinach, oraz pamiętając o regule Hunda. Podane w tabeli konfiguracje elektronowe dla kilku cząsteczek i ich niektórych jonów tłumaczą w prosty sposób obserwowane różnice w charakterze wiązań występujących w cząsteczkach i ich jonach. Na przykład cząsteczka H_2 różni się od swojego dodatniego jonu tym, że na orbitalu wiążącym ma dwa elektrony, a nie jeden. Wiązanie w H_2 ma więc dokładnie taki sam charakter, jak wiązanie w jonie H_2^+. Jest ono jednak znacznie silniejsze w neutralnej cząsteczce, gdyż wiązanie to pochodzi od dwóch elektronów. Dlatego też odległość między jądrami w H_2 jest mniejsza od odległości między jądrami w H_2^+. Jak widać więc dla utworzenia kowalencyjnego wiązania chemicznego nie jest niezbędna para elektronów. Niemniej wiązanie dwuelektronowe odgrywa w pewnym sensie wyróżnioną rolę, gdyż jest silniejsze od wiązania jednoelektronowego. Jest to jednak różnica ilościowa, a nie jakościowa.

wiązanie dwuelektronowe

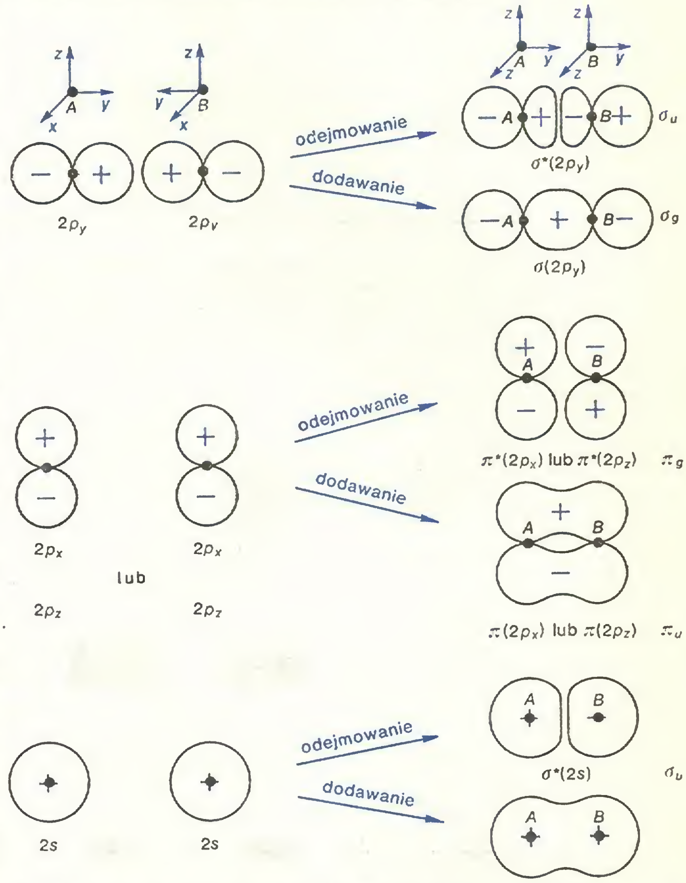
jon He_2^+

Rozpatrzmy z kolei jon He_2^+. W jonie tym dwa elektrony znajdują się na orbitalu wiążącym oraz jeden na orbitalu antywiążącym. Mamy więc w tym wypadku wiązanie trójelektronowe. Jest ono jednak słabsze od wiązania dwuelektronowego, gdyż efekty energetyczne jednego z elektronów znajdującego się na MO wiążącym i elektronu na MO antywiążącym znoszą się wzajemnie (zob. wzór (20)) — ściślej biorąc efekt antywiążący przeważa trochę nad efektem wiążącym. Jon dodatni cząsteczki helu jest układem trwałym, a jego wiązanie pochodzi w zasadzie od jednego elektronu znajdującego się na orbitalu wiążącym. Zrozumiały jest również fakt, że w przedstawionym schemacie jak i w rzeczywistości cząsteczka He_2 będzie układem nietrwałym (efekt wiążący dwóch elektronów na orbitalu $1\sigma_g$ jest znoszony przez efekt antywiążący dwóch elektronów na orbitalu $1\sigma_u$). Dwa atomy helu odpychają się więc wzajemnie (odpychanie walencyjne) nie tworząc cząsteczki.

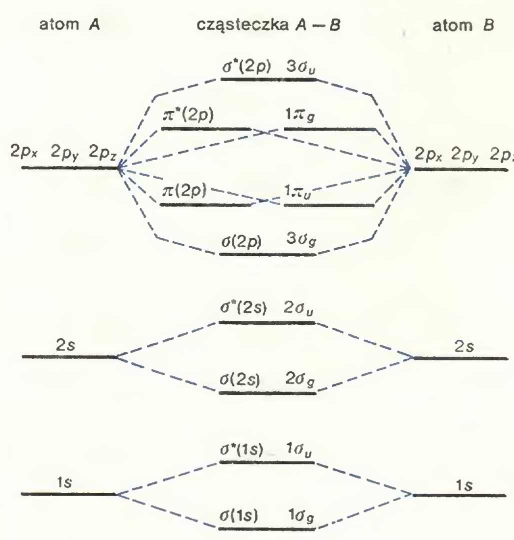
wiązanie trójelektronowe

Na przykładzie cząsteczki Li_2 można wytłumaczyć znany dobrze z chemii fakt, że elektrony powłok wewnętrznych nie biorą udziału w tworzeniu wiązań chemicznych. W miarę zbliżania się dwóch atomów litu przy tworzeniu cząsteczki przenikają się najpierw orbitale $2s$ (są one bardziej rozmyte w przestrzeni niż orbitale $1s$) i tworzy się orbital wiążący i antywiążący. Koncentracja ujemnego ładunku elektronowego w obszarze między jądrami cząsteczki powoduje dalsze zbliżanie się atomów. Jednocześnie jednak nastę-

cząsteczka Li_2



Rys. 29. Przekroje konturów orbitali molekularnych utworzonych w wyniku dodawania lub odejmowania orbitali atomowych $2s$ i $2p$ umieszczonych w punktach położeń jąder A i B. Kontury orbitali molekularnych utworzonych z orbitali atomowych $1s$ są podobne do konturów orbitali $\sigma(2s)$



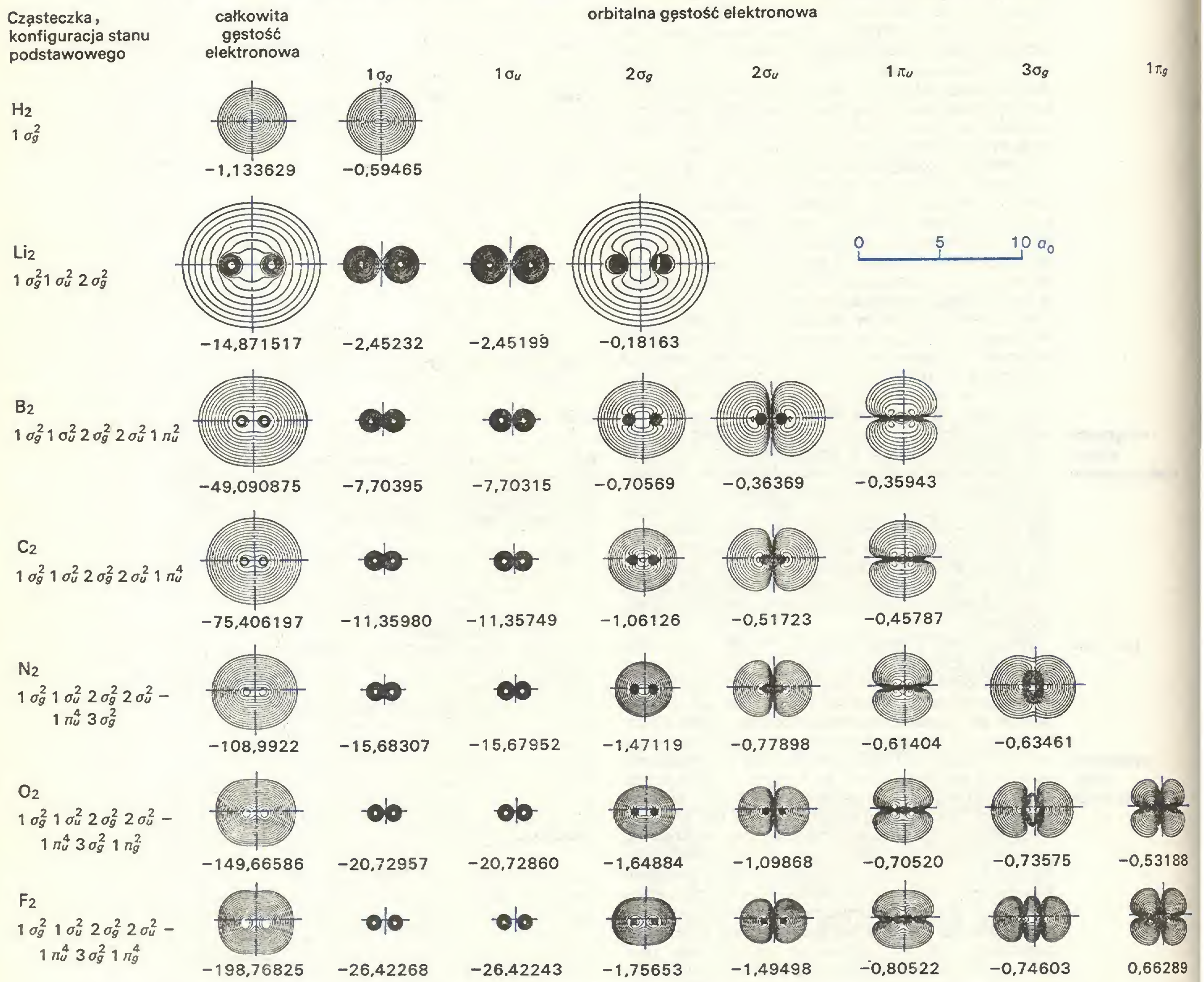
Rys. 30. Poziomy energetyczne rozdzielonych atomów oraz dwuatomowej cząsteczki homojądrowej

Konfiguracje elektronowe cząsteczek i jonów

Cząsteczka lub jon	Konfiguracja elektronowa	D_0 lub D_e (w nawiasie), eV	R_e, nm
H_2^+	$1\sigma_g$	(2,79)	0,106
H_2	$1\sigma_g^2$	(4,75)	0,073
He_2^+	$1\sigma_g^2\ 1\sigma_u$	(ok. 3)	
He_2	$1\sigma_g^2\ 1\sigma_u^2$		
Li_2	$1\sigma_g^2\ 1\sigma_u^2\ 2\sigma_g^2$	(1,05)	0,267
N_2^+	$1\sigma_g^2\ 1\sigma_u^2\ 2\sigma_g^2\ 2\sigma_u^2\ 1\pi_u^4\ 3\sigma_g$	8,72	0,121
N_2	$1\sigma_g^2\ 1\sigma_u^2\ 2\sigma_g^2\ 2\sigma_u^2\ 1\pi_u^4\ 3\sigma_g^2$	9,76	0,1108
O_2^+	$1\sigma_g^2\ 1\sigma_u^2\ 2\sigma_g^2\ 2\sigma_u^2\ 3\sigma_g^2\ 1\pi_u^4\ 1\pi_g$	6,48	0,1123
O_2	$1\sigma_g^2\ 1\sigma_u^2\ 2\sigma_g^2\ 2\sigma_u^2\ 3\sigma_g^2\ 1\pi_u^4\ 1\pi_g^2$	5,08	0,1207
F_2	$1\sigma_g^2\ 1\sigma_u^2\ 2\sigma_g^2\ 2\sigma_u^2\ 1\pi_u^4\ 3\sigma_g^2\ 1\pi_g^4$	(1,68)	0,142

puje zwiększenie się odpychania kulombowskiego pomiędzy jądrami atomów litu oraz pojawia się odpychanie walencyjne elektronów na orbitalach $1\sigma_g$ i $1\sigma_u$. Dlatego w cząsteczce litu mamy w zasadzie pojedyncze wiązanie (wiązanie σ) pochodzące od dwóch elektronów znajdujących się na orbitalu wiążącym $2\sigma_g$.

Elektrony w cząsteczce opisane orbitalami moleku-



Rys. 31. Warstwice gęstości elektronowych 7 homojądrowych cząsteczek dwuatomowych (wiązania kowalencyjne) otrzymane z obliczeń metodą HF (zob. rys. 9, podpis). Zewnętrzne warstwice odpowiadają gęstości $6{,}1\cdot10^{-5}\ e/a_0^3$. Warstwice gęstości elektronowych cząsteczek He_2, Be_2 oraz Ne_2, które nie są stabilne w stanie podstawowym, nie zostały wykreślone

wiązanie π, elektron π

larnymi typu π będziemy nazywali elektronami π, a wiązanie elektronowe opisane orbitalem π będziemy nazywali wiązaniem π. Często zdarza się, że obok wiązań π występują w cząsteczce wiązania σ. Na przykład w cząsteczce N_2 (zob. podane w tab. Konfiguracje elektronowe cząsteczek i jonów, zob. rys. 31) mamy wiązanie potrójne utworzone przez sześć elektronów: dwa wiązania π oraz jedno wiązanie σ (efekty wiążące elektronów na orbitalach $1\sigma_g$ i $2\sigma_g$ są znoszone przez efekty antywiążące elektronów na orbitalach odpowiednio $1\sigma_u$ i $2\sigma_u$).

Uważny czytelnik zapewne zwróci uwagę na fakt, że dla cząsteczek N_2, N_2^+ oraz F_2 sekwencja orbitali $3\sigma_g$ oraz $1\pi_u$ została odwrócona w porównaniu z rys. 30. W pierwszych dwóch przykładach zostało to uczynione wbrew obliczeniom (najdokładniejsze obliczenia przewidują sekwencję taką jak na rys. 30), bowiem stwierdzono doświadczalnie dla tych cząsteczek, że wyższym energetycznie orbitalem jest $3\sigma_g$ a nie $1\pi_u$. Natomiast dla cząsteczki F_2 obliczenia Hartree'ego–Focka podają odwróconą kolejność wspomnianych orbitali (rys. 31).

Z podanego schematu orbitali molekularnych wynika, że jonizacja cząsteczki N_2 prowadzi do oderwania elektronu znajdującego się na orbitalu wiążącym $3\sigma_g$ i wobec tego jon N_2^+ ma słabsze wiązanie (D_0 = 8,72 eV) niż neutralny układ N_2 (D_0 = 9,76 eV). Jeśli chodzi o cząsteczkę tlenu, to jej stan podstawowy jest trypletem ze względu na degenerację orbitali antywiążących $1\pi_g$ (reguła Hunda, zob. rys. 7). Wyjaśnienie tego faktu było jednym z większych osiągnięć teorii MO. Widzimy dalej, że zjonizowanie niektórych cząsteczek wzmacnia wiązanie między atomami (porównaj wartości D_0 oraz R_e dla O_2 z odpowiednimi wartościami dla O_2^+), gdyż odrywając elektron z orbitalu antywiążącego $1\pi_g$ zmniejszamy efekt antywiążący w cząsteczce. Zrozumiałą jest również rzeczą, że w jonach ujemnych O_2^- oraz O_2^{2-} efekty antywiążące są większe niż w O_2, a tym samym w jonach wiązania są słabsze i dłuższe (dla O_2, O_2^- oraz O_2^{2-} wartości R_e w nm wynoszą odpowiednio 0,121; 0,126 oraz 0,149).

Cząsteczki dwuatomowe heterojądrowe

Wiele trudności napotykamy przy tworzeniu orbitali molekularnych z orbitali atomowych dla dwuatomowych cząsteczek heterojądrowych. Nie występuje już w tym wypadku płaszczyzna symetrii prostopadła do osi cząsteczki i wobec tego współczynników w rozwinięciu $\psi = c_A\varphi_A + c_B\varphi_B$ należy szukać metodą wariacyjną. Kłopotliwy jest również dobór orbitali atomowych φ_A oraz φ_B, gdyż różnego typu AO w różnych atomach mają różne energie. Korzysta się tutaj jednak z takich orbitali atomowych, aby ich kombinacje liniowe były jak najbardziej efektywne. Oznacza to m.in., że energie określonych φ_A i φ_B nie powinny wiele różnić się między sobą. Dlatego też tworząc orbitale molekularne na przykład dla cząsteczki LiH powinniśmy pamiętać, że energia orbitalu $1s_H$ jest bliższa energii orbitalu $2s_{Li}$, aniżeli energii $1s_{Li}$ (por. rys. 10). Dwa elektrony znajdujące się na orbitalu atomowym 1*s* litu nie odgrywają istotnej roli w tworzeniu wiązania w cząsteczce LiH (rola tych elektronów sprowadza się w zasadzie do ekranowania jądra Li), natomiast wiązanie w cząsteczce jest głównie utworzone przez elektron 1*s* wodoru i elektron 2*s* litu — para ta zajmuje w stanie podstawowym cząsteczki orbital molekularny postaci $\psi = c_A 1s_H^A + c_B 2s_{Li}^B$. Cząsteczki heterojądrowe mają, podobnie jak homojądrowe, symetrię osiową, dlatego nadal obowiązuje podział orbitali molekularnych na orbitale typu σ oraz π (nie występuje tutaj jednak środek symetrii). Dlatego wspomniany wyżej orbital w LiH ma symetrię σ.

cząsteczka LiH

Przy tworzeniu orbitali molekularnych dla innych heterojądrowych cząsteczek dwuatomowych sytuacja jest na ogół bardziej złożona. Na przykład w cząsteczce CO orbitale 1*s* zarówno węgla jak i tlenu są skoncentrowane blisko odpowiednich jąder i praktycznie nie nakrywają się zupełnie z orbitalami 1*s* drugiego atomu. Nakrywanie się ich z orbitalami 2*s* lub $2p_y$ (przy założeniu, że oś *y* pokrywa się z osią cząsteczki drugiego atomu) jest trochę większe, jednak różnica pomiędzy energiami orbitali (por. rys. 10), np. 1*s* tlenu i 2*s* lub $2p_y$ węgla, jest tak duża, że i w tym wypadku nie mamy efektywnej kombinacji liniowej orbitali atomowych. Tak więc, elektrony powłok wewnętrznych cząsteczek heterojądrowych nie biorą również udziału w tworzeniu wiązania chemicznego. Natomiast orbitale atomowe 2*s* i 2*p* obu atomów tworzą już efektywne kombinacje liniowe. Ogólnie można powiedzieć, że wiązania są tworzone przez elektrony walencyjne znajdujące się na orbitalach walencyjnych (orbitale walencyjne — orbitale atomowe o największej dla danego atomu wartości głównej liczby kwantowej *n* w stanie podstawowym).

elektrony walencyjne, orbitale walencyjne

W celu otrzymania orbitali molekularnych dla cząsteczki CO tworzymy efektywne kombinacje liniowe walencyjnych orbitali atomowych. Z orbitali $2p_x$ tlenu oraz $2p_x$ węgla otrzymujemy wiążący i antywiążący orbital molekularny typu π. Podobnie z orbitali $2p_z$ otrzymujemy orbitale π. Natomiast gdy chcemy otrzymać orbitale molekularne typu σ z orbitali atomowych 2*s* i $2p_y$ poszczególnych atomów postępujemy inaczej. W przypadku cząsteczki heterojądrowej nie potrafimy rozstrzygnąć, które z czterech orbitali (orbitale 2*s* i $2p_y$ atomu tlenu oraz orbitale 2*s* i $2p_y$ atomu węgla) będą tworzyły najbardziej efektywną kombinację liniową (z porównania energii orbitalnych nie możemy skorzystać, gdyż energie tych czterech orbitali są bliskie sobie). Dlatego też orbitale molekularne σ dla cząsteczki CO zakładamy w postaci

cząsteczka CO

$$\psi_i = c_{i1}2s_C + c_{i2}2p_{yC} + c_{i3}2s_O + c_{i4}2p_{yO},$$

gdzie wskaźniki C i O odnoszą się odpowiednio do atomów węgla i tlenu. Z czterech orbitali atomowych można utworzyć cztery orbitale molekularne, różniące się współczynnikami (oraz energiami orbitalnymi) — dwa z nich są orbitalami wiążącymi, a dwa orbitalami antywiążącymi.

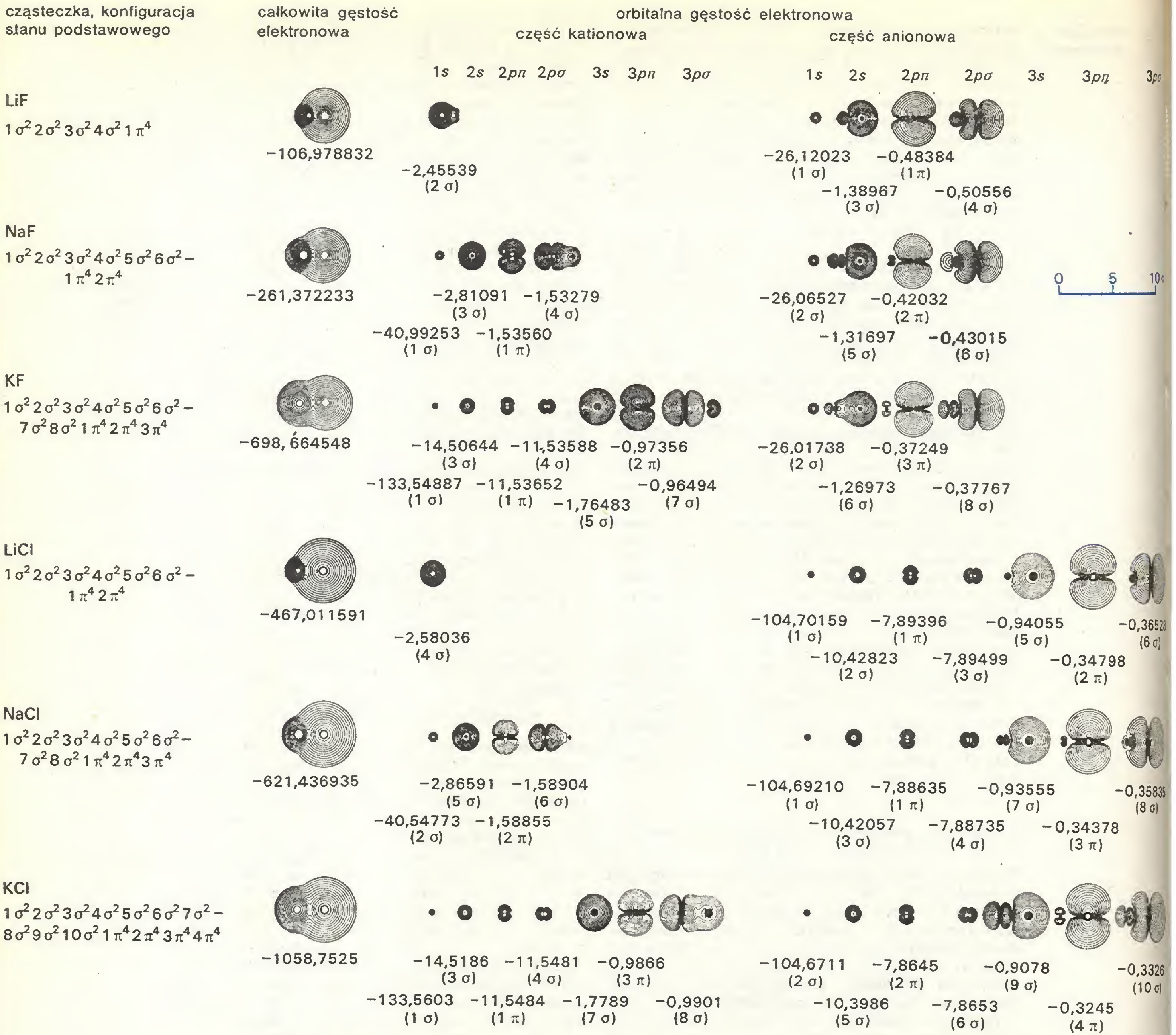
Trudno jest nam ustalić konfigurację elektronową cząsteczek heterojądrowych, ponieważ m.in. dla różnych cząsteczek kolejność orbitali może być różna. Zazwyczaj dzieli się po prostu wszystkie orbitale na orbitale σ i π, a w obrębie każdej grupy numeruje się je kolejno 1σ, 2σ, ..., 1π, 2π, ... według wzrastającej energii orbitalnej (podobnie postąpiono w przypadku cząsteczek homojądrowych, por. rys. 31). Dla cząsteczki CO ustalono (korzystając z wyników doświadczalnych i teoretycznych) następującą konfigurację elektronową CO: $1\sigma^2 2\sigma^2 3\sigma^2 4\sigma^2 1\pi^4 5\sigma^2$.

ustalanie konfiguracji

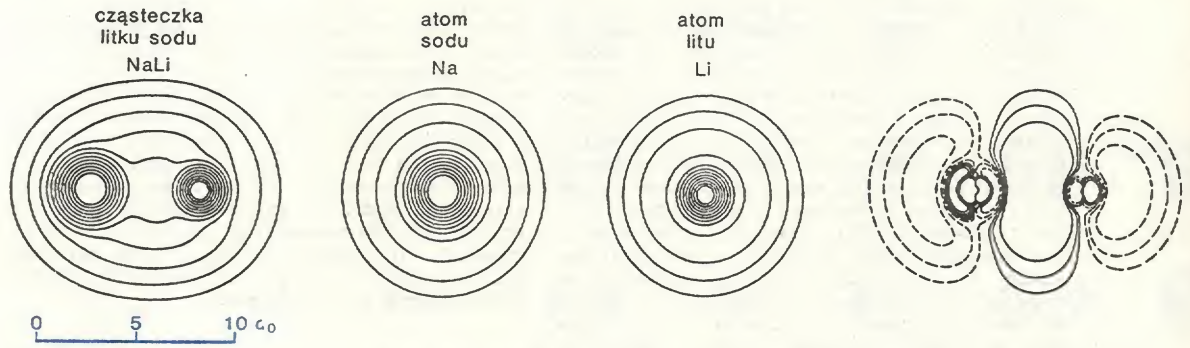
Chcąc uzyskać możliwie najdokładniejsze wyniki ilościowe dla cząsteczek, nie należy ograniczać się do przybliżania orbitali molekularnych za pomocą dwóch tylko orbitali atomowych, jak to uczyniono w przypadku cząsteczek homojądrowych. Dla cząsteczek dwuatomowych, ze względu na występującą w nich symetrię osiową, można obliczyć dokładne orbitale HF, które dają najlepszy obraz cząsteczki w przybliżeniu jednoelektronowym. Rysunki 31 i 32 przedstawiają warstwice gęstości elektronowych dla kilku dwuatomowych homo- i heterojądrowych cząsteczek. Widać tu wyraźnie, jak poszczególne orbitale molekularne zmieniają się przy przejściu od cząsteczki do cząsteczki. Widać również wyraźnie, jak zmieniają się energie poszczególnych orbitali molekularnych w serii cząsteczek.

Wiązanie chemiczne

Obliczenia gęstości elektronowych w układach dwuatomowych są pomocne w zrozumieniu koncepcji



Rys. 32. Warstwice gęstości elektronowych 6 halogenków metali alkalicznych (wiązania jonowe) otrzymane z obliczeń metodą Hartree'ego-Focka. Warstwice gęstości elektronowych dla poszczególnych orbitali uporządkowano ze względu na ich jonowe pochodzenie (zob. rys. 9, podpis)



Rys. 33. Warstwice gęstości elektronowych (obliczone z funkcji HF) stabilnej cząsteczki NaLi i atomów tworzących tę cząsteczkę. Obok podano warstwice „różnic gęstości elektronowych", tzn. różnic między gęstością stabilnej cząsteczki i gęstościami niezaburzonych atomów umieszczonych względem siebie w odległości odpowiadającej położeniu równowagi cząsteczki. Linie ciągłe oznaczają obszary dodatnich wartości różnic gęstości (w wyniku utworzenia cząsteczki z atomów gęstość elektronowa w tym obszarze zwiększyła się), linie przerywane oznaczają obszary ujemnych wartości różnic gęstości (w wyniku utworzenia cząsteczki z atomów gęstość elektronowa w tym obszarze zmniejszyła się). Zob. rys.9, podpis

wiązania chemicznego. Możemy bowiem analizować wiązanie chemiczne badając różnice w gęstościach elektronowych cząsteczki i atomów (rys. 33). W przypadku stabilnej cząsteczki gęstość elektronowa w obszarze między jądrami cząsteczki jest większa w porównaniu do gęstości elektronowej izolowanych atomów.

Za pomocą obliczeń możemy obserwować zmiany zachodzące w gęstościach elektronowych poszczególnych atomów przy ich wzajemnym zbliżaniu się (praktycznie prawie przy ciągłej zmianie odległości międzyjądrowej, rys. 35). Szczególnie interesujący jest przypadek cząsteczki o wiązaniu jonowym LiF (rys. 35B). Rozkład gęstości elektronowej w cząsteczce przy dużych odległościach międzyjądrowych odpowiada rozkładowi gęstości elektronowej w izolowanych neutralnych atomach litu i fluoru (rys. 9). Rozkład ten ulega gwałtownej zmianie już przy odległości ok. $13{,}9a_0$ (ok. 0,74 nm); (rys. 35B, punkt B). Tworzą się wówczas jony Li^+ oraz F^-, które prawie nie zmienione zbliżają się nadal do położenia równowagi przy ok. $2{,}94a_0$ (ok. 0,156 nm). A więc układ jonowy Li^+F^- tworzy się przy znacznie większej odległości międzyjądrowej niż położenie równowagi cząsteczki! W układzie He_2 dwa atomy helu odpychają się, w przeciwieństwie do dwóch atomów wodoru w układzie H_2. Gęstość elektronowa w układzie He_2 jest wypychana z obszaru międzyjądrowego. Takie zachowanie się atomów jest charakterystyczne dla atomów gazów szlachetnych posiadających zapełnione powłoki elektronowe.

wiązanie jonowe

Pomiary a obliczenia

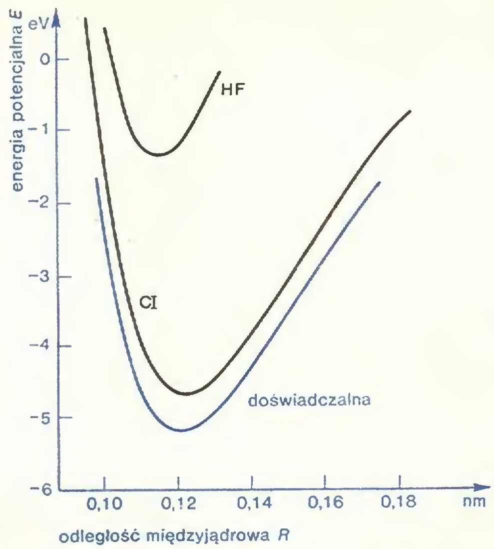
Podobnie jak w obliczeniach Hartree'ego–Focka dla atomów, tak i dla cząsteczek obliczone energie HF stanowią przeszło 99% energii dokładnej cząsteczki (np. dla Li_2: $\mathcal{E}_{HF} = -14{,}8715$ j.at.; $E_{dośw} = -14{,}9944$ j.at.; dla F_2: $\mathcal{E}_{HF} = -198{,}7683$ j.at., $E_{dośw} = -199{,}670$ j.at.). Znacznie gorsze wyniki uzyskuje się przy obliczaniu energii wiązania lub energii dysocjacji cząsteczek. Na przykład obliczona metodą HF energia dysocjacji cząsteczki O_2 wynosi 1,43 eV, co stanowi zaledwie 27% wartości doświadczalnej 5,21 eV. Dla cząsteczki F_2, której doświadczalna energia wiązania wynosi 1,68 eV, obliczenia metodą HF wykazują, że jest ona nietrwała (ujemna energia wiązania −1,37 eV!).

Przyczyną tych rozbieżności jest stosowanie przybliżenia jednoelektronowego, w którym pomija się korelację elektronów. Energia korelacji dla F_2 jest większa o ok. 3 eV od energii korelacji dla dwóch izolowanych atomów fluoru — wystarcza to, aby otrzymać za pomocą metody HF absurdalny wynik dla energii wiązania cząsteczki F_2. Wspominaliśmy już, że energię korelacji można uwzględnić w obliczeniach, posługując się metodą oddziaływania konfiguracyjnego. Połączenie metody CI z procedurą SCF dało dla F_2 energię wiązania 1,67 eV w doskonałej zgodności z doświadczeniem. Podobnie dla cząsteczki O_2 uwzględnienie oddziaływania konfiguracyjnego daje nie tylko dobry wynik dla energii dysocjacji cząsteczki (4,72 eV, tj. 91% wartości doświadczalnej), ale przebieg krzywej energii potencjalnej jest również w dobrej zgodności z krzywą doświadczalną (rys. 34).

W wielu przypadkach obliczenia Hartree'ego–Focka dają wartości teoretyczne, które są w dobrej zgodności z danymi doświadczalnymi. Na przykład przewidywane położenia równowagi dla halogenków alkalicznych, obliczone momenty dipolowe (charakteryzujące niejednorodność rozkładu ładunku elektrycznego w cząsteczkach), jak również wartości wielu wielkości spektroskopowych są w zasadzie w bardzo dobrej zgodności z doświadczeniem (tabela).

Nie zawsze jednak taka zgodność występuje, zwłaszcza gdy zmierzone wartości doświadczalne są małe, np. gdy moment dipolowy cząsteczki jest mniejszy od 1 D (1 debaja). Dobrym przykładem ilustrującym występujące tu trudności jest przypadek cząsteczki CO. Wartość doświadczalna momentu dipolowego cząsteczki jest −0,112 D (znak − oznacza kierunek momentu dipolowego C^-O^+). Moment dipolowy obliczony metodą HF różni się od doświadczalnego momentu o ok. 0,4 D i ma przeciwny znak (C^+O^-), wynosi bowiem 0,274 D. Dopiero uwzględnienie w obliczeniach oddziaływania konfiguracyjnego poprawia teoretyczny wynik do wartości −0,077 D.

moment dipolowy cząsteczki CO



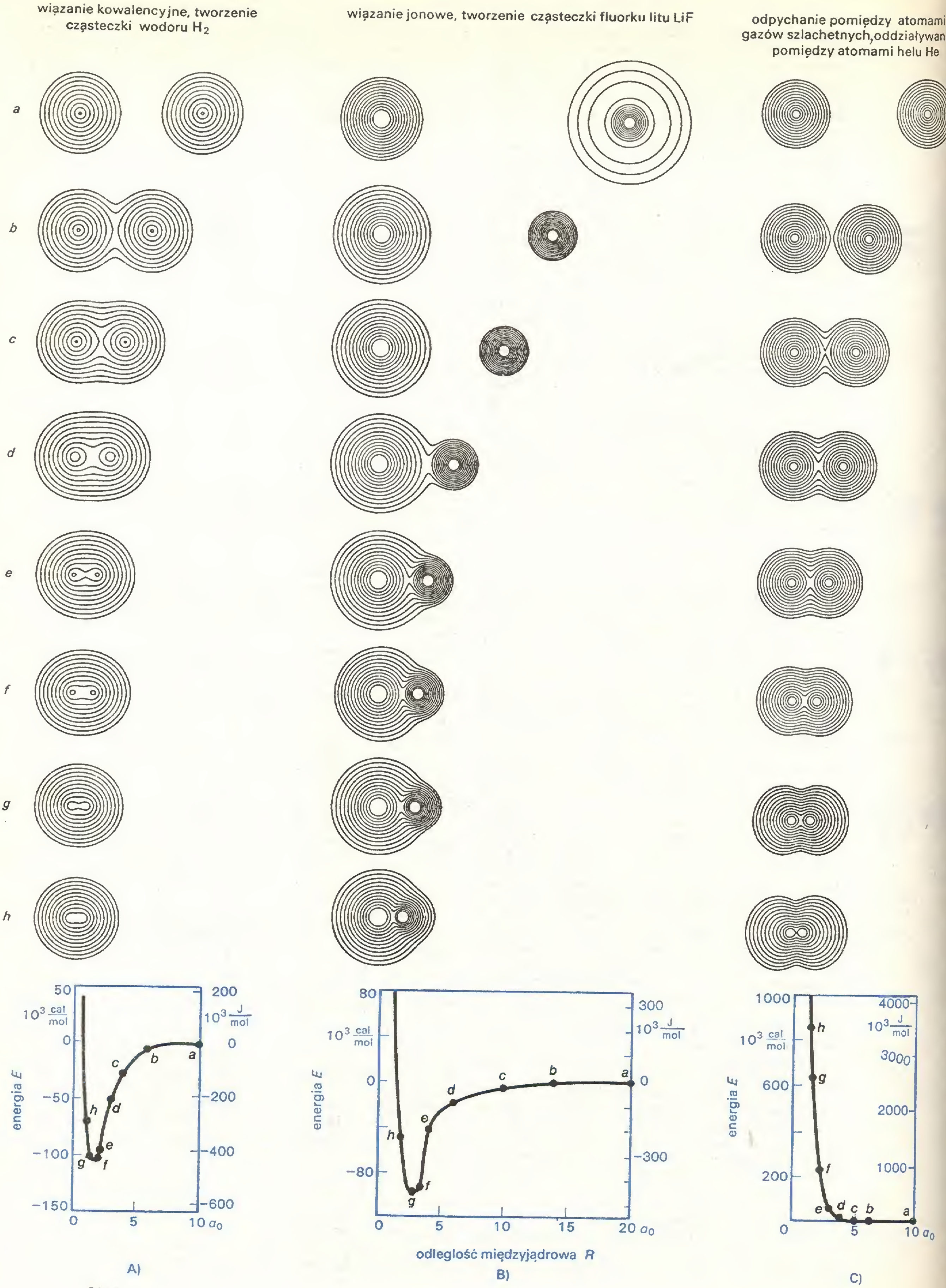
Rys. 34. Porównanie doświadczalnej krzywej energii potencjalnej dla stanu podstawowego cząsteczki O_2 z krzywą obliczoną metodą Hartree'ego–Focka (HF) oraz obliczoną przy uwzględnieniu oddziaływania konfiguracyjnego; wg H.F. Scheafer III, J. Chem. Phys. 54, 2207 (1971)

Momenty dipolowe i stałe spektroskopowe halogenków alkalicznych otrzymane metodą Hartree'ego–Focka. Wyniki porównano z danymi doświadczalnymi (wartości w nawiasach)

Wielkość	LiCl	NaF	NaCl	KF
Moment dipolowy μ, D (debaj)	7,256 (7,128)	8,367 (8,206)	9,138 (9,002)	8,720 (8,6)
Energia wiązania D_e, eV	3,83 (4,84)	3,05 (4,95)	3,18 (4,22)	3,05 (5,09)
Odległość międzyjądrowa R_e, a_0,	3,825 (3,819)	3,628 (3,639)	4,485 (4,460)	4,188 (4.104)
Częstość drgań podstawowych, ω_e, cm^{-1}	672 (641)	558 (536)	378 (365)	449 (426)
Stała rotacyjna B_e, cm^{-1}	0,7042 (0,7065)	0,439 (0,437)	0,2155 (0,2181)	0,2686 (0,2799)
Poprawka anharmoniczna $\omega_e x_e$, cm^{-1}	4,67 (4,2)	4,39 (3,83)	2,59 (2,05)	2,48 (2,43)
Poprawka do poziomów rotacyjnych α_e, cm^{-1}	0,00444 —	0,00468 (0,00457)	0,00180 (0,00162)	0,00214 (0,00233)
Stała siłowa wiązania, k_e, mdyn/nm	1,55 (1,41)	1,91 (1,76)	1,17 —	1,51 (1,31)

Wg R.L. Matca, J. Chem. Phys. 47, 4595, 5295 (1967); 48, 335 (1968); 49, 1264 (1968).

Porównanie teoretycznych wartości z danymi doświadczalnymi wskazuje, że obliczenia HF dla układów dwuatomowych mogą niejednokrotnie dawać wyniki, które różnią się od danych doświadczalnych nie więcej niż o ok. 10–20%. Taka dokładność w wielu wypadkach jest wystarczająca dla celów porównawczych. Dlatego też kilka lat temu powstał pomysł zbu-



wiązanie kowalencyjne, tworzenie cząsteczki wodoru H2
wiązanie jonowe, tworzenie cząsteczki fluorku litu LiF
odpychanie pomiędzy atomami gazów szlachetnych, oddziaływanie pomiędzy atomami helu He
a
b
c
d
e
f
g
h
energia E
10³ cal/mol
10³ J/mol
50
0
−50
−100
−150
200
0
−200
−400
−600
0
5
10 a0
A)
80
0
−80
300
0
−300
0
5
10
15
20 a0
odległość międzyjądrowa R
B)
1000
600
200
4000
3000
2000
1000
0
5
10 a0
C)

dowania „instrumentu obliczeniowego", za pomocą którego chemik lub fizyk, którzy nie są specjalistami z zakresu programowania i nie znają szczegółów metod matematycznych potrzebnych do rozwiązywania równań HF, mogliby otrzymywać teoretyczne wartości określonych wielkości fizykochemicznych cząsteczki. Szczególne znaczenie ma to wówczas, gdy wykonanie pomiarów jest bardzo trudne (np. wymagające budowania specjalnej aparatury) lub nawet niebezpieczne (praca ze związkami trującymi). Chemik lub fizyk może wówczas zdecydować czy wykonywać pomiary, czy też oprzeć się na wartościach teoretycznych (wprawdzie przybliżonych, ale stosunkowo łatwo osiągalnych). Kilka lat temu A.C. Wahl i jego współpracownicy z Argonne National Laboratory (USA) zbudowali system obliczeniowy nazwany BISONem.

Za pomocą tego systemu można obecnie dla dowolnego układu dwuatomowego obliczać funkcje falowe Hartree'go–Focka, a następnie uzyskiwać informacje o gęstościach elektronowych w stanie podstawowym oraz w stanach wzbudzonych, obliczać parametry cząsteczek (np. momenty dipolowe) czy też stałe spektroskopowe, jak również analizować krzywe energii potencjalnej. Wyniki końcowe w zależności od życzenia mogą być wydrukowane na papierze, przedstawione na ekranie kineskopu lub za pomocą filmu (np. zmiany w czasie gęstości ładunku elektronowego przy tworzeniu cząsteczki). Rys. 9 oraz 31–33 i 35 zostały właśnie otrzymane za pomocą systemu BISON. Co ciekawsze, BISON został tak skonstruowany, aby fizyk czy chemik mógł porozumieć się z „przyrządem obliczeniowym" za pomocą mowy ludzkiej. Muszą oni jednak używać pewnych słów kodu zrozumiałego dla systemu BISON. Typowa konwersacja pomiędzy chemikiem a BISONem może wyglądać następująco:

BISON

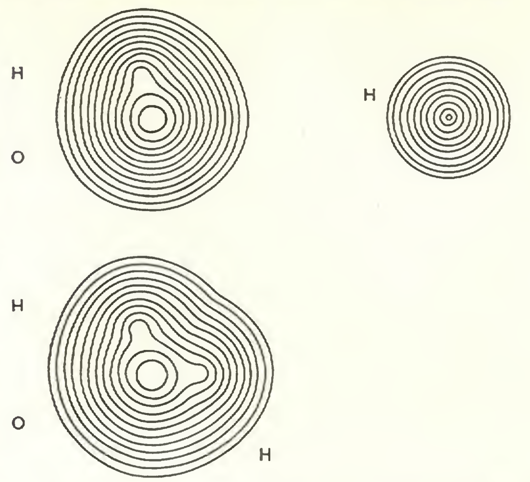
Chemik: Chciałbym znać wartość momentu dipolowego najniższego energetycznie stanu dubletowego π cząsteczki tlenku wapnia (CaO).
BISON: Przy jakiej odległości międzyjądrowej?
Chemik: Przy odległości równowagi.
BISON: Jeżeli należy obliczyć położenie równowagi, czas obliczeń będzie wynosił ok. 4 h. Prawdopodobny błąd momentu dipolowego będzie wynosił ok. 20%. Czy mam wykonywać obliczenia?

Jeżeli chemik odpowie „Tak", system obliczeniowy zacznie wykonywać operacje obliczeniowe.

Dokładność wyników teoretycznych uzyskiwanych za pomocą systemu BISON w niektórych wypadkach nie jest oczywiście zadowalająca. Dlatego też kontynuowane są prace nad modyfikacją tego systemu. W przyszłości chemik lub fizyk będzie mógł uzyskiwać za pomocą tego systemu (lub podobnego) wyniki teoretyczne, które będą bardziej dokładne od dotychczasowych (system zostanie udoskonalony m.in. przez uwzględnienie oddziaływania konfiguracyjnego). Niemniej już dzisiaj chemik kwantowy może w wielu wypadkach uzyskiwać wyniki teoretyczne, które dostarczają cennych wskazówek doświadczalnikom. I to nie tylko dla małych układów cząsteczek dwuatomowych, ale również dla znacznie większych — dużych układów wieloatomowych.

Małe układy wieloatomowe

Wspominaliśmy już o tym, że znajdowanie orbitali molekularnych HF dla cząsteczek zbudowanych z więcej niż dwóch atomów stwarza duże trudności numeryczne. Praktycznie dla takich układów znajduje się orbitale molekularne przybliżone przez liniową kombinację funkcji jednoelektronowych (13). Niemniej takie orbitale molekularne znajdowane metodą SCF



Rys. 36. Tworzenie cząsteczki wody z atomu wodoru i cząsteczki OH. Warstwice gęstości elektronowej OH, H oraz H_2O obliczone zostały metodą HF. Wewnętrzne warstwice dla OH i H_2O odpowiadają gęstości $1{,}0e/a_0^3$, a dla atomu wodoru gęstości $0{,}25e/a_0^3$; warstwice zewnętrzne odpowiadają gęstości $4{,}9\cdot10^{-4}$ e/a_0^3; wg A.C. Wahl, U. Blukis, J. Chem. Educ. 45, 787 (1968)

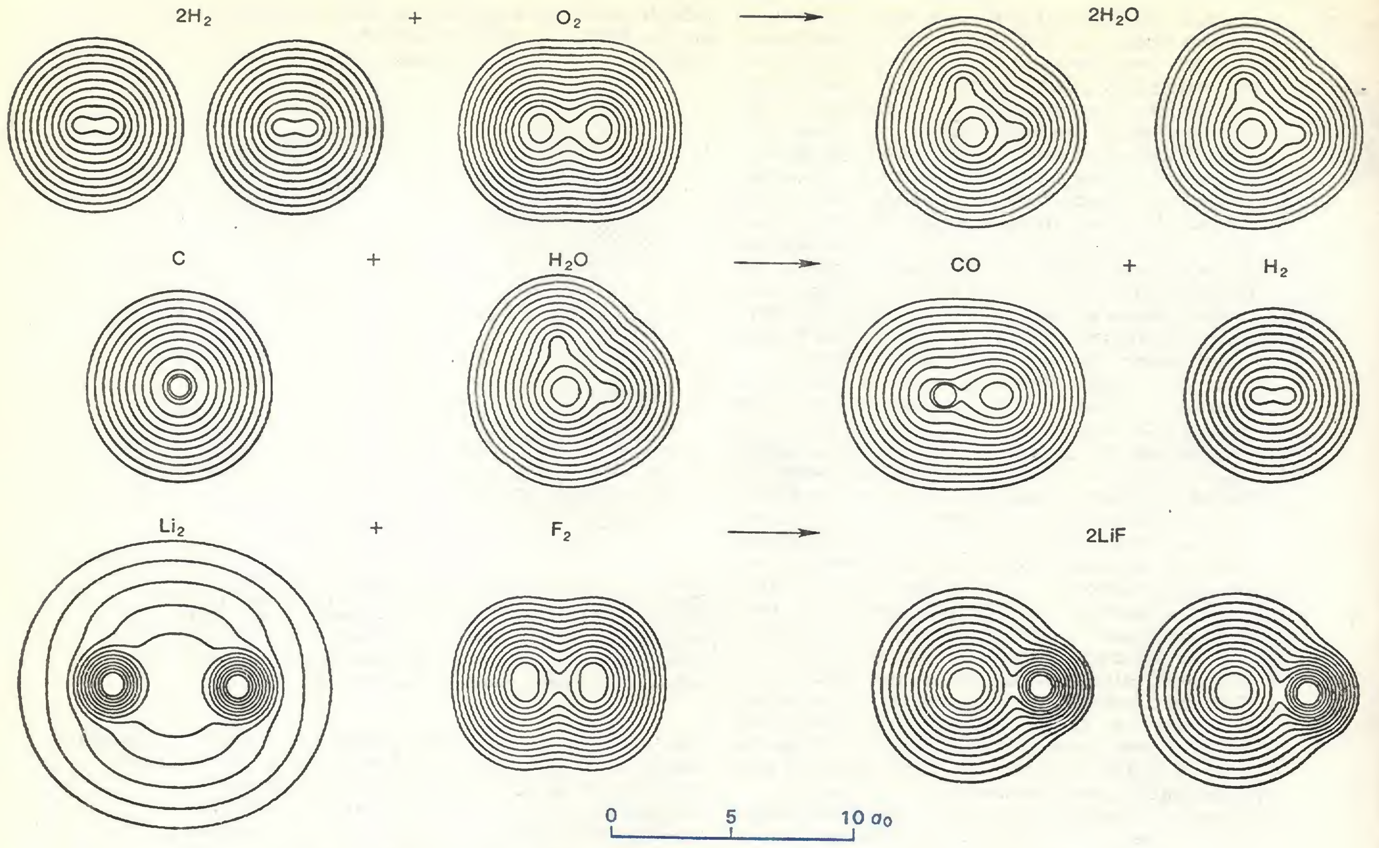
cząsteczka wody

dają np. poprawny opis całkowitego ładunku elektronowego w cząsteczce. Można więc przyjąć, że na przykład rys. 36 poprawnie przedstawia rozkład ładunku elektronowego w cząsteczce wody. Dzięki obliczeniom teoretycznym możliwe jest szczegółowe przeanalizowanie zmian rozkładu ładunku elektronowego cząsteczki; wówczas, gdy znajduje się ona w sąsiedztwie innej cząsteczki (lub atomu), w wyniku oddziaływania między nimi nastąpi reakcja chemiczna (rys. 37).

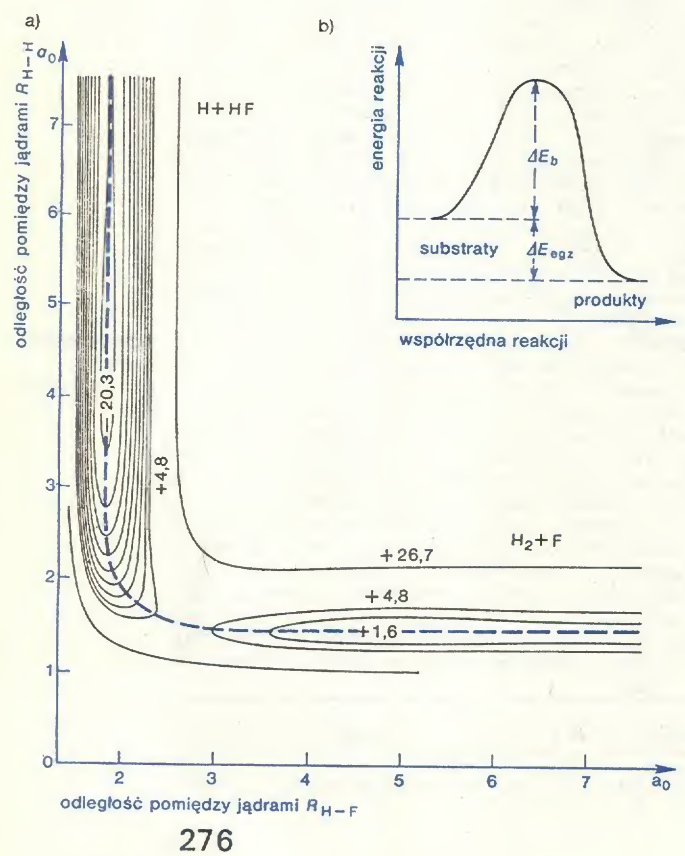
Metody chemii kwantowej mogą być również z powodzeniem zastosowane do badania reakcji chemicznych za pomocą tzw. powierzchni energii potencjalnej układów. Określenie tych powierzchni jest pierwszym krokiem przy teoretycznej dyskusji na temat szybkości reakcji chemicznych. Ogólne własności powierzchni energii potencjalnych można prześledzić na przykładzie liniowej reakcji $F+H_2 \rightarrow FH+H$. Rysunek 38 przedstawia względną energię całkowitą układu trzech atomów (fluor oraz dwa atomy wodoru) przy ich różnych wzajemnych odległościach. Energię tę obliczano przez odjęcie energii substancji reagujących (tzn. energii atomu F i energii cząsteczki H_2) od energii całkowitej układu: jeden atom fluoru + dwa atomy wodoru. Obliczenia dla wspomnianej reakcji wykonano w przybliżeniu jednoelektronowym metodą SCF oraz z uwzględnieniem efektów korelacji metodą SCF CI. Rysunek 38 przedstawia powierzchnię energii potencjalnej uzyskaną za pomocą tego drugiego, bardziej dokładnego, sposobu. Linie stałej energii naniesione są jako linie warstwicowe. Powierzchnia energii potencjalnej utworzona jest z dwóch długich wąskich dolin przedstawiających trwałe cząsteczki H_2 i HF. Obszar odpowiadający substancjom reagującym (stan początkowy) oddzielony jest od obszaru odpowiadającego produktom (stan końcowy) obszarem o wyższej energii, w wyniku czego reakcja ma energię aktywacji (odpowiada jej bariera reakcji ΔE_b zaznaczona na rys. 38). Obliczona wysokość bariery metodą SCF CI wynosi 5,8 kcal/mol (metoda SCF bez oddziaływa-

powierzchnia energii potencjalnej

Rys. 35. Sekwencja warstwic całkowitej gęstości elektronowej w układach H_2, LiF i He_2 przy różnej odległości międzyjądrowej R. A) Wiązanie kowalencyjne: tworzenie cząsteczki wodoru H_2. B) wiązanie jonowe: tworzenie cząsteczki fluorku litu LiF. C) odpychanie między atomami gazów szlachetnych na przykładzie oddziaływania między atomami helu. Wewnętrzna warstwica dla H_2 odpowiada gęstości $0{,}25e/a_0^3$, a wewnętrzne warstwice dla LiF i He_2 odpowiadają gęstości $1{,}0e/a_0^3$. Warstwice zewnętrzne odpowiadają gęstości $4{,}9\cdot10^{-4}e/a_0^3$. Wykresy u dołu ukazują zmiany całkowitej energii cząsteczek wraz ze zmianą R



Rys. 37. Trzy reakcje chemiczne scharakteryzowane przy pomocy warstwic obrazujących różnice w gęstości elektronowych substratów i produktów reakcji; obliczenia przeprowadzone zostały metodą HF



nia konfiguracyjnego daje wysokość bariery równą aż 34,3 kcal/mol), podczas gdy wyznaczona doświadczalnie energia aktywacji wynosi ok. 1,7 kcal/mol (ok. 7 kJ/mol). Przy tak małych wartościach energii aktywacji nie można jej porównywać bezpośrednio z wysokością bariery reakcji. Prawdopodobnie doświadczalna bariera reakcji ma wysokość rzędu 5 kcal/mol (ok. 20 kJ/mol) lub nieco mniej. Wskazuje to na bardzo dobrą zgodność wyników obliczeń metodą SCF CI z doświadczeniem w porównaniu z obliczeniami SCF nie uwzględniającymi efektów korelacji. Obliczenia SCF CI dają jeszcze jeden bardzo interesujący wynik — wskazują one mianowicie, że reakcja $F+H_2 \rightarrow FH+H$ jest reakcją egzotermiczną. Obliczona bowiem energia końcowa reakcji jest niższa od energii początkowej o 20,4 kcal/mol, co stosunkowo dobrze zgadza się z odpowiednią wartością doświadczalną wynoszącą 31,2 kcal/mol. Warto w tym miejscu podkreślić, że obliczenia SCF bez CI nie przewidują, że reakcja jest egzotermiczna — różnica pomiędzy energią stanu końcowego a energią stanu początkowego reakcji została obliczona jako dodatnia (+0,6 kcal/ /mol). Uwzględnienie efektów korelacyjnych w obliczeniach, w których porównuje się różnice wielkości, ma w wielu przypadkach istotne znaczenie (zwiększenie rozmiarów oddziaływania konfiguracyjnego z

reakcja $F+H_2 \rightarrow FH+H$

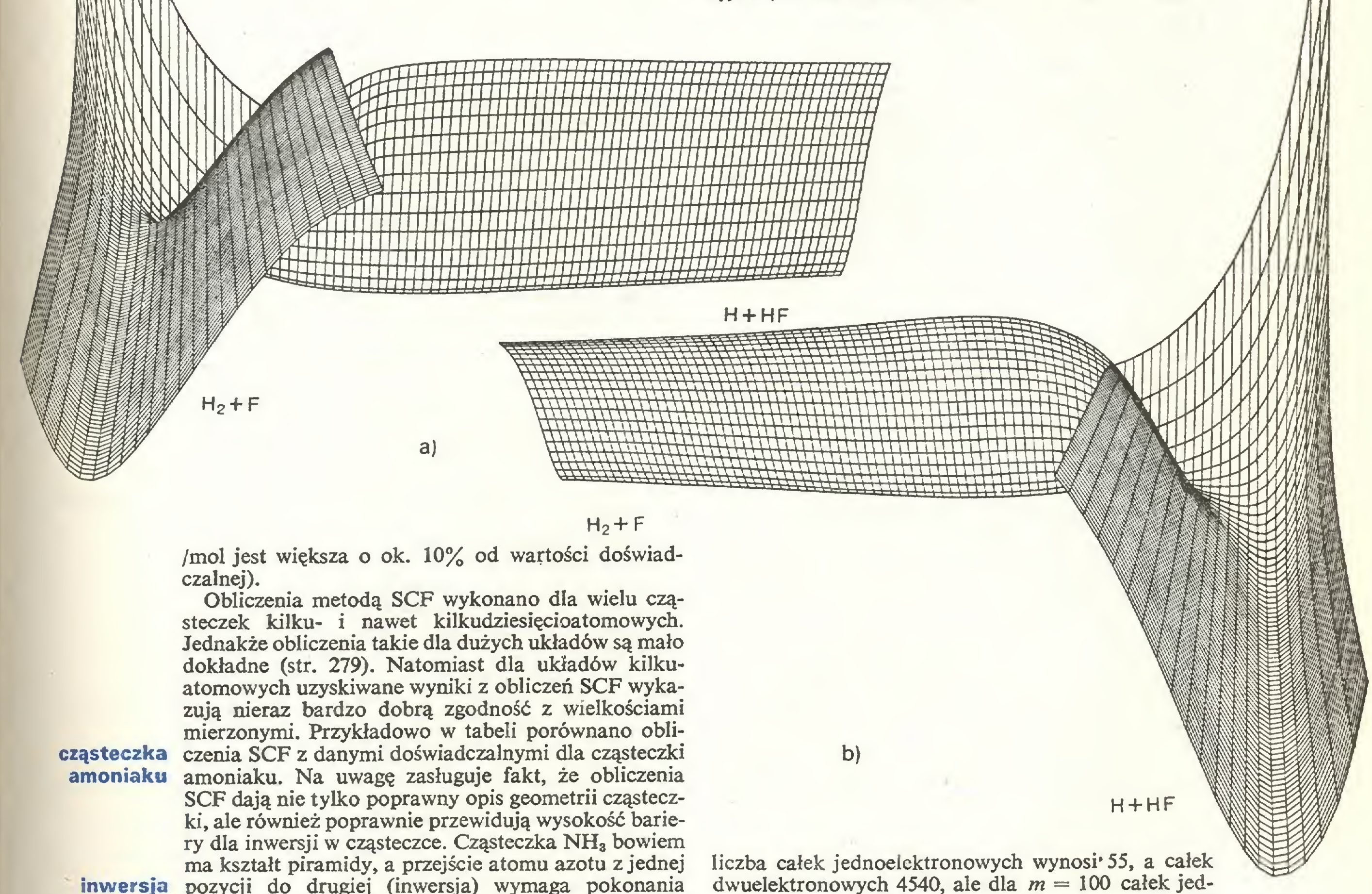
Rys. 38. Teoretyczny opis reakcji chemicznej: a) Tradycyjna dwuwymiarowa powierzchnia energii potencjalnej liniowego układu FH_2 wyznaczona na podstawie obliczeń SCF CI. Energia podana jest w kcal/mol. Odległość F-H zmienia się od $1,4a_0$ do $8,0a_0$, a odległość H-H od $1,0a_0$ do $7,6a_0$. Linia przerywana oznacza „drogę" reakcji odpowiadającą najmniejszej energii. b) Schematyczne przedstawienie zależności energii reakcji od współrzędnej reakcji, czyli od odległości wzdłuż drogi reakcji o najmniejszej energii (linia przerywana na rys. a). Zaznaczono wysokość bariery ΔE_b (energię aktywacji) oraz energię ΔE_{egz} charakteryzującą egzotermiczność reakcji

Porównanie wyników obliczeń metodą SCF dla cząsteczki NH_3 z danymi doświadczalnymi

Wielkość	Wartość	
	obliczona metodą SCF	doświadczalna
Długość wiązania N—H, nm	0,1000	0,10116
Kąt między wiązaniami H—N—H, stopnie	107,2	106,8
Moment dipolowy μ, D	1,66	1,48
Energia całkowita, j.at.	−56,2219	−56,578
Bariera dla inwersji, kcal/mol	5,1	5,8

Wg A. Rauk, L.C. Allen, E. Clementi, J. Chem. Phys. 52, 4133 (1970).

214 konfiguracji do 338 konfiguracji poprawiło zgodność obliczonej energii ΔE_{egz} z wartością doświadczalną — wartość obliczona $\Delta E_{egz} = -34{,}4$ kcal/



/mol jest większa o ok. 10% od wartości doświadczalnej).

Obliczenia metodą SCF wykonano dla wielu cząsteczek kilku- i nawet kilkudziesięcioatomowych. Jednakże obliczenia takie dla dużych układów są mało dokładne (str. 279). Natomiast dla układów kilkuatomowych uzyskiwane wyniki z obliczeń SCF wykazują nieraz bardzo dobrą zgodność z wielkościami mierzonymi. Przykładowo w tabeli porównano obliczenia SCF z danymi doświadczalnymi dla cząsteczki amoniaku. Na uwagę zasługuje fakt, że obliczenia SCF dają nie tylko poprawny opis geometrii cząsteczki, ale również poprawnie przewidują wysokość bariery dla inwersji w cząsteczce. Cząsteczka NH_3 bowiem ma kształt piramidy, a przejście atomu azotu z jednej pozycji do drugiej (inwersja) wymaga pokonania bariery potencjału, którą można zarówno zmierzyć, jak i obliczyć. Jak widać na przykładzie cząsteczki NH_3, obliczenia SCF z powodzeniem zostały zastosowane do opisu tej bariery (mogą więc one być zastosowane do badań konformacyjnych).

cząsteczka amoniaku

inwersja

Metody nieempiryczne i empiryczne chemii kwantowej

Metoda SCF MO, w której orbitale molekularne poszukiwane są w postaci (13), może być zastosowana do dowolnego układu chemicznego. Jednakże dla dużych układów obliczenia są kosztowne, a wyniki uzyskiwane, jak już wspominaliśmy kilkakrotnie, są mało dokładne. Aby uzyskać bowiem dla danego układu jak najlepszy obraz orbitalny, tzn. aby wyniki uzyskane metodą SCF MO były jak najbardziej zbliżone do wyników metody Hartree'ego–Focka, przy szukaniu orbitali molekularnych w postaci (13) musimy używać jak najdłuższej bazy (jak największe m w (13)). Z punktu widzenia rachunkowego nie jest to jednak proste. Przyjęcie dużej liczby funkcji bazy powoduje, że dla układu wieloatomowego rośnie liczba całek jedno- i dwuelektronowych występujących w równaniach SCF (12). Gdy liczba funkcji bazy wynosi m, wówczas liczba całek jednoelektronowych równa jest $m(m+1)/2$, natomiast liczba całek dwuelektronowych wynosi $[m(m+1)/4][m(m+1)/2+1]$. Jak widać liczba tych ostatnich całek gwałtownie rośnie wraz ze wzrostem liczby funkcji bazy (wprawdzie dla $m = 10$

Rys. 39. Trójwymiarowa powierzchnia energii potencjalnej (kanał reakcji) układu liniowego FH_2: a) kanał reakcji widziany od strony wejścia ($H_2 + F$), b) kanał reakcji widziany od strony wyjścia ($FH+H$). Zob. rys. 38

liczba całek jednoelektronowych wynosi 55, a całek dwuelektronowych 4540, ale dla $m = 100$ całek jednoelektronowych mamy 5050, a dwuelektronowych już 12 751 250). Obliczanie dużej liczby całek jest bardzo pracochłonne nawet dla szybkiego komputera.

Czytelnik zdaje sobie oczywiście sprawę, że rozmiar użytej bazy zależy od tego, czy zbieżność rozwinięcia (13) jest szybka, czy też nie, tzn. czy używając określonej bazy (wybrana postać funkcji bazy oraz ich liczba) wyniki uzyskane metodą SCF MO są zbliżone do wyników metody HF. Okazuje się, że używając orbitali atomowych jako bazy uzyskujemy stosunkowo szybką zbieżność rozwinięcia (13). Pojawia się jednak wówczas dodatkowy kłopot. Dla cząsteczek zawierających więcej niż dwa atomy obliczanie niektórych całek występujących w równaniach (12) jest skomplikowane. Można uniknąć tego kłopotu używając w roz-

zbieżność rozwinięcia

winięciu (13) nie orbitali atomowych (AO mają wykładniczą zależność od odległości elektronu od jądra postaci $e^{-Z'r/a_0}$, zob. (14)), ale tzw. funkcji gaussowskich postaci $e^{-\alpha r^2}$. Niestety, unikając kłopotu z obliczeniem szeregu wspomnianych całek w (12), traci się zbieżność rozwinięcia (13). Innymi słowy, aby zreprodukować wyniki obliczeń uzyskane dla określonej cząsteczki metodą SCF przy użyciu AO musimy, używając jako bazy funkcji Gaussa, użyć znacznie większego rozwinięcia (13). A to z kolei zwiększa liczbę całek jedno-, a szczególnie dwuelektronowych.

Pomimo trudności typu numerycznego, wykonano obliczenia za pomocą metody SCF MO dla wielu złożonych układów, np. dla zasad kwasów nukleinowych i dla ich pary guanina-cytozyna (układ 136 elektronów), a ostatnio dla 2,4,7-trójnitrofluorenonu zawierającego aż 160 elektronów. E. Clementi i jego współpracownicy, którzy wykonali te obliczenia przyznają, że wyniki obliczeń są bardzo przybliżone, a jednocześnie bardzo pracochłonne. Na przykład dla pary guanina-cytozyna należało obliczyć ok. 70 mld całek dwuelektronowych. Pomimo tego, że autorzy ci wykonywali obliczenia na bardzo szybkim komputerze IBM 360/195 (obliczał ok. 10^5 całek na 1 s), obliczenia SCF trwały ok. 8 dni. Autorzy wspominają, że wykonywane przez nich 4 lata wcześniej obliczenia na IBM 7094 dla reakcji $NH_3 + HCl \rightleftarrows NH_4Cl$ przebiegały ok. 400 razy wolniej (a więc gdyby chciano wówczas wykonać obliczenia dla pary guanina-cytozyna, musiano by zużyć na to 3200 dni!). W ciągu kilku lat nastąpił więc bardzo duży postęp w technice obliczeniowej. Ale nie tylko w technice obliczeniowej. Autorzy powyższych obliczeń w ciągu kilku lat zmodyfikowali i udoskonalili także program obliczeniowy SCF MO. Połączenie myśli matematycznej programistów (matematyków, chemików kwantowych) z jednoczesnym rozwojem komputerów doprowadziło do takiego przyspieszenia obliczeń.

Obliczenia HF czy też SCF MO, o których wspominaliśmy dotąd, są obliczeniami nieempirycznymi. W obliczeniach tego typu nie korzysta się z danych doświadczalnych — wychodząc z założeń i postulatów mechaniki kwantowej wykonuje się obliczenia aż do uzyskania wyników, które można porównać z danymi doświadczalnymi. Jedynymi wartościami doświadczalnymi stosowanymi w teorii są stałe uniwersalne, takie jak ładunek i masa elektronu, czy też stała Plancka. Teorie takie w chemii kwantowej nazywane są często teoriami *ab initio*, bowiem w teoriach tych obliczane są wszystkie całki występujące od momentu rozpoczęcia obliczeń (czyli *ab initio*).

teorie *ab initio*

Mimo dynamicznego rozwoju w dziedzinie komputerów, wysoki koszt obliczeń *ab initio* i ich niewystarczająca dokładność dla dużych układów chemicznych powodują, że nadal w chemii kwantowej używane są metody przybliżone, szczególnie metody półempiryczne. W metodach tych niektóre całki występujące w teorii wyznacza się z danych doświadczalnych. Omówiliśmy poprzednio w sposób jakościowy wiązanie chemiczne w układach dwuatomowych (str. 267). Zamiast obliczać wartość całki rezonansowej β występującej w wyrażeniu na energię orbitalu wiążącego (wzór (20)), możemy tak dobrać jej wartość, aby pewne wyniki, np. energię cząsteczki lub jej potencjał jonizacyjny otrzymać zgodne z doświadczeniem. Postępowanie takie umożliwia nieraz wyznaczenie innych wielkości teoretycznych, już bez odwoływania się do doświadczenia. W taki sposób sformułowana teoria nosi nazwę teorii półempirycznej, a całki, które traktuje się jako wielkości dobierane, nazywamy parametrami teorii. Teorie półempiryczne są często bardziej przydatne od teorii *ab initio*, szczególnie w odniesieniu do dużych cząsteczek. Jednak rozmiar układów chemicznych, dla których obliczenia *ab initio* stają się użyteczne, zwiększa się z każdym rokiem.

teorie półempiryczne

W ciągu kilkudziesięcioletniej historii chemii kwantowej opracowano tak dużo wersji różnych metod półempirycznych, że nie sposób je tu wszystkie omówić. Wspomnijmy jedynie o najbardziej charakterystycznych metodach.

Najstarsza metoda półempiryczna chemii kwantowej wyrosła na gruncie badań własności cząsteczek organicznych. Elektrony walencyjne atomów wchodzących w skład tych cząsteczek tworzą charakterystyczne dwie grupy wiązań — wiązania σ oraz wiązania π. Te ostatnie wiązania utworzone są przez układ elektronów π odpowiedzialny za wiele charakterystycznych własności cząsteczek organicznych. W przybliżeniu π-elektronowym rozpatruje się jedynie układ elektronów π cząsteczki, tzn. znajduje się funkcję falową i energię dla tych tylko elektronów. Zakłada się, że elektrony π poruszają się w pewnym potencjale pochodzącym od jąder cząsteczki i układu elektronów σ. W najprostszym sformułowaniu zaproponowanym przez E. Hückla (metoda HMO — Hückel MO) poszukuje się orbitali molekularnych dla elektronów π cząsteczki w postaci LCAO (jako bazę funkcji w (13) wybiera się jedynie orbitale walencyjne atomów oddających elektrony π do układu). Drastycznym przybliżeniem wprowadzonym w metodzie HMO jest pominięcie oddziaływania między elektronami π w hamiltonianie π-elektronowym — przyjmuje się go w postaci $\hat{H} = \sum_{i=1}^{n} \hat{H}^{ef}(i)$, gdzie suma przebiega po wszystkich elektronach π cząsteczki. Współczynniki c_{pi} rozwinięcia MO na AO (13) oblicza się metodą Ritza (por. wzory (8) oraz (9)). Parametrami empirycznymi teorii są całki kulombowskie $\alpha_p (\equiv H_{pp} = \int \varphi_p^* \hat{H}^{ef} \varphi_p dv)$ oraz całki rezonansowe β_{pq} $(\equiv H_{pq} = \int \varphi_p^* \hat{H}^{ef} \varphi_q dv)$. Rozwiązując układ równań typu (9), otrzymujemy orbitale molekularne ψ_i oraz odpowiadające im energie orbitalne ε_i.

przybliżenia π-elektronowe

metoda HMO — Hückel MO

Gdy uwzględniamy w hamiltonianie π-elektronowym bezpośrednio oddziaływanie między elektronami (tzn. w hamiltoniaanie π-elektronowym występuje wyraz $\sum_{i<j} e^2/r_{ij}$, gdzie suma przebiega po wszystkich elektronach π), orbitale molekularne ψ_i są nadal przybliżone, tak jak w metodzie HMO w postaci LCAO (13), ale znajduje się je przez rozwiązanie równań SCF. Parametrami empirycznymi metody są w zasadzie całki rezonansowe typu β_{pq}, jednak dodatkowo wiele całek występujących w elementach macierzowych operatora Hartree'go–Focka $\hat{F}$ w (12) przybliża się przez wartości doświadczalne (np. przez potencjały jonizacyjne czy też powinowactwa elektronowe poszczególnych stanów walencyjnych atomów) lub też przyjmuje się na nie przybliżoną postać (np. wzory ekstrapolacyjne na całki dwuelektronowe). Metoda półempiryczna SCF LCAO MO jest szczególnie użyteczna przy interpretacji elektronowych widm absorpcyjnych cząsteczek organicznych. Okazuje się jednak, że w tym celu należy użyć konfiguracyjnego oddziaływania (zob. Korelacja elektronów) — konfiguracje wzbudzone dla cząsteczki tworzy się z orbitali molekularnych uzyskanych z rozwiązań równań SCF. Taką wersję metody półempirycznej nazywa się metodą SCF LCAO MO CI, a jej bardzo popularny wariant z przybliżeniami wyprowadzonymi przez trzech chemików kwantowych nosi nazwę metody Parisera–Parra–Pople'a (lub metody PPP), od nazwisk autorów metody.

Obie wspomniane metody można rozszerzyć na elektrony walencyjne cząsteczki (i to oczywiście nie tylko dla cząsteczek organicznych, ale dla dowolnego układu chemicznego). Rozszerzenie metody HMO nosi nazwę rozszerzonej teorii Hückla (EHT — Extended Hückel Theory) — najbardziej popularna wersja tej metody została opracowana przez R. Hoffmana w latach 1963–64. W metodzie tej, podobnie jak w HMO, hamiltonian układu elektronów walencyjnych jest określony w postaci $\hat{H} = \sum_i \hat{H}^{ef}(i)$ (tutaj suma przebiega po wszystkich elektronach walencyjnych układu), a orbitale molekularne (będące kombi-

rozszerzona teoria Hückla — EHT

nacją liniową orbitali walencyjnych atomów tworzących cząsteczkę) znajduje się metodą Ritza. W metodzie EHT sposób oceny całek α_p oraz β_{pq} jest inny niż w metodzie HMO, pominiemy jednak tutaj szczegóły parametryzacji.

Jeśli w hamiltonianie dla elektronów walencyjnych uwzględni się bezpośrednie oddziaływania pomiędzy elektronami, to orbitale molekularne znajduje się wówczas za pomocą procedury SCF. Niektórzy badacze wprowadzili niezależnie od siebie różne przybliżenia na poszczególne całki występujące w elementach macierzowych operatora $\hat{F}$ — przybliżeń tych jest dużo, i nie będziemy ich tutaj wymieniać. Najbardziej popularne z tych metod są: metoda CNDO/2 opracowana przez J.A. Pople'a i G.A. Segala (1966), oraz wersja tej metody z dołączonym konfiguracyjnym oddziaływaniem — metoda CNDO/S CI (J. Del Bene i H.H. Jaffe, 1968), metoda INDO (J.A. Pople i współpracownicy 1967) oraz metoda MINDO/2 (M.J.S. Dewar i E. Haselbach, 1970). Metody te, uwzględniające elektrony walencyjne i oparte na procedurze SCF, nazywane są często metodami AVE (*All Valence Electrons*).

metody AVE (CN-DO/2; INDO; MINDO/2)

Półempiryczne teorie miały i mają nadal bardzo duże znaczenie dla chemii. Niekorzystną sytuację stwarza fakt, że jest tych teorii za dużo. Każda ze wspomnianych metod (HMO, SCF MO CI, EHT, AVE) ma wiele wariantów różniących się sposobem oceny parametrów półempirycznych czy też sposobem przybliżeń wprowadzonych do oceny całek. Przyczyną takiej sytuacji jest to, że nie można sformułować teorii półempirycznej w ten sposób, aby interpretowała ona wszystkie fakty doświadczalne. Wartość wyznaczonych parametrów półempirycznych określonej metody zależy od tego, do interpretacji jakich własności fizykochemicznych cząsteczki została użyta teoria. Na przykład metoda CNDO/2 odnosi duże sukcesy w interpretacji geometrii cząsteczek, natomiast przewidywane widma elektronowe cząsteczek nie pokrywają się z doświadczeniem. Poprawną interpretację widm elektronowych otrzymuje się wówczas, gdy nie tylko dołączy się konfiguracyjne oddziaływanie, ale gdy dodatkowo w metodzie CNDO/2 zmieni się niektóre całki (dokonać więc trzeba zmiany parametryzacji). Taka zmieniona metoda CNDO/S CI poprawnie opisuje własności spektroskopowe cząsteczek, ale nie jest już ona w stanie opisać poprawnie ich geometrii stanu podstawowego. Nic więc dziwnego, że chemicy kwantowi rozbudowują obliczenia *ab initio*, aby uwolnić się od niepewności spowodowanej wprowadzoną parametryzacją.

CNDO/S CI

Dlatego też należy z dużą ostrożnością podchodzić do wyników obliczeń półempirycznych. Niemniej, kilkudziesięcioletnia praktyka tych obliczeń wskazuje, że metody półempiryczne w wielu przypadkach opisują poprawnie korelacje pomiędzy własnościami serii podobnych cząsteczek.

Duże układy

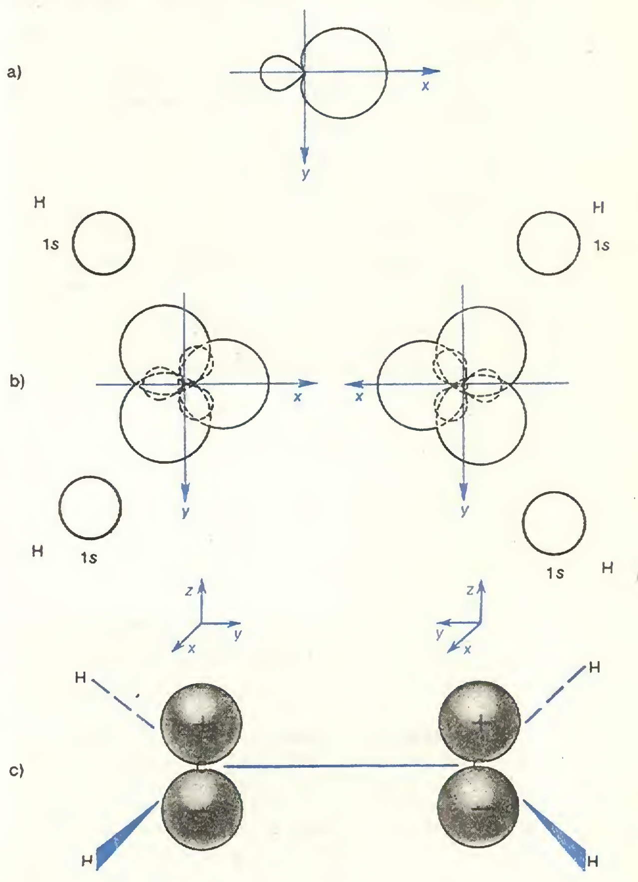
Przez duże układy będziemy rozumieli tutaj wieloelektronowe układy chemiczne, np. cząsteczkę organiczną lub nieorganiczną, kompleks tych cząsteczek, kryształ lub polimer. Ograniczymy nasze rozważania w zasadzie jedynie do cząsteczek organicznych (układy te były przedmiotem najbardziej intensywnych badań teoretycznych). Są to na ogół układy płaskie (lub też układy, których niektóre części są płaskie), odznaczające się dużą stabilnością energetyczną. Zanim omówimy niektóre zastosowania chemii kwantowej do badania własności tych układów, opiszemy jakościowo strukturę elektronową najprostszych cząsteczek, rozpoczynając od skrótowego przedstawienia pojęcia orbitali zhybrydyzowanych na przykładzie cząsteczki etylenu.

Orbitale zhybrydyzowane

Cząsteczka etylenu C_2H_4 jest układem płaskim i symetrycznym — odległości między środkami atomów węgla i wodoru (długość wiązania C—H) są takie same, a kąty H—C—H utworzone między dwiema prostymi liniami wyrysowanymi od środka atomu węgla do atomu wodoru wzdłuż wiązań (kąty wiązań) są prawie równe 120°. W najprostszym schemacie jednoelektronowym (str. 251) wiązania w etylenie utworzone są przez elektrony walencyjne atomów wchodzących w skład cząsteczki (elektrony $2s$ i $2p$ dwóch atomów węgla i elektrony $1s$ czterech atomów wodoru). Omawiając konfiguracje elektronowe atomów (zob. Atomy wieloelektronowe) podaliśmy dla węgla konfigurację stanu podstawowego $1s^2 2s^2 2p^2$. Jest to konfiguracja elektronowa stanu podstawowego swobodnego atomu węgla. Gdyby taką konfigurację przyjąć dla węgla w cząsteczce, nie można by było wytłumaczyć w prostym schemacie równocenności szeregu wiązań cząsteczki. W cząsteczce atom nie jest swobodny; jest on otoczony innymi atomami, które powodują zmianę rozkładu elektronów walencyjnych atomu w porównaniu z atomem swobodnym. Mówimy, że atom w cząsteczce znajduje się w określonym stanie walencyjnym. Przez stan walencyjny atomu będziemy rozumieli pewien hipotetyczny rozkład elektronów atomu. Stan walencyjny określonego atomu węgla np. w cząsteczce etylenu to taki rozkład elektronów tego atomu, który powstałby w wyniku usunięcia wszystkich pozostałych atomów z cząsteczki z ich elektronami do nieskończoności przy niezmienionym ukierunkowaniu wiązań. W stanie walencyj-

cząsteczka etylenu

stan walencyjny atomu



Rys. 40. Kształty orbitali zhybrydyzowanych i niezhybrydyzowanych atomów węgla i wodoru w cząsteczce etylenu: a) przekrój powierzchni granicznej orbitalu zhybrydyzowanego typu sp^2; b) przekroje powierzchni granicznych orbitali zhybrydyzowanych sp^2 atomów węgla i orbitali $1s$ atomów wodoru; c) orbitale $2p_z$ atomów węgla

orbitale zhybrydyzowane (hybrydy)

nym atomu elektrony znajdują się na tzw. zhybrydyzowanych orbitalach, które można przedstawić jako liniową kombinację AO swobodnego atomu (każdy orbital zhybrydyzowany ma więc częściowo cechy różnych orbitali swobodnego atomu). W przypadku atomu węgla w cząsteczce etylenu mamy do czynienia z hybrydyzacją trygonalną (lub hybrydyzacją typu sp^2) — w wyniku mieszania jednego orbitala typu $2s$ i dwóch orbitali typu $2p$ otrzymujemy trzy równoważne orbitale (hybrydy) sp^2:

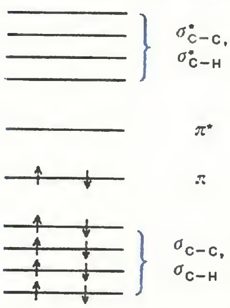
$$h_i = c_{1i}\varphi_{2s}+c_{2i}\varphi_{2p_x}+c_{3i}\varphi_{2p_y}, \quad (i = 1, 2, 3).$$

Podczas hybrydyzacji tego typu jeden z orbitali $2p$ (tutaj $2p_z$) pozostaje niezmieniony.

Na rys. 40 przedstawiono przekroje powierzchni granicznych orbitali sp^2. Okazuje się, że orbitale sp^2 są symetryczne względem odbicia w płaszczyźnie cząsteczki (przy przyjęciu układu współrzędnych jak na rys. 40 jest nią płaszczyzna xy). Natomiast orbital $2p_z$ jest antysymetryczny względem odbicia w płaszczyźnie cząsteczki.

cząsteczka etylenu

Korzystając więc z orbitali zhybrydyzowanych można bardzo łatwo wytłumaczyć tworzenie się wiązań w cząsteczce etylenu. Cząsteczka ta ma pięć wiązań typu σ. Jedno z tych wiązań, wiązanie C—C, utworzone jest przez dwa zhybrydyzowane orbitale sp^2 atomów węgla cząsteczki i opisane jest wiążącym orbitalem molekularnym (analogicznie jak w przypadku cząsteczek dwuatomowych orbital molekularny jest kombinacją liniową dwóch zhybrydyzowanych orbitali atomowych sp^2 zlokalizowanych na dwóch atomach węgla). Cztery pozostałe wiązania typu σ (są to wiązania C—H) opisane są orbitalami molekularnymi utworzonymi z kombinacji liniowych orbitala atomowego $1s$ atomu wodoru i jednego zhybrydyzowanego orbitala sp^2 atomu węgla. W cząsteczce etylenu — oprócz wspomnianych pięciu wiązań — występuje dodatkowe wiązanie. Jest ono utworzone wskutek silnego przenikania się orbitali atomowych $2p_z$ atomów węgla. To dodatkowe wiązanie opisane jest wiążącym orbitalem molekularnym utworzonym przez kombinację liniową orbitali atomowych $2p_z$ (por. rys. 29) zlokalizowanych na atomach węgla cząsteczki (jest to wiązanie typu π). W cząsteczce etylenu występuje więc wiązanie podwójne pomiędzy atomami węgla — jedno wiązanie σ i jedno wiązanie π. Własności przestrzenne obu wiązań są oczywiście odmienne. Oprócz omówionej powyżej hybrydyzacji typu sp^2 bardzo użyteczne w chemii są hybrydyzacje typu sp^3 oraz typu sp. Nie będziemy jednak ich tu omawiali.



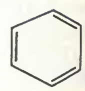
Rys. 41. Układ orbitali molekularnych cząsteczki etylenu. Strzałkami zaznaczono obsadzenie orbitali przez elektrony walencyjne (konfiguracja stanu podstawowego)

Z rozważań energetycznych wynika, że orbitale molekularne opisujące wiązania σ w cząsteczce mają niższą energię aniżeli orbitale molekularne typu π. Każdy z tych orbitali molekularnych w stanie podstawowym cząsteczki jest zajęty przez dwa elektrony o przeciwnie skierowanych spinach (por. rys. 41). Nad orbitalami wiążącymi znajdują się orbitale molekularne antywiążące, nieobsadzone w stanie podstawowym przez elektrony.

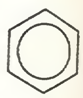
W podobny sposób, w jaki rozpatrywaliśmy tworzenie orbitali molekularnych w etylenie, można rozważać orbitale molekularne cząsteczek bardziej złożonych, jak benzenu, naftalenu itp. Komplikacje wiążą

cząsteczka benzenu

się wyłącznie ze zwiększeniem się liczby elektronów walencyjnych. I tak np. w benzenie (30 elektronów walencyjnych) mamy sześć atomów węgla o hybrydyzacji sp^2 i sześć atomów wodoru. Każdy atom węgla tworzy jedno wiązanie σ_{C-H} z atomem wodoru oraz dwa wiązania σ_{C-C} z sąsiednimi atomami węgla. Mamy więc w benzenie 6 wiązań σ_{C-H} i 6 wiązań σ_{C-C}. Każdy z atomów węgla ma po jednym niezhybrydyzowanym orbitalu atomowym typu $2p_z$. Z orbitali tych możemy utworzyć więc wiązania typu π. W jaki sposób? Czy w taki sam sposób jak w etylenie? W cząsteczce etylenu dwa orbitale $2p_z$ tworzą wiązanie typu π i wspólnie z wiązaniem σ_{C-C} w cząsteczce zostaje utworzone wiązanie podwójne. Fakt ten zaznaczony jest we wzorze strukturalnym etylenu przez podwójną kreskę pomiędzy atomami węgla $\mathrm{H_2C{=}CH_2}$. W benzenie powinniśmy w zasadzie wyróżnić trzy wiązania — benzen powinien mieć trzy wiązania podwójne (struktura a). Struktura taka jest jednak sprzeczna z własnościami benzenu. Na przykład wszystkie długości wiązań pomiędzy atomami węgla są jednakowe, cząsteczka ma wysoką symetrię D_{6h}. Fakty te związane są z własnościami elektronów opisanych orbitalami atomowymi typu $2p_z$. Okazuje się, że są one bardzo „ruchliwe”, w przeciwieństwie do elektronów tworzących wiązania σ nie ograniczają swojej obecności do obszaru, gdzie tworzą wiązania. Nie można po prostu odróżnić wiązań π (struktura b). Orbitale atomowe $2p_z$ poprzez liniowe kombinacje tworzą orbitale molekularne. W cząsteczce benzenu sześć orbitali atomowych $2p_z$ tworzy sześć liniowo niezależnych orbitali molekularnych: trzy z nich są wiążącymi orbitalami, a trzy orbitalami antywiążącymi.



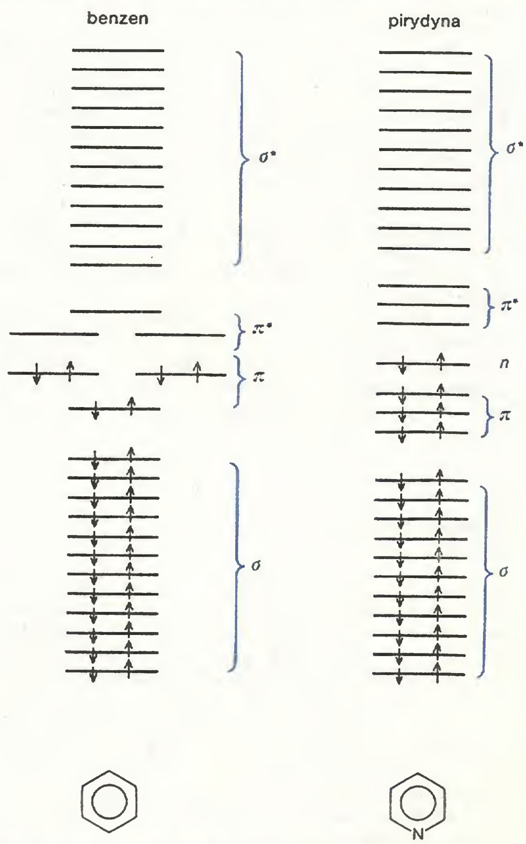
a



b

elektrony typu n

W cząsteczkach organicznych zawierających tlen lub azot, oprócz elektronów σ i π rozróżnia się jeszcze



Rys. 42. Układ orbitali molekularnych w cząsteczkach benzenu i pirydyny. Strzałkami zaznaczono obsadzenie orbitali molekularnych przez elektrony walencyjne atomów (konfiguracja stanu podstawowego)

cząsteczka pirydyny

jeden rodzaj elektronów — elektrony typu *n*. Własności tych ostatnich elektronów omówimy na przykładzie cząsteczki pirydyny (rys. 42). Swobodny atom azotu w stanie podstawowym ma konfigurację elektronową $1s^2 2s^2 2p^3$ — ma więc on pięć elektronów walencyjnych. W pierścieniu pirydynowym orbitale atomowe azotu mają, podobnie jak orbitale węgla, hybrydyzację sp^2. Dwa elektrony walencyjne atomu azotu opisane hybrydami sp^2 tworzą wiązanie σ_{C-N} z elektronami sp^2 sąsiednich atomów węgla. Trzeci orbital sp^2 azotu nie tworzy żadnego wiązania, gdyż w jego sąsiedztwie nie ma innego atomu ani grupy. Orbital ten nazywamy orbitalem niewiążącym (od ang. *nonbonding*) lub po prostu orbitalem *n*. Na orbitalu tym znajduje się para elektronów *n* mających przeciwnie skierowane spiny. Warto podkreślić, że orbital niewiążący opisujący elektrony *n* jest symetryczny względem odbicia w płaszczyźnie cząsteczki (jest to zrozumiałe, gdyż orbital *n* jest po prostu orbitalem zhybrydyzowanym sp^2). Ostatni, piąty elektron walencyjny azotu, jest opisany orbitalem $2p_z$ i jest po prostu elektronem π. Elektron ten wchodzi w sprzężenie z pięcioma elektronami π pochodzącymi od atomów węgla cząsteczki. Mamy w ten sposób utworzony, podobnie jak w benzenie, układ sześciu elektronów π. I zupełnie w ten sam sposób jak w benzenie z sześciu orbitali atomowych $2p_z$ możemy utworzyć sześć niezależnych orbitali molekularnych π, z których trzy najniższe energetycznie będą obsadzone przez sześć elektronów w stanie podstawowym cząsteczki pirydyny. Na uwagę zasługują pewne różnice pomiędzy orbitalami molekularnymi benzenu i pirydyny. W benzenie, mającym wysoką symetrię, współczynniki rozwinięcia orbitali molekularnych na orbitale atomowe $2p_z$ można łatwo wyznaczyć, korzystając z własności symetrii układu. Tak więc postać takich przybliżonych orbitali molekularnych benzenu jest wyznaczona przez topologię układu. Natomiast dla pirydyny, wykorzystując jedynie własności symetrii cząsteczki nie można wyznaczyć postaci orbitali molekularnych. Dalsza różnica polega na tym, że w benzenie występują orbitale molekularne π zdegenerowane, natomiast w pirydynie nie występuje degeneracja orbitali π.

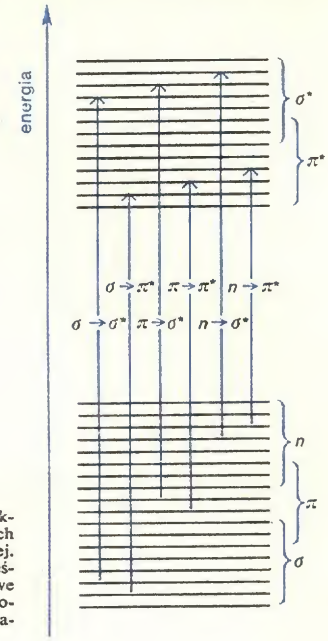
orbital niewiążący (orbital *n*)

Pomiędzy układem wiążących orbitali molekularnych π pirydyny a układem antywiążących orbitali π^* znajduje się orbital niewiążący *n* zajęty w stanie podstawowym cząsteczki przez dwa elektrony. Tak wysokie energetycznie położenie orbitalu *n* związane jest z tym, że nie tworzy on żadnego wiązania.

Posługując się orbitalami zhybrydyzowanymi można bardzo łatwo określić, ile elektronów typu σ i π ma dana cząsteczka, a tym samym można określić, ile orbitali molekularnych typu σ i π można zbudować dla niej. Hybrydyzacja orbitali atomowych jest tylko procedurą matematyczną i chociaż jest ona bardzo użyteczna przy opisie jakościowym wiązań w cząsteczce, nie jest wcale konieczna. Zgodnie bowiem ze schematem przybliżenia jednoelektronowego, każdy elektron walencyjny cząsteczki opisany jest orbitalem molekularnym, który może być przybliżony przez LCAO. Jako rozwinięcia MO na LCAO można użyć orbitali atomowych walencyjnych atomów tworzących cząsteczkę (taką bazę używa się w obliczeniach półempirycznych), ale baza może być oczywiście większa. W tym sformułowaniu orbital molekularny opisujący elektron σ lub π jest rozciągnięty po całej cząsteczce, i nie charakteryzuje określonego wiązania pomiędzy dwoma atomami cząsteczki.

Niezależnie od rodzaju przybliżenia użytego do opisu orbitali molekularnych wszystkie MO cząsteczki organicznej (jeśli jest ona płaska) dzielimy na dwa typy — orbitale σ (tutaj wchodzą też orbitale *n*, które mają symetrię σ) i orbitale π. Orbitale zajęte przez elektrony w konfiguracji podstawowej nazywamy orbitalami wiążącymi, natomiast orbitale nie zajęte przez elektrony w tej konfiguracji nazywamy orbitalami antywiążącymi. Względne rozmieszczenie energetyczne orbitali molekularnych σ, π oraz *n* w cząsteczce pokazane jest na rys. 43. Obliczenia półempiryczne typu EHT lub AVE, jak również obliczenia *ab initio* pokazują, że w złożonych cząsteczkach organicznych niektóre orbitale molekularne σ znajdują się pomię-

układ orbitali molekularnych



Rys. 43. Schematyczny układ orbitali molekularnych w cząsteczce organicznej. Strzałki oznaczają określone przejścia elektronowe związane z absorpcją promieniowania elektromagnetycznego

dzy orbitalami π — w takich układach nie ma więc ścisłego podziału obszarów energii orbitalnych na obszar energii orbitali σ, π oraz *n*. W ogólności jednak orbitale σ mają niższe energie niż orbitale π, a te z kolei niższe niż orbitale *n*. Schemat energetyczny orbitali molekularnych przedstawiony na rys. 43 tłumaczy niektóre fakty z elektronowej spektroskopii molekularnej (→ Spektroskopia molekularna — Widma elektronowe cząsteczek). Tak więc promieniowanie elektromagnetyczne padające na cząsteczkę powoduje różnego typu przejścia elektronowe, mające swoje charakterystyczne własności. Zadaniem spektroskopii doświadczalnej, jak również teoretycznej jest m.in. określenie natury stanów wzbudzonych (tzn. rodzaju przejścia: przejście $\pi \rightarrow \pi^*$ czy $n \rightarrow \pi^*$ itd.) cząsteczek organicznych (zob. rys. 15).

przejścia $\pi \rightarrow \pi^*$ $n \rightarrow \pi^*$

Własności fizykochemiczne cząsteczki

Korzystając z określonej metody półempirycznej lub *ab initio* można dla danej cząsteczki wyznaczyć jej orbitale molekularne, a tym samym można wyznaczyć przybliżoną funkcję falową opisującą konfigurację stanu podstawowego cząsteczki. Za pomocą tej funk-

Kąty między wiązaniami w różnych cząsteczkach

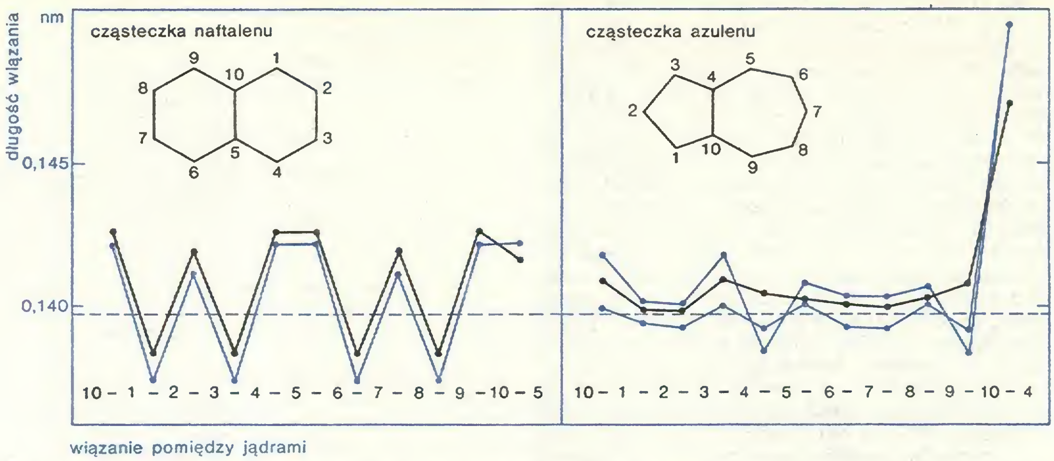
Cząsteczka	Kąt	Wartość, stopnie	
		doświadczalna	teoretyczna (metoda CNDO/2)
NH_3	H—N—H	103	107
H_2O	H—O—H	104	107
CO_2	O—C—O	180	180
NF_3	F—N—F	104	103
CH_3	H—C—H	120	120
NH_3	H—N—H	107	107
BF_3	F—B—F	120	120
H_2CO	H—C—H	117	115
F_2CO	F—C—F	108	109
C_2H_4	H—C—C	122	123

Wg G. Klopman, B. O'Leary, Topics Curr. Chem. 15, 445 (1970); B. Roos, I. Fischer-Hjalmars, Kemisk Tidskrift, No. 12, 10 (1970).

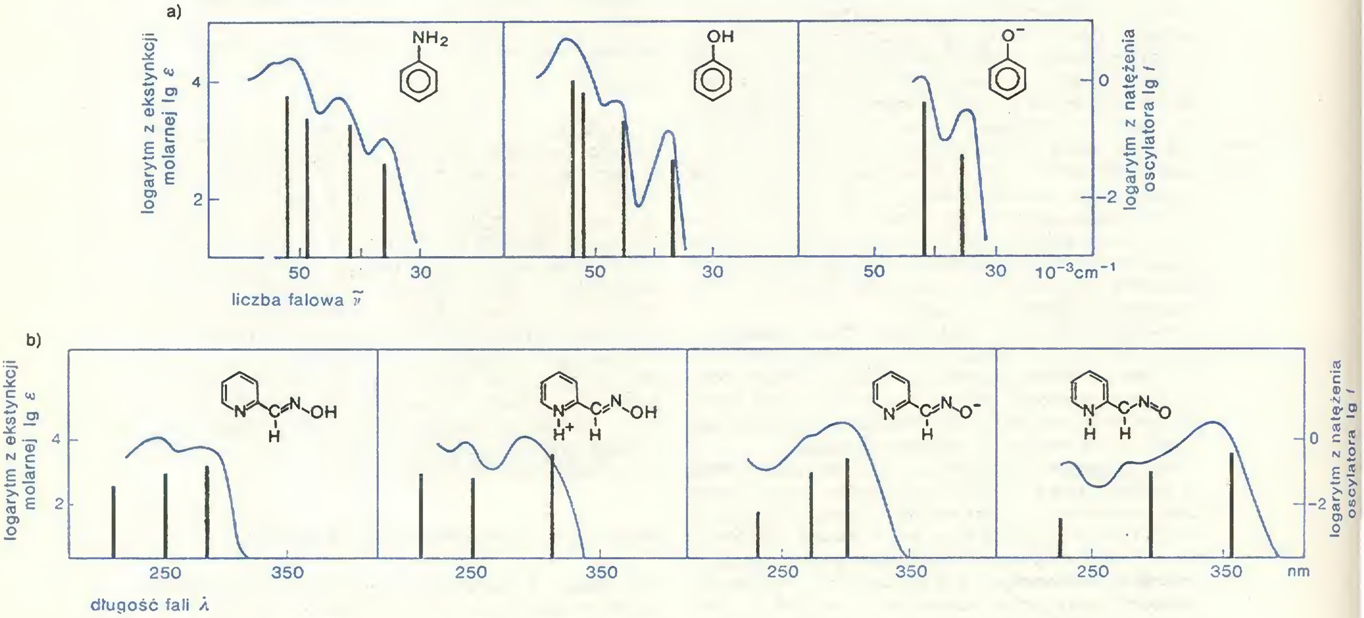
cji falowej możemy pokusić się o interpretację wielu własności fizykochemicznych cząsteczki.

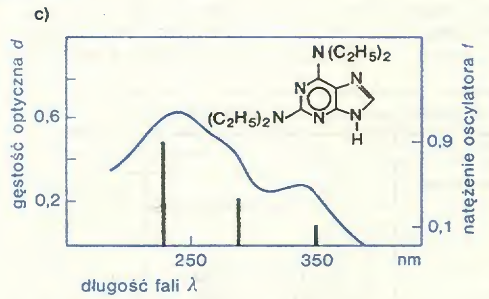
W zależności od geometrii układu zawierającego elektrony π, delokalizacja tych ostatnich w cząsteczce zachodzi w różny sposób. Długości wiązań w cząsteczkach organicznych są więc zależne od tej delokalizacji. Rysunek 44 podaje np. porównanie pomiędzy doświadczalnymi i teoretycznymi długościami wiązań w naftalenie i azulenie. Z rysunku tego widać również, że pomimo tego, iż obie cząsteczki mają po 10 elektronów π, to różnice w topologii cząsteczek mają istotne znaczenie dla długości wiązań w cząsteczkach.

Metody półempiryczne opisujące własności elektronów π cząsteczki mogą być użyte do interpretacji geometrii, tj. długości wiązań, nie dla całej cząsteczki, ale tylko tych ich części, w których występuje delokalizacja elektronów π. Nie można więc na podstawie tych metod obliczyć np. długości wiązań C—H w cząsteczkach. Można to uczynić natomiast na podstawie metod półempirycznych typu AVE. Metody te (a szczególnie metoda CNDO/2 lub MINDO/2) w wielu wypadkach interpretują poprawnie nie tylko wartości długości wiązań, ale również i wartości kątów w różnego typu cząsteczkach.



Rys. 44. Długości wiązań w cząsteczce naftalenu i azulenu. Linie niebieskie — wartości doświadczalne, linie czarne — wartości obliczone metodą PPP; linia pozioma odpowiada wartości 0,1397 nm (długość wiązania C—C w cząsteczce benzenu). W różnych pomiarach otrzymano różne wartości długości wiązań w azulenie (dwie linie niebieskie); wg I. Fischer-Hjalmars, B. Roos, Kemisk Tidskrift, No. 11, 28 (1970)

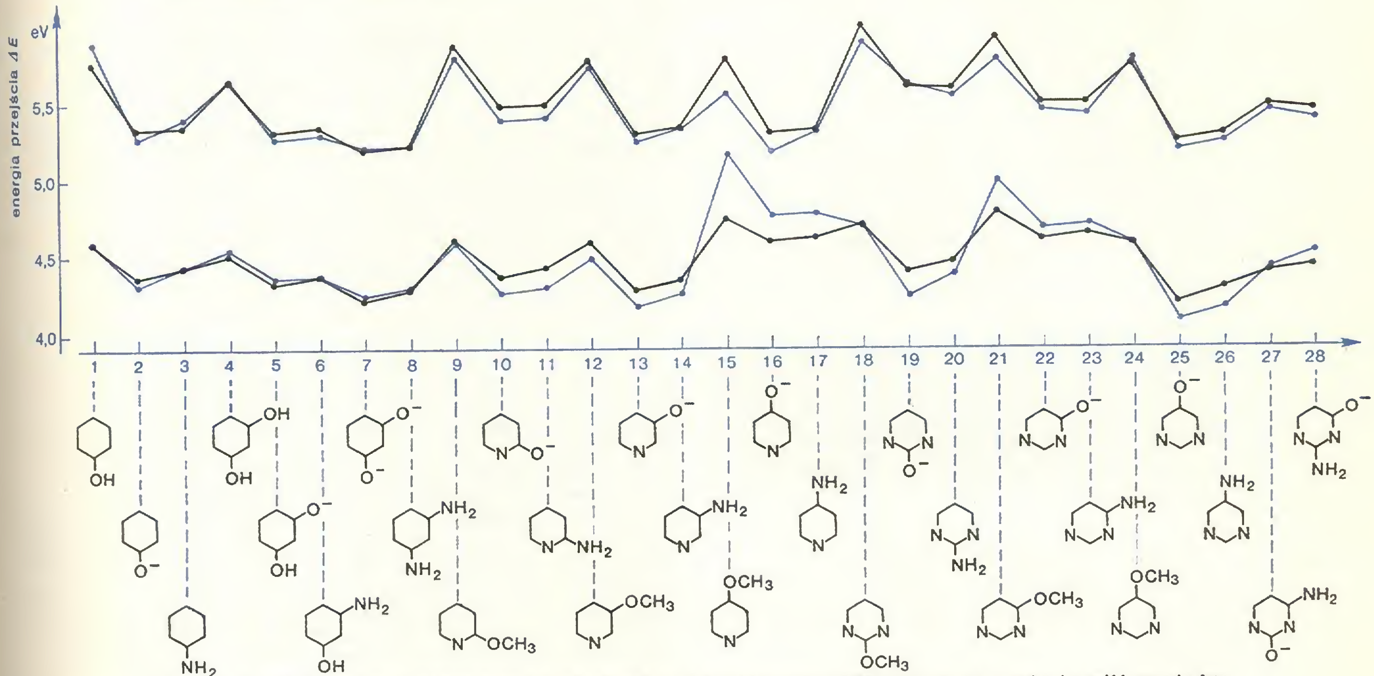




Rys. 45. Widma absorpcyjne w zakresie nadfioletu z kilku cząsteczek organicznych. Obliczone metodą PPP położenia maksimów pasm absorpcyjnych oznaczone są liniami pionowymi (wysokość linii jest proporcjonalna do obliczonego natężenia oscylatora f lub do $\lg f$); związek między ekstynkcją molarną ε i natężeniem f jest postaci: $\lg \varepsilon = 4 - \lg f$; a) anilina, fenol i anion fenolu; wg M. Tichý, R. Zahradník, J. Phys. Chem. 73, 534 (1969); b) forma laktimowa pirydyno-2-aldoksymu, jego kation i anion oraz forma laktamowa związku; wg J.S. Kwiatkowski, W. Herzog (Boniewicz), B. Szczodrowska, J. Molec. Struct., 26, 333 (1975); c) 2,6-dwu(bisetyloamino)puryna; wg J.S. Kwiatkowski, Bull. Inst. Phys. Res. Atom. (Bulgaria) 21, 327 (1971)

Metody półempiryczne interpretują poprawnie także wartości innych wielkości fizykochemicznych cząsteczek, jak np. momentów dipolowych, potencjałów jonizacyjnych, ciepła tworzenia cząsteczek jak również wartości całego szeregu wielkości charakteryzujących widma tak jądrowego, jak i elektronowego rezonansu paramagnetycznego.

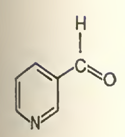
Gdy chcemy zinterpretować elektronowe widma absorpcyjne w zakresie nadfioletu szczególnie użyteczna okazuje się metoda PPP. Rysunek 45 podaje krzywe absorpcyjne kilku związków organicznych wraz z położeniami przewidywanych przez teorię położeń pasm absorpcyjnych. Teoria nie tylko interpretuje poprawnie położenia pasm absorpcyjnych cząsteczek, ale również położenia pasm absorpcyjnych ich anionów i kationów oraz różnych form tautomerycznych cząsteczek. Metoda PPP jest szczególnie użyteczna przy interpretacji przesunięć pasm absorpcyjnych w serii podobnych cząsteczek (rys. 46).



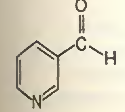
Rys. 46. Korelacja pomiędzy obserwowanymi (linie niebieskie) i obliczonymi metodą PPP (linie czarne) energiami przejść typu singlet--singlet serii cząsteczek organicznych; wg M. Berndt, J. S. Kwiatkowski, Theor. Chim. Acta 17, 35 (1970)

badanie konformacji

Metody chemii kwantowej stosowane są w wielu wypadkach do badania konformacji cząsteczek. Konformację cząsteczki można badać bezpośrednio obliczając zmianę energii całkowitej cząsteczki wraz ze zmianą geometrii części cząsteczki (np. przy obrocie określonej grupy wokół wiązania) lub też w sposób pośredni przez obliczenie wielu własności cząsteczki dla jej różnych konformacji i porównanie tych własności z własnościami doświadczalnymi. Na przykład dane doświadczalne pokazują, że cząsteczka 3-formylopirydyny występuje w roztworze jako mieszanina dwóch form cis (Ia) i trans (Ib), a mierzony moment dipolowy cząsteczki (ściślej biorąc zmierzony moment dipolowy mieszaniny obu form) wynosi 2,37 D. Momenty dipolowe obliczone dla formy (Ia) i (Ib) wynoszą odpowiednio 1,13 D i 3,90 D. Przyjmując, że w roztworze występuje mieszanina obu form w proporcji ok. 40% formy (Ia) i ok. 60% formy (Ib), otrzymuje się obliczony moment dipolowy dla mieszaniny w zgodności z wartością doświadczalną. Obliczona w ten sposób zawartość procentowa obu form w mieszaninie jest w bardzo dobrej zgodności z wartością wyznaczoną doświadczalnie za pomocą pomiarów jądrowego rezonansu magnetycznego (30% formy Ia, 70% formy Ib).



Ia



Ib

Biochemia kwantowa

Rozpatrywanie podstawowych procesów biologicznych jako procesów zachodzących na poziomie molekularnym lub submolekularnym pociągnęło za sobą rozwój teorii tych zjawisk — rozwój biologii molekularnej. W ostatnich kilkunastu latach zaczęto coraz częściej stosować metody mechaniki kwantowej do zagadnień biologicznych, co spowodowało nawet wydzielenie odrębnego działu nazywanego biologią kwantową (lub w węższym znaczeniu biochemią kwantową, obejmującą badanie metodami chemii kwantowej własności cząsteczek o znaczeniu biochemicznym). Ten nowy dział obejmuje badania teoretyczne nad własnościami wielu leków (farmakologia kwantowa), modelowych białek i składników kwasów nukleinowych (biochemia kwantowa kwasów nukleinowych). Badania z dziedziny biochemii kwantowej prowadzone są w wielu laboratoriach na całym świecie, a niektóre laboratoria zajmują się wyłącznie tymi badaniami.

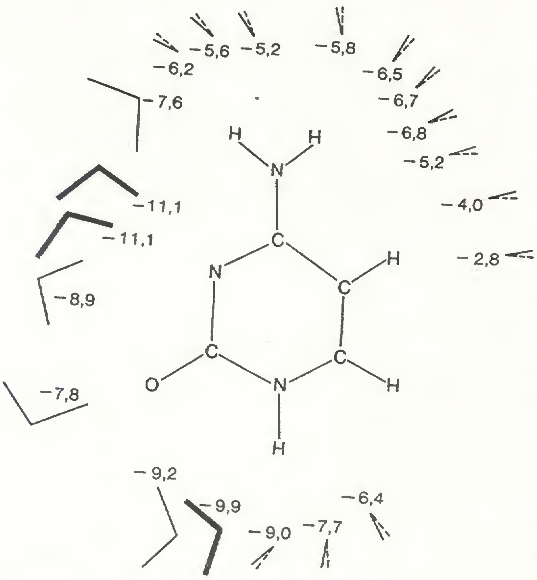
badanie kwasów nukleinowych

Wyniki zastosowań metod chemii kwantowej do opisu struktury elektronowej zasad kwasów nukleinowych, ich licznych pochodnych i oddziaływań między nimi są dobrą ilustracją szczególnie żywej współpracy fizyków i chemików teoretyków z biologami. Kwasy nukleinowe są odpowiedzialne za przekazywanie informacji genetycznej (→ Kwasy nukleinowe). Zmiany informacji genetycznej są przyczyną powstawania mutacji, które mogą być spowodowane np. niewielką zmianą konfiguracji atomów w kwasie dezoksyrybonukleinowym (DNA); są to więc zjawiska, którymi rządzą prawa mechaniki kwantowej. Nic więc dziwnego, że próbowano i próbuje się nadal rozwiązać problemy teoretyczne związane z budową oraz funkcją kwasów nukleinowych.

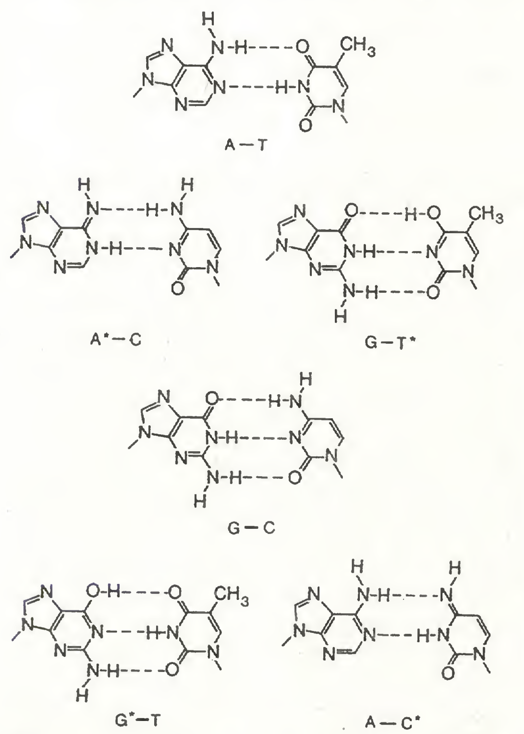
W celu zbadania struktury elektronowej podstawowych składników kwasów nukleinowych (cytozyny, uracylu, tyminy, guaniny, adeniny, jak również wielu ich pochodnych) użyto wszystkich znanych metod obliczeniowych chemii kwantowej (tak półempirycznych jak i nieempirycznych). Pozwoliło to wyjaśnić wiele problemów związanych z własnościami stanu podstawowego tych cząsteczek. Na przykład, na długo przed

moment dipolowy cytozyny i guaniny

zmierzeniem momentów dipolowych i potencjałów jonizacyjnych zasad kwasów nukleinowych wykonano obliczenia wartości tych wielkości, a późniejsze pomiary potwierdziły obliczone wartości. Jednak momenty dipolowe cytozyny i guaniny nie zostały dotąd zmierzone, a informacje o nich można uzyskać tylko z obliczeń. Ma to istotne znaczenie dla zrozumienia oddziaływań pomiędzy zasadami w łańcuchu DNA.



Rys. 47. Energia oddziaływania (w kcal/mol) między cząsteczką cytozyny i cząsteczką wody przy różnych wzajemnych ustawieniach. Linie grube oznaczają miejsca najbardziej prawdopodobne (największa energia oddziaływania) na przyłączenie cząsteczki wody do cytozyny. Linie ciągłe oznaczają ustawienie obu cząsteczek w tej samej płaszczyźnie, linie przerywane oznaczają ustawienie cząsteczki wody prostopadle do płaszczyzny cytozyny; wg G.N.J. Port, A. Pullman, FEBS Letters 31, 70 (1973)



Rys. 48. Komplementarne pary adenina-tymina (A-T) oraz guanina-cytozyna (G-C) w DNA; podano też pary nieprawidłowe, powstające gdy jedna z zasad występuje w rzadkiej formie tautomerycznej (oznaczona gwiazdką)

Obliczenia kwantowe dla stanu podstawowego zasad DNA pozwoliły wyjaśnić wiele problemów związanych z reaktywnością chemiczną tych cząsteczek jak również przewidzieć najbardziej prawdopodobne miejsca w zasadach na przyłączenie cząsteczki wody (rys. 47).

tautomeria

Obliczenia kwantowe dla zasad kwasów nukleinowych są bardzo ważne w związku z obserwowaną w zasadach tautomerią. W proponowanej przez Watsona i Cricka strukturze DNA zasady połączone w komplementarnych parach występują w ściśle określonych formach tautomerycznych (formy aminowe i laktamowe). Jeśli jedna z zasad wystąpi w rzadkiej formie tautomerycznej, następuje wówczas zmiana własności kompleksowania i w łańcuchu DNA zostanie wbudowany niewłaściwy nukleotyd (powstaje wówczas błąd w tworzeniu pary zasad a w konsekwencji organizm syntetyzować może niewłaściwe białka, powodując „chorobę" organizmu). Rysunek 48 przedstawia na przykładzie par adenina-tymina oraz guanina-cytozyna powiązanie zmian w tautomerii ze zmianami w tworzeniu komplementarnych par zasad. Prawdopodobieństwo występowania określonej zasady DNA w rzadkiej formie tautomerycznej jest bardzo małe. Zwiększa się ono w istotny sposób pod wpływem działania wielu czynników, np. promieniowania elektromagnetycznego, czynników jonizujących czy też wskutek działania mutagenów chemicznych na zasady. W ostatnim wypadku, w wyniku działania hydroksyloaminy lub hydrazyny na cytozynę, powstają między innymi pochodne cytozyny, dla których prawdopodobieństwo występowania w rzadkiej formie jest o wiele większe niż w wypadku niepodstawionej cytozyny. Fakt ten tłumaczy występowanie licznych mutacji pod wpływem działania tych związków na DNA.

Znając własności poszczególnych zasad można zająć się zagadnieniem własności par zasad, znaczeniem wiązań wodorowych, za pomocą których połączone są pary zasad, zagadnieniem stabilności podwójnego heliksu DNA i wreszcie strukturą kwasów nukleinowych. Wykonano np. obliczenia oddziaływań pomiędzy parami zasad przy zmiennej orientacji zasad. Pokazano, że w wypadku pary guanina-cytozyna najbardziej prawdopodobne ustawienie zasad jest takie, jakie jest obserwowane w podwójnym heliksie DNA. Natomiast dla pary adenina-uracyl oprócz ustawienia obserwowanego w spirali DNA możliwe było inne ustawienie wzajemne zasad, co znalazło potem potwierdzenie w kryształach pochodnych adeniny i uracylu.

wiązania wodorowe

Wiele prac z dziedziny chemii kwantowej zostało poświęconych badaniu własności szczególnego typu wiązań zwanych wiązaniami wodorowymi, za pomocą których połączone są komplementarne pary zasad w DNA (rys. 48). Wiązanie wodorowe X—H···Y powstaje wówczas, kiedy atom wodoru H połączony jest z silnie elektroujemnym atomem X (np. atomem azotu lub tlenu), natomiast atom Y ma tzw. wolną parę elektronów (np. wiązania wodorowe N—H···O lub N—H···N pomiędzy zasadami DNA na rys. 48). Wiązania wodorowe są kilkadziesiąt razy słabsze od przeciętnych wiązań chemicznych (np. energia wiązania wodorowego dla dwóch cząsteczek wody wynosi ok. 5 kcal/mol $\approx$ 21 kJ/mol), niemniej są one jednym z czynników decydujących o komplementarności zasad i mają specyficzny wpływ na trwałość pary. Poprzez wiązanie wodorowe zachodzi częściowa delokalizacja elektronów π obu zasad; jest to między innymi przyczyną większej stabilności pary w stosunku do układu izolowanych dwóch zasad. Oprócz prac pogłębiających wiadomości dotyczące oddziaływań między poszczególnymi zasadami prowadzone są także badania większych zespołów zasad, takich jak polinukleotydy. Jednakże wyciąganie ogólnych wniosków dotyczących własności DNA z tego typu badań jest jeszcze przedwczesne.

Zrozumienie roli i funkcji jaką pełnią kwasy nukleinowe w żywym organizmie powoduje nadal ogromne

zainteresowanie strukturą tych kwasów. Nic więc dziwnego, że biologia kwantowa rozwija się bardzo intensywnie, i należy spodziewać się, że będzie rozwijała się w najbliższej przyszłości. Być może badania kwantowe nad strukturą kwasów nukleinowych i białkami pomogą wyjaśnić procesy starzenia się organizmu, czy też wyjaśnią przyczyny powstawania chorób. Może powstanie więc „kwantowa teoria raka", a może powstanie „medycyna kwantowa", która nie tylko znajdzie skuteczne lekarstwo przeciw rakowi, ale również przeciw innym chorobom.

Podany w niniejszym artykule przegląd niektórych zastosowań mechaniki kwantowej w chemii z konieczności jest powierzchowny i niepełny. Zostały pominięte tutaj zastosowania teorii kwantowych do interpretacji budowy związków kompleksowych. Nie omówiono szczegółowo zastosowania metod kwantowych do badania własności kwasów nukleinowych, pominięto zastosowanie tych metod do badania białek i witamin oraz zastosowanie w farmakologii. Nie wspomniano zupełnie o teorii reakcjił fotochemicznych itp. zagadnieniach.

Autor uzna jednak, że artykuł ten spełni swoje zadanie, jeśli Czytelnik wyrobi sobie nowy pogląd na budowę cząsteczek i nowy sposób myślenia o układach chemicznych.

C.A. Coulson *Wiązania chemiczne*, Warszawa 1963; A. Gołębiewski *Chemia kwantowa związków nieorganicznych*, Warszawa 1969; A. Gołębiewski *Chemia kwantowa związków organicznych*, Warszawa 1973; W. Kołos *Kwantowe teorie w chemii i biologii*, Wrocław 1971; W. Kołos *Chemia kwantowa*, Warszawa 1978; J. S. Kwiatkowski, B. Pullman, Adv. Heterocycl. Chem. 18, 200 (1975); A. Pullman Topics Curr. Chem. 31, 45 (1972); B. Pullman, A. Pullman *Quantum Biochemistry*, New York 1963; A. J. Sadlej *Elementarne metody chemii kwantowej*, Warszawa 1966.

SPEKTROSKOPIA

Oddziaływanie promieniowania elektromagnetycznego lub korpuskularnego z materią we wszystkich jej makro- i mikroskopowych formach jest podstawą całej spektroskopii, niezależnie od tego czy jej przedmiotem badań jest struktura izolowanego elementu mikro- lub makroukładu, czy też zjawiska i procesy zachodzące w tych układach. Najważniejszą konsekwencją takiego oddziaływania jest fakt, że energia jest pochłaniana lub wysyłana przez materię w pewnych dyskretnych, skończonych porcjach — kwantach energii. Badając częstość lub długość fali pochłanianego czy wysyłanego promieniowania możemy wyznaczyć wielkość zmian energii w procesie oddziaływania promieniowania z materią, a tym samym określić dyskretne poziomy energetyczne badanego układu. Każdy układ materialny ma charakterystyczne dla siebie poziomy energetyczne, co umożliwia podział spektroskopii na podstawowe działy: spektroskopię jądrową, której przedmiotem badań są poziomy energetyczne i własności jąder atomowych; spektroskopię atomową, historycznie najstarszy dział spektroskopii, zajmujący się badaniem poziomów energetycznych atomów; spektroskopię molekularną, badającą poziomy energetyczne i strukturę cząsteczek oraz oddziaływania międzycząsteczkowe; spektroskopię kryształów, której przedmiotem badań jest struktura energetyczna kryształów oraz czynniki określające i wpływające na tę strukturę.

Spektroskopia wraz z jej zastosowaniami stanowi jeden z najbardziej podstawowych działów fizyki, chemii i innych dyscyplin pokrewnych. Poniżej omówione zostaną podstawy i zastosowania wybranych działów spektroskopii oraz metody badań spektroskopowych.

Spektroskopia atomowa

Tadeusz Skaliński

Promieniowanie optyczne wysyłane przez pary atomowe wzbudzone do świecenia przez wyładowanie elektryczne, przez ogrzanie do wysokiej temperautry lub optycznie zawiera w sobie ogromną ilość informacji o atomach, które to promieniowanie wysłały. Aby tę informację wyzyskać, winniśmy promieniowanie rozłożyć w widmo, tj. przeanalizować, jak przebiega w nim zależność natężenia od częstości drgań. W odróżnieniu od promieniowania o widmie ciągłym, wysyłanego przez rozżarzone ciała stałe, promieniowanie par atomowych ma widmo liniowe.

Badanie tych widm ujawniło liczne prawidłowości, których interpretacja pozwoliła na wyciągnięcie daleko idących wniosków o strukturze atomu. Ten zakres zagadnień jest przedmiotem badań spektroskopii atomowej.

Atom wodoropodobny i jego widmo

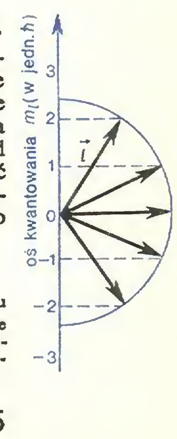
atom wodoropodobny w mechanice kwantowej

Zagadnienie wyznaczenia energii stanów stacjonarnych atomu o jednym elektronie i ładunku jądra $+Ze$ może być rozwiązane w sposób konsekwentny metodami relatywistycznej mechaniki kwantowej (równanie Diraca). Ujęcie nierelatywistyczne, w którym punktem wyjścia jest równanie Schrödingera, prowadzi do wyznaczenia funkcji własnych i wartości własnych operatora energii (→ Chemia kwantowa). Energia stanów stacjonarnych

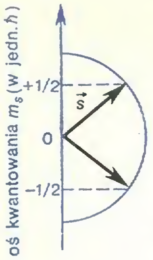
$$W_n = -\frac{2\pi^2 e^4 m Z^2}{h^2 n^2}, \qquad (1)$$

gdzie e, m — oznaczają ładunek i masę elektronu, $h = 6{,}626176 \cdot 10^{-34}\,\mathrm{J \cdot s}$ — stałą Plancka; $n = 1, 2, 3...$ — główną liczbę kwantową.

W nierelatywistycznej mechanice kwantowej wprowadza się jeszcze dwie liczby kwantowe: l — orbitalnego momentu pędu elektronu $\hbar\vec{l}$ ($|\vec{l}| = \sqrt{l(l+1)} \equiv l^*$, $\hbar = h/2\pi$) oraz m_l — rzutu $\vec{l}$ na kierunek kwantowania; rzut ten jest równy $\hbar m_l$ (rys. 1). Liczby kwantowe n, l, m_l (wszystkie całkowite) spełniają następujące warunki: $0 \leqslant l \leqslant n-1$ i $|m_l| \leqslant l$. Z wyrażenia na W_n widać, że w tym ujęciu energia stanu nie zależy od l (mówimy o zwyrodnieniu względem l). Zależność energii stanu od orientacji momentu orbitalnego ujawnia się dopiero po przyłożeniu pola zewnętrznego (zob. str. 288).



Rys. 1. Skwantowane ustawienie wektora $\vec{l}$ względem kierunku wyróżnionego; $l = 2$; $l^* = \sqrt{6}$ (promień okręgu). Zaznaczone położenia odpowiadają wartościom $m_l = 2, 1, 0, -1, -2$. Kierunek składowej prostopadłej do osi kwantowania nie jest określony — wszystkie są jednakowo prawdopodobne



Rys. 2. Skwantowane ustawienia wektora $\vec{s}$ względem kierunku wyróżnionego; $s = {}^1/_2$; $s^* = \sqrt{3}/2$ (promień okręgu). Te dwie możliwości ustawienia spinu względem kierunku wyróżnionego zostały potwierdzone w doświadczeniu nad odchylaniem atomów srebra biegnących w poprzek niejednorodnego pola magnetycznego

W okresie powstawania mechaniki kwantowej analiza struktury widm (głównie metali alkalicznych) doprowadziła Uhlenbecka i Goudsmitha do wniosku, że każdy elektron jest obdarzony własnym, wewnętrznym momentem pędu — spinem. Jest on stały i jego liczba kwantowa $s = {}^1/_2$. Podobnie jak w wypadku momentu orbitalnego spin elektronu wynosi $\vec{s}\hbar$, przy czym $|\vec{s}| = \sqrt{s(s+1)} \equiv s^* = {}^1/_2\sqrt{3}$. Rzut spinu na oś kwantowania jest równy $m_s\hbar = \pm{}^1/_2\hbar$ (rys. 2). Spin sumuje się z orbitalnym momentem pędu, tworząc całkowity moment pędu elektronu (jest to specyficzny sposób sumowania, zarazem wektorowy i spełniający reguły kwantowe). Podkreślić tu należy, że spin elektronu nie wynika z rozważań nierelatywistycznej mechaniki kwantowej i w tym schemacie jest wielkością dołączoną fenomenologicznie. To dołączenie znajduje swój wyraz w dodaniu do zespołu liczb kwantowych opisujących stan elektronu liczby kwantowej $m_s = \pm{}^1/_2$ oraz w przedstawieniu funkcji stanu elektronu w postaci $\Psi = \psi(n, l, m_l)\cdot\alpha(m_s)$. Spin elektronu (ogólniej — spin każdej cząstki elementarnej) jest wielkością fizyczną charakterystyczną dla świata atomowego i nie ma analogii w fizyce klasycznej.

spin elektronu

Widmo wodoru

Podstawowa zależność fizyki kwantowej wiążąca różnicę energii stanów układu fizycznego ΔW z częstością wysłanego (lub pochłoniętego) promieniowania ma postać:

$$\Delta W = h\nu = h\tilde{\nu}c,$$

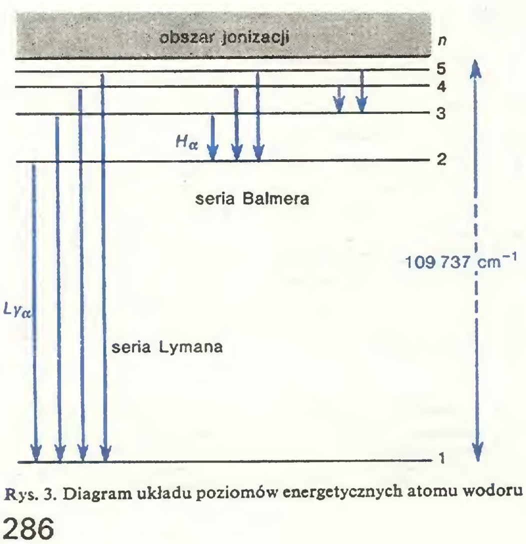
przy czym ν jest częstością promieniowania, $\tilde{\nu} = \nu/c$ — liczbą falową (w cm^{-1}). Jeśli wyrażenie $2\pi^2 e^4 m/h^3 c$ oznaczymy przez R, to energia $W_n = -RZ^2hc/n^2$. R jest bardzo ważną stałą fizyczną; jest to tzw. stała Rydberga, ma ona wartość 109737,31 cm^{-1}. Wyrażona w skali liczb falowych różnica energii między stanem początkowym i i końcowym f wynosi zatem:

stała Rydberga

$$\tilde{\nu}_{if} = -RZ^2\left(\frac{1}{n_i^2} - \frac{1}{n_f^2}\right).$$

Gdy $Z = 1$, przy $n_f = 1$ (stan podstawowy) i $n_i = 2, 3, 4, \dots$ (n może przy przejściu zmieniać się o dowolną liczbę jednostek), otrzymamy liczby falowe kolejnych linii pierwszej serii widmowej wodoru (serii Lymana), leżącej w dalekim nadfiolecie (rys. 3). Gdy $n_i \to \infty$, to $\tilde{\nu} \to +R$; stała Rydberga wyraża więc w liczbach falowych energię wiązania elektronu w stanie podstawowym atomu wodoru. Podobnie — przyjmując stan o $n_f = 2$ jako końcowy, otrzymamy dla $n_i = 3, 4, 5\dots$ kolejne linie serii Balmera (il. 131, tabl. 33) itd.

serie widmowe



Rys. 3. Diagram układu poziomów energetycznych atomu wodoru

Atom wieloelektronowy

Analiza struktury widma atomu wieloelektronowego sprowadza się (przez analogię z widmem wodoru) do wyznaczenia energii stanów stacjonarnych, obliczenia ich funkcji stanu i ustalenia reguł dozwalających na przejścia między różnymi stanami. Zagadnienie to jest jednak bardzo skomplikowane (zagadnienie wielu ciał) i można je rozwiązać jedynie wprowadzając daleko posunięte uproszczenia.

Z zasady Pauliego, wg której elektrony należące do atomu (ogólniej — do określonego mikroukładu fizycznego) muszą być w różnych stanach kwantowych (tj. stanach opisanych przez nie powtarzające się w danym układzie kombinacje liczb kwantowych n, l, m_l i m_s), oraz z ograniczeń nałożonych na te liczby wynika, że gdy elektrony wypełniają powłoki o określonym n w powłoce może być $2n^2$ elektronów. Ponadto elektrony obsadzają kolejne stany o najniższej energii, a ponieważ energia stanu zależy głównie od liczby kwantowej n, będziemy mieli do czynienia z kolejnym wypełnianiem powłok od pierwszej ($n = 1$) począwszy. Można wykazać, że wypełnione powłoki charakteryzują się szczególną trwałością i słabą aktywnością chemiczną. To samo, acz w mniejszym stopniu, odnosi się do wypełnionych podpowłok (tj. zbioru stanów o określonych n i l). Wypadkowe momenty pędu (orbitalnego i spinowego) oraz związane z nimi momenty magnetyczne w wypełnionych powłokach i podpowłokach zerują się.

Przybliżenia i założenia upraszczające są następujące:

a) rozważamy stany wzbudzenia jednoelektronowego;

b) oddziaływanie elektrostatyczne elektronu optycznego z pozostałymi elektronami przedstawiamy za pomocą obliczonego w sposób uproszczony potencjału;

c) uwzględniamy oddziaływanie magnetyczne między momentami orbitalnymi i spinami elektronów powłoki walencyjnej;

d) uwzględniamy reguły wyboru ustalające możliwości przejścia promienistego między określonymi poziomami.

W spektroskopii atomowej stosuje się następującą konwencję: wartości liczby kwantowej orbitalnego momentu pędu elektronu l — 0, 1, 2, 3, 4, 5, 6 ... zastępuje się umownymi oznaczeniami s, p, d, f, g, h, i ... Oznaczenie literowe poprzedza się główną liczbą kwantową, np. $3p$ oznacza elektron w stanie o $n = 3$ i $l = 1$. Przy opisie konfiguracji powłoki elektronowej górny wskaźnik oznacza liczbę elektronów w określonym stanie (np. $1s^2\,2s^2\,2p$ oznacza konfigurację, w której pierwsza podpowłoka jest wypełniona dwoma elektronami, a druga zawiera dwa elektrony $2s$ i jeden elektron $2p$).

Momenty orbitalne i spiny elektronów powłoki walencyjnej atomu wieloelektronowego po dodaniu dają całkowity moment pędu powłoki elektronowej. Jeśli nie ma oddziaływań zewnętrznych, moment ten pozostaje niezmienny, przy czym (podobnie jak i momenty poszczególnych elektronów) jest skwantowany; jego składowa wzdłuż osi kwantowania wynosi $\hbar M_J$, a wielkość całkowitego momentu pędu jest równa $\hbar\vec{J}$, gdzie $|\vec{J}| = \sqrt{J(J+1)} \equiv J^*$. Liczby kwantowe M_J i J charakteryzują stan atomu. J jest liczbą dodatnią (całkowitą lub połówkową) lub zerem. M_J przybiera jedną z $2J+1$ wartości $-J, -J+1, \dots, J-1, J$.

całkowity moment pędu powłoki elektronowej

Ze sposobu otrzymania rzutu J na oś kwantowania wynika, że gdy liczba elektronów w powłoce walencyjnej jest nieparzysta, M_J jest liczbą połówkową (nieparzysta liczba rzutów poszczególnych spinów, z których każdy jest równy $\pm 1/2$), gdy jest ona parzysta, M_J jest całkowite. Ponieważ zaś J jest równe maksymalnej wartości M_J, więc ta sama uwaga dotyczy J.

Ważną cechą charakterystyczną kwantowania przestrzennego całkowitego momentu pędu powłoki elektronowej jest to, że ustalona jest wyłącznie wielkość składowej wzdłuż osi kwantowania, orientacja składowej poprzecznej pozostaje nieoznaczona — wszystkie jej położenia są jednakowo prawdopodobne. Można ten stan zobrazować modelowo przyjmując, że wektor $\vec{J}$ wykonuje jednostajną precesję dokoła osi kwantowania.

reguły wyboru dla J i M_J

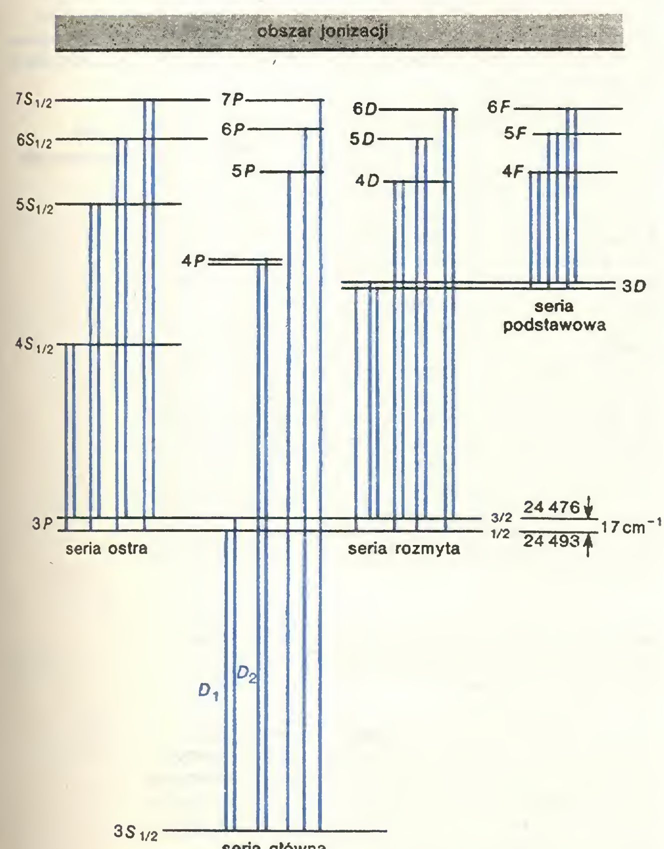
Przy przejściach promienistych jednofotonowych reguły wyboru na J i M_J są następujące: $\Delta J = 0, \pm 1$ (przejście $J = 0 \rightarrow J = 0$ jest wzbronione), $\Delta M_J = 0, \pm 1$. Ten symboliczny zapis należy rozumieć następująco: tylko te przejścia promieniste są w atomie dozwolone, w których liczby kwantowe J i M_J stanów początkowego i końcowego spełniają podane wyżej warunki.

widma elektronów o jednym elektronie walencyjnym

W metalach alkalicznych (Li, Na, K, Rb, Cs) oraz w jonach o strukturze izoelektronowej z nimi (Be^+, Mg^+, Ca^+, Ba^+, Sr^+, B^{2+}, Al^{2+} itp.) mamy jeden elektron na zewnątrz wypełnionych powłok elektronowych. Całkowity orbitalny moment pędu elektronów atomu L jest więc równy momentowi orbitalnemu l elektronu walencyjnego ($l = L$, a stany atomu odpowiednio do wartości L będą oznaczone literami S, P, D, F, G, H, I...).

W układzie, w którym jest określony wypadkowy orbitalny moment pędu powłoki elektronowej, obowiązują następujące reguły wyboru dla liczby L: $\Delta L = \pm 1$ ($\Delta L = 0$ wzbronione!). Sprzężenie wypadkowego momentu orbitalnego $\vec{L}$ ze spinem $\vec{S}$ daje całkowity moment powłoki $\vec{J}$. Stosownie do tego będziemy mieli następujące ciągi termów: nS, nP, nD... Każdy z nich (z wyjątkiem termów nS) rozszczepia się na dwa poziomy struktury subtelnej o $J = L \pm 1/2$ ($nP_{1/2}$ i $nP_{3/2}$; $nD_{3/2}$ i $nD_{5/2}$ itd.). Termy S są pojedyncze o $J = 1/2$($nS_{1/2}$). Rysunek 4 ukazuje omówione tu układy termów w atomie sodu. Stanem podstawowym jest $3S_{1/2}$.

reguły wyboru dla L

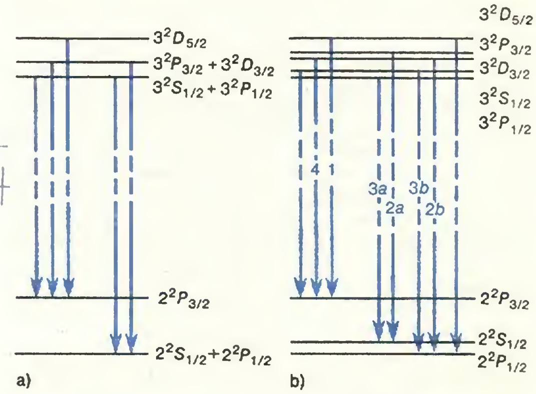


Rys. 4. Schemat poziomów atomu sodu i serie występujące w jego widmie. Rozszczepienia stanów $3P$, $4P$ i $3D$ są wielokrotnie powiększone

Struktura subtelna poziomów atomu wodoru

Relatywistyczna mechanika kwantowa Diraca prowadzi do wniosku, że struktura subtelna poziomów wodoru wykazuje zwyrodnienie względem L: energia zależy jedynie od liczb kwantowych n i J. Wyrażenie na energię stanu stacjonarnego W_n, wzór (1), zostaje uzupełnione w pierwszym przybliżeniu wyrazem poprawkowym rzędu wielkości α^2 ($\alpha = e^2/\hbar c \approx 1/137$ — jest to tzw. stała struktury subtelnej Sommerfelda) i zależnym od n i J. Obliczoną zgodnie z tym wzorem strukturę poziomów wodoru o $n = 2$ i $n = 3$ oraz strukturę pierwszej linii serii Balmera H_α ukazuje rys. 5. Zgodnie z regułami wyboru linia ta winna mieć 5 składowych, których metodami spektroskopii tradycyjnej nie można rozdzielić (wskutek rozszerzenia linii wywołanej efektem Dopplera). Mimo to jednak już prace H.G. Kuhna i G.W. Seriesa oraz R.C. Williamsa nad H i D, prowadzone metodami interferometrycznymi, wykazały pewne odstępstwa od przewidywań teoretycznych. Badania doświadczalne były stymulowane przez równoczesny rozwój elektrodynamiki kwantowej, która analizując efekty związane z fluktuacjami w zerowym polu promieniowania, przewidywała dodatkowe przesunięcia wzajemne poziomów w stosunku do położeń przewidzianych przez teorię Diraca. To przesunięcie nazwano przesunięciem Lamba i Retherforda. Dodatkowa poprawka do wzoru, w której występuje zależność od liczby kwantowej L, jest rzędu α^3. Z jej postaci wynika, że największych przesunięć można oczekiwać dla stanów S i małych wartości liczby kwantowej głównej n. Uwzględnienie przesunięcia prowadzi do komplikacji układu poziomów i struktury subtelnej linii H_α (rys. 5). W szczególności teoria przewidywała podwyższenie energii poziomu $2^2S_{1/2}$ w stosunku do $2^2P_{1/2}$ o $1057{,}99 \pm 0{,}16$ MHz. Pierwsze dokładne pomiary przesunięcia Lamba przeprowadzono łącząc metody spektroskopii optycznej i mikrofalowej. Ich zasadę w ogromnym uproszczeniu przedstawia rys. 6. Zauważymy najpierw, że stan $2^2S_{1/2}$ jest stanem metatrwałym, tzn. mimo, że jest to stan wzbudzony, reguły wyboru zakazują przejścia promienistego do stanu podstawo-

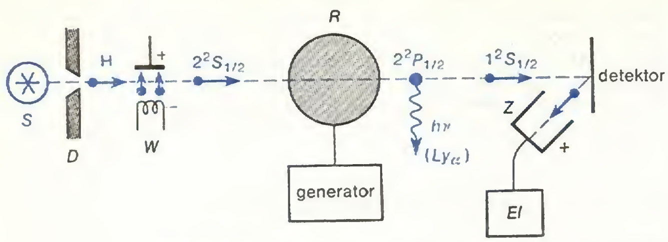
stała struktury subtelnej α



Rys. 5. Struktura linii H_α wodoru: a) wg teorii Diraca; b) z uwzględnieniem przesunięcia Lamba

przesunięcie Lamba

wego $1^2S_{1/2}$ (odpowiadałoby to przejściu o $\Delta L = 0$). Czas życia atomów w tym stanie jest bardzo długi i energia wzbudzenia (przekraczająca 10 eV) może być



Rys. 6. Schemat doświadczenia Lamba–Retherforda: *S* źródło atomów wodoru, *D* układ diafragm wydzielających smukłą wiązkę atomów, *W* urządzenie do przyspieszania elektronów i wzbudzania atomów wodoru przez zderzenia, *R* obszar pomiarowy, w którym można wytwarzać zmienne pole elektromagnetyczne, *Z* — elektroda zbiorcza, *El* — układ elektrometryczny

doświadczenie Lamba–Retherforda

stracona jedynie niemal w procesie bezpromienistym (zderzenia). W aparaturze Lamba i Retherforda wiązka atomów wodoru była wzbudzona przez zderzenia z elektronami do stanów o $n = 2$. Z tej mieszaniny atomów w stanach: podstawowym $1^2S_{1/2}$ i wzbudzonych $2^2S_{1/2}$, $2^2P_{1/2}$, $2^2P_{3/2}$ atomy w stanach $2P$ wysyłały w czasie ok. 1 ns foton linii Ly_α (pierwszej linii serii Lymana) i powracały do stanu podstawowego. W swej dalszej drodze przechodziły przez obszar pomiarowy (obszar oddziaływania z polem mikrofalowym) i dochodziły do układu detekcyjnego, nie wywołując żadnych efektów (podobnie jak atomy, które nie zostały w ogóle wzbudzone). Inaczej było z atomami w stanie metatrwałym $2^2S_{1/2}$. Gdy dochodziły one do detektora, kosztem ich energii wzbudzenia wyrzucane były elektrony rejestrowane następnie w układzie elektrometrycznym. W doświadczeniu chodziło o to, by przez przyłożenie w obszarze pomiarowym oscylującego pola elektromagnetycznego o częstości odpowiadającej rozszczepieniu poziomów $2^2S_{1/2}$–$2^2P_{1/2}$ (ok. 1058 MHz) wymusić przejście do stanu promienistego, a z niego, po emisji fotonu, otrzymać atom w stanie podstawowym, nie oddziałujący już z detektorem. Wymuszenie przejścia ma charakter bardzo ostrego rezonansu, co sprawia, że badany efekt jest niezmiernie czuły na dostrojenie. Można więc wyznaczyć szukane rozszczepienie z dokładnością lepszą niż 0,1 MHz. Otrzymany wynik $1057{,}77 \pm 0{,}10$ MHz doskonale się zgadza z podaną wyżej przewidywaną wartością teoretyczną.

Dzięki rozwojowi różnych technik spektroskopii współczesnej (omawianych dalej) stało się możliwe rozszerzenie zakresu badań na wyższe stany wzbudzone atomu wodoru i na wielokrotnie zjonizowane atomy różnych pierwiastków cięższych.

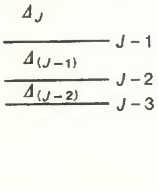
widma atomów o kilku elektronach walencyjnych

W układach wieloelektronowych istnieją różne typy sprzężeń momentów elektronowych w powłoce walencyjnej atomu. Z nich dwa najprostsze to:

1) Sprzężenie $\overrightarrow{LS}$ (Russela–Saundersa). Z tym typem sprzężenia mamy do czynienia wówczas, gdy oddziaływania wzajemne między momentami orbitalnymi poszczególnych elektronów oraz między spinami znacznie górują nad oddziaływaniami spin-orbita. Wówczas tworzy się wypadkowy moment orbitalny

sprzężenie *LS*

powłoki walencyjnej $\vec{L} = \vec{l_1} + \vec{l_2} + \ldots$ oraz wypadkowy spin $\vec{S} = \vec{s_1} + \vec{s_2} + \ldots$ Oba momenty wypadkowe dodają się, tworząc w ogólności dla każdego L liczbę $2S+1$ możliwych kombinacji prowadzących do różnych wartości J (rozszczepienie multipletowe termów). Rozszczepienia multipletowe wykazują regularności. Za przykład może służyć reguła odległościowa Landégo: rozszczepienie pary kolejnych poziomów Δ_J: $\Delta_{(J-1)} : \Delta_{(J-2)} = J : (J-1) : (J-2)$. Reguła ta ma duże znaczenie przy analizie nadsubtelnej struktury linii widmowych.



multiplet

Terminu „multiplet" używa się na ogół w dwóch różnych znaczeniach, odnosi się go albo do termu energii, albo do układu linii widmowych. Pierwsze znaczenie odpowiada sytuacji omawianej w tekście. Gdy np. w rtęci, mającej w powłoce walencyjnej dwa elektrony, spiny są równoległe ($S = 1$), wówczas każdy z termów L (np. $P, D \ldots$) rozszczepi się na $2S+1 = 3$ poziomy o $J = L-1$, L, $L+1$ (np. 6^3P_0, 6^3P_1, 6^3P_2). Mamy trypletowy układ poziomów. Natomiast grupę linii widmowych powstałych z przejść między poziomami dwu różnych rozszczepionych na multiplety termów (np. $6^3D_{J'}$–6^3P_J) nazwiemy multipletem widmowym.

W układach wieloelektronowych sumowanie spinów prowadzi, zależnie od ich orientacji, do różnych wartości wypadkowej S, co oznacza, że w określonym atomie może występować kilka układów termów o różnej krotności. W układzie dwuelektronowym wypadkowa S może być równa bądź 0, bądź 1 i termy są odpowiednio pojedyncze (układ singletowy) lub potrójne (układ trypletowy). Ważnym warunkiem uzupełniającym reguły wyboru dla przejść promienistych jest zakaz interkombinacji $\Delta S = 0$. Przejściu promienistemu nie może towarzyszyć przeorientowanie się spinu. Ta reguła nie jest bezwzględnie zachowana w atomach ciężkich (np. linia 253,7 nm rtęci). Jednak przejścia promieniste zachodzące z jej naruszeniem są znacznie mniej prawdopodobne niż inne, przy których reguła ta jest spełniona.

sprzężenie $\{j, j\}$

2) Sprzężenie $\{j, j\}$. Oddziaływania spin-orbita znacznie górują nad oddziaływaniami wzajemnymi momentów orbitalnych oraz spinów. W konsekwencji dla każdego elektronu oddzielnie tworzą się najpierw wypadkowe momenty całkowite $\vec{l_1} + \vec{s_1} = \vec{j_1}$; $\vec{l_2} + \vec{s_2} = \vec{j_2} \ldots$, a następnie moment całkowity tworzy się przez złożenie $\vec{j_1} + \vec{j_2} + \ldots = \vec{J}$. Nie możemy teraz mówić ani o wypadkowym momencie orbitalnym, ani o wypadkowym spinie powłoki walencyjnej. Umowny zapis stanu atomu w wypadku sprzężenia $\{j, j\}$ jest: $n_1 l_1 n_2 l_2 \{j_1, j_2\}_J$.

stany metatrwałe

W atomach mających więcej niż jeden układ termów napotykamy stany wzbudzone (stany metatrwałe), z których promienisty powrót do stanów niżej położonych (w szczególności do stanu podstawowego) nie jest dozwolony ani pośrednio, ani bezpośrednio, np. stany 2^3S_1 i 2^1S_0 w He lub 6^3P_0 i 6^3P_2 w rtęci. Atomy w takich stanach mają czasy życia od ok. 10^{-2} s do 1 s. Dezaktywacja ich zachodzi głównie w sposób bezpromienisty przy zderzeniach ze ściankami naczynia, w którym gaz jest zamknięty. Jednakże w szczególnych warunkach można obserwować przejścia promieniste z tych stanów. Są to tzw. linie wzbronione, o bardzo małym w porównaniu z liniami dozwolonymi natężeniu. Występują one np. w widmach mgławic lub w świeceniu górnych warstw atmosfery (w zorzy polarnej). Atomy w stanach metatrwałych odgrywają istotną rolę w wyładowaniach elektrycznych w gazach rozrzedzonych. Szczególnie w gazach szlachetnych, w mieszaninie He i Ne, atomy He w stanach metatrwałych stanowią rezerwuar energii przekazywanej do stanu $3s_2$ neonu stanowiącego stan wyjściowy do akcji laserowych.

Rozszczepienie poziomów energetycznych atomu w polu magnetycznym (zjawisko Zeemana)

Badania rozszczepienia poziomów atomowych, a co za tym idzie — i linii widmowych w polu magnetycznym dostarczają wielu cennych informacji pomocnych przy identyfikacji i systematyzacji widm i poziomów energetycznych. Doprowadziły one również do stworzenia nowych metod spektroskopii atomowej.

Z momentami pędu elektronu (orbitalnym i spinem) są związane elementarne momenty magnetyczne (→ Teoria magnetyzmu):

$$\vec{\mu}_l = -\frac{e\hbar}{2mc}\vec{l} \quad \text{i} \quad \vec{\mu}_s = -\frac{e\hbar}{mc}\vec{s}.$$

Zjawisko Zeemana należy rozpatrywać jako oddziaływanie między polem zewnętrznym i tymi elementarnymi momentami. Zasadniczym czynnikiem charakteryzującym sytuację fizyczną po wprowadzeniu pola magnetycznego jest wyznaczenie kierunku kwantowania jednakowego dla wszystkich atomów zbioru. Drugim istotnym czynnikiem, który należy wziąć pod uwagę przy analizie tego zjawiska jest stosunek wielkości oddziaływania między polem magnetycznym i momentami atomowymi do wielkości oddziaływań wewnątrzatomowych sprzęgających ze sobą różne momenty.

Ograniczymy się przy tym do rozpatrzenia tylko przypadków krańcowych: a) pola silnego — gdy energia oddziaływania poszczególnych momentów z polem jest duża w porównaniu z energią sprzężeń między tymi momentami; wówczas to sprzężenia wewnątrzatomowe zostają przełamane i każdy z momentów ustawia się niezależnie względem pola; b) pola słabego — gdy energia oddziaływania z polem stanowi tylko niewielkie zaburzenie sprzężeń wewnątrzatomowych, które zostają zachowane; skutkiem oddziaływania z polem jest tylko skwantowane ustawianie się momentu wypadkowego względem pola.

Moment magnetyczny powłoki walencyjnej o jednym elektronie wynosi

$$\vec{\mu} = \vec{\mu}_l + \vec{\mu}_s = -\frac{e\hbar}{2mc}(\vec{l}+2\vec{s}) - \mu_B(\vec{l}+2\vec{s}),$$

gdzie $\mu_B = \frac{e\hbar}{2mc}$ jest magnetonem Bohra (moment magnetyczny elektronu w stanie p; $l = 1$).

atom o jednym elektronie walencyjnym w silnym polu

W nieobecności pola kierunki $\vec{\mu}$ różnych atomów są rozłożone chaotycznie w przestrzeni. Gdy natomiast wprowadzimy silne pole $\vec{H}$, sprzężenie $\vec{L}\vec{S}$ zostanie przełamane i każdy z momentów ustawi się niezależnie w sposób skwantowany względem pola $\vec{H}$. Rzuty $\vec{l}$ i $\vec{s}$ będą opisane przez liczby kwantowe m_l i $m_s = \pm 1/2$. Tak więc rzut $\vec{\mu}$ na kierunek $\vec{H}$:

$$\mu_H = -\mu_B(m_l+2m_s),$$

a energia oddziaływania z polem

$$W_H = -\vec{\mu}\cdot\vec{H} = \mu_B H(m_l+2m_s).$$

Każdej wartości l odpowiada $2l+1$ wartości m_l, a ponadto dwie wartości $m_s = \pm 1/2$, zatem poziom ulega rozszczepieniu. Wartości nawiasu otrzymane przy sumowaniu różnych m_l i m_s częściowo się powtarzają i otrzymujemy ostatecznie $2l+3$ różnych wartości liczbowych. Energia całkowita $W = W_0 + W_H$, gdzie W_0 jest energią w nieobecności pola. Liczby falowe linii składowych struktury zeemanowskiej przejścia między poziomami i i f opisuje wzór:

$$\tilde{\nu}^{if} = (W^i - W^f)/hc = \frac{1}{hc}[(W_0^i + W_H^i) - (W_0^f + W_H^f)].$$

Wprowadzając na W_H wyrażenia zależne od m_l i m_s, stosując reguły wyboru $\Delta m_l = 0$ lub ± 1 i $\Delta m_s = 0$(!) (zachowanie orientacji spinu w przejściu promienistym) otrzymujemy:

$$\tilde{\nu}^{if} = \tilde{\nu}_0^{if} + \frac{eH}{4\pi mc^2}\Delta m_l = \tilde{\nu}_0^{if} \pm LH\Delta m_l;$$

$L = 4{,}68604\ \mathrm{m^{-1}T^{-1}}$ jest tzw. jednostką Lorentza. Otrzymujemy w ten sposób rozszczepienie linii na 3 składowe odpowiednio do wartości Δm_l:

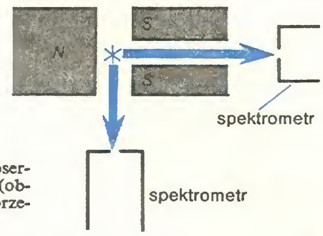
$$\tilde{\nu}^{if} = \tilde{\nu}_0^{if} \pm LH, \text{ gdy } m_l = \pm 1 \text{ (składowe } \sigma)$$

$$\tilde{\nu}^{if} = \tilde{\nu}_0^{if}, \text{ gdy } m_l = 0 \text{ (składowa } \pi).$$

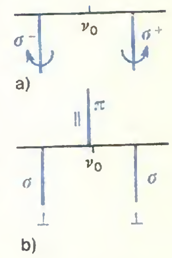
Jest to tzw. tryplet Lorentza — jak to przewidywała klasyczna teoria tego zjawiska. Rozsunięcie składowych bocznych jest proporcjonalne do natężenia pola magnetycznego.

obserwacja zjawiska Zeemana

Aby zaobserwować zjawisko Zeemana, winniśmy umieścić źródło promieniowania między biegunami silnego elektromagnesu. Stosowane są dwa typy obserwacji: podłużna — promieniowania rozchodzącego się wzdłuż linii sił pola (do obserwacji służy kanał przewiercony w rdzeniu elektromagnesu), i poprzeczna — prostopadle do linii sił (rys. 7). W każdym z tych



Rys. 7. Dwa sposoby obserwacji zjawiska Zeemana (obserwacja podłużna i poprzeczna)



Rys. 8. Rozszczepienie zeemanowskie w silnym polu magnetycznym i rodzaje polaryzacji linii: a) obserwacja podłużna, b) obserwacja poprzeczna

przypadków otrzymuje się inny obraz składowych (rys. 8). Przy obserwacji podłużnej obserwuje się tylko składowe σ. Są one spolaryzowane kołowo, o przeciwnych zwrotach (rys. 8a). Przy obserwacji poprzecznej mamy wszystkie trzy składowe: składowa nie przesunięta (π) jest spolaryzowana liniowo, równolegle do kierunku $\vec{H}$; składowe σ są również spolaryzowane liniowo, prostopadle do $\vec{H}$ (rys. 8b).

polaryzacja składowych

Tego typu struktura nosi czasem nazwę normalnego zjawiska Zeemana. Nazwa ma pochodzenie historyczne — rozszczepienie na trzy składowe wynikało z teorii elektronów Lorentza, przeto struktury, które były zgodne z tą teorią, nazywano normalnymi, wszystkie inne stanowiły zjawisko anomalne.

zjawisko Zeemana w słabym polu

Słabe pole zewnętrzne nie narusza sprzężeń wewnątrzatomowych i stanowi w stosunku do nich jedynie niewielkie zaburzenie. Na przykład przy sprzężeniu $\vec{L}\vec{S}$ całkowity moment powłoki elektronowej $\vec{J}$ pozostaje wielkością dobrze określoną, a skutkiem oddziaływania z polem jest taka orientacja $\vec{J}$ względem pola $\vec{H}$, że rzut $\vec{J}$ na kierunek pola wynosi $M_J\hbar$. Energia oddziaływania z polem będzie więc (podobnie jak poprzednio) proporcjonalna do rzutu M_J całkowitego momentu pędu na kierunek pola:

$$W_H = \frac{e\hbar H}{2mc} g M_J = \mu_B H g M_J,$$

gdzie g jest stałą charakteryzującą dany poziom i wyrażającą się w wypadku sprzężenia $\vec{L}\vec{S}$ bardzo prosto przez liczby kwantowe L, S i J:

$$g = 1 + \frac{J(J+1)+S(S+1)-L(L+1)}{2J(J+1)}$$

(w innych rodzajach sprzężeń wyrażenie na g jest znacznie bardziej złożone). Stałe g różnych poziomów są w ogólności różne, zależne od L, S i J; a więc:

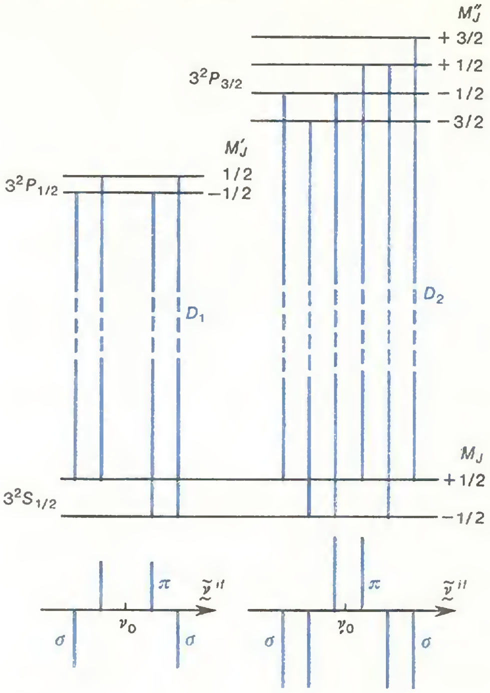
$$\tilde{\nu}^{if} = \tilde{\nu}_0^{if} + LH(g^i M_J^i - g^f M_J^f).$$

Reguły wyboru na M_J są następujące: $\Delta M_J = 0$ lub ± 1. Przejściu o $\Delta M_J = 0$ odpowiadają składowe π, a przejściu o $\Delta M_J = \pm 1$ — składowe σ. Ich warunki obserwacji i stan polaryzacji jest podobny jak w silnym polu.

rozszczepienie dubletu sodu

Struktura zeemanowska stanów o tych samych liczbach kwantowych L, S i J, a więc stanów odpowiadających tej samej sekwencji termów (np. 6^3P_1, 7^3P_1, 8^3P_1 itd.), jest identyczna. Różni się natomiast struktura stanów tworzących określony multiplet (np. 6^3P_0, 6^3P_1, 6^3P_2), bo różne są wartości J (chociaż L i S są takie same). Rozszczepienia dubletu sodu w słabym

polu przedstawia rys. 9. Trzem stanom $3^2S_{1/2}$, $3^2P_{1/2}$, $3^2P_{3/2}$ odpowiadają wartości $g' = 2$, $g'' = {}^2/_3$ i $g''' = {}^4/_3$. Charakterystyczną cechą otrzymanej struktury



Rys. 9. Struktura zeemanowska dubletu rezonansowego sodu w słabym polu magnetycznym

składowe σ i π

jest zgrupowanie składowych σ na zewnątrz struktury, a składowych π — w obszarze bliskim położenia linii w polu zerowym, ν_0. Ze wzrostem natężenia pola składowe σ odsuwają się coraz dalej, a wskutek postępującego przełamywania sprzężenia wewnątrzatomowego grupują się, tworząc w końcu obraz taki jak w silnym polu. Widać wreszcie, że ponieważ w słabym polu o obrazie zeemanowskim decyduje całkowity moment pędu powłoki elektronowej, nie ma potrzeby rozpatrywania oddzielnie sytuacji, gdy w powłoce walencyjnej jest jeden elektron lub gdy jest ich kilka.

Nadsubtelna struktura linii widmowych

Badanie widm za pomocą urządzeń spektralnych o wielkiej zdolności rozdzielczej (interferometry) ujawniło, że linie widmowe mają niejednokrotnie strukturę bardzo złożoną, rozszczepiają się na kilka, a czasem nawet kilkanaście składowych. Strukturze tej nadano nazwę nadsubtelnej struktury linii, a jej zbadanie nie tylko pozwoliło na odkrycie nowych i ważnych efektów związanych z oddziaływaniem elektronów powłoki z jądrem atomu, lecz również doprowadziło do wyznaczenia podstawowych parametrów charakteryzujących jądro.

szerokość linii widmowych

Badanie struktury nadsubtelnej utrudnia fakt, że linie widmowe mają skończoną szerokość spektralną, co wynika z nałożenia się kilku efektów. Wspomnimy tu o dwóch. Szerokość naturalna jest związana z nieokreślonością energii ΔW stanu atomowego, zgodnie z zasadą Heisenberga $\Delta W\tau \geqslant \hbar$, gdzie τ jest czasem życia atomu w tym stanie. Przechodząc do częstości, otrzymujemy $\Delta\nu = \Delta W/h \geqslant (1/2\pi\tau)$ i przy typowych czasach życia stanów wzbudzonych ($\tau = 10^{-9}$ s) $\Delta\nu \approx 1{,}5\cdot 10^8$ Hz. Jedynie stan podstawowy i stany metatrwałe można uważać za spektralnie ostre. Szerokość dopplerowska linii wiąże się z tym, że promieniowanie jest wysyłane przez zbiór atomów o chaotycznie rozłożonych prędkościach ruchu termicznego. Szerokość ta $\Delta\nu_D = 8{,}6\cdot 10^{-7}\,\nu_0\sqrt{T/\mu}$ zależy więc od częstości przejścia ν_0, od temperatury bezwzględnej gazu T i jego masy cząsteczkowej μ. Szerokość dopplerowska jest dla przejść optycznych co najmniej o rząd wielkości większa od szerokości naturalnej, którą możemy wówczas w pierwszym przybliżeniu pominąć. Przy badaniach nadsubtelnej struktury stosujemy więc źródła promieniowania tak zbudowane, by rozszerzenie dopplerowskie linii zminimalizować.

spin jądra

Podobieństwo między strukturą hipermultipletów (tak nazwano multiplety nadsubtelnej struktury) i multipletów struktury subtelnej nasunęło hipotezę (W. Pauli 1924 r.), że jądro atomowe jest obdarzone własnym momentem pędu (spinem jądrowym), z którym jest związany również moment magnetyczny. Spin jądrowy $\hbar\vec{I}$ jest opisany przez liczbę kwantową I ($|\vec{I}| = \sqrt{I(I+1)} \equiv I^*$).

Jądra o nieparzystej liczbie masowej mają w ogólności spin połówkowy ($I = {}^1/_2, {}^3/_2 \ldots$). Gdy liczba masowa jest parzysta, spin jądra jest na ogół równy zeru. Wśród jąder trwałych niewiele jest wyjątków od tej reguły: ^{2}H($I = 1$), ^{6}Li ($I = 1$), ^{14}N($I = 1$), ^{40}K ($I = 4$).

Oddziaływanie między spinem jądrowym $\vec{I}$ i całkowitym momentem pędu powłoki elektronowej $\vec{J}$ prowadzi do wytworzenia całkowitego momentu pędu atomu $\vec{F} = \vec{I}+\vec{J}$ (podobnie jak sprzężenie momentów $\vec{L}$ i $\vec{S}$ w $\vec{J}$). Z niezwykłej małości rozszczepienia nadsubtelnego wnioskujemy, że oddziaływanie między $\vec{I}$ i $\vec{J}$ jest bardzo słabe i nie narusza istniejących w powłoce elektronowej sprzężeń. Analiza widm nadsubtelnej struktury ujawniła prawidłowości analogiczne do widm struktury subtelnej. Ułatwiło to poważnie rozwikłanie tych struktur i wyznaczenie na tej podstawie spinu jądrowego $\vec{I}$. Przez wiele lat analiza nadsubtelnej struktury była jedynym źródłem informacji o spinie jądrowym i większość danych dotyczących jąder trwałych pochodzi z tych właśnie pomiarów.

wpływ elektrycznego momentu kwadrupolowego na strukturę nadsubtelną

W dotychczasowych rozważaniach traktowano jądro jako twór punktowy. Nawet wprowadzenie spinu jądrowego nie wymagało dodatkowych założeń o jego skończonych rozmiarach. Jednakże staranne wyznaczenie rozszczepień nadsubtelnych wykazało, że w wielu wypadkach napotykamy wyraźne odstępstwa od wzmiankowanych poprzednio reguł odległościowych. Te odstępstwa powiązano z modyfikacją wyrażenia na energię oddziaływania elektrostatycznego między jądrem i elektronem walencyjnym, wynikającą ze skończonych rozmiarów jądra i z faktu, że jądro (a zatem i rozkład ładunku elektrycznego) nie jest kuliste. Jądro jest elipsoidą obrotową (spłaszczoną lub wydłużoną; → Modele jądrowe). Natężenie pola elektrycznego pochodzącego od powłoki elektronowej zmienia się w obszarze jądra i występuje dodatkowe oddziaływanie, zależne od stanu powłoki elektronowej. Wskutek tego oddziaływania rozszczepienia nadsubtelne ulegają modyfikacji. Aby efekty można było zaobserwować, musi być spełniony warunek $I \geqslant 1$ i $J \geqslant 1$. Analiza modyfikacji struktury nadsubtelnej pozwala na obliczenie elektrycznego momentu kwadrupolowego jądra (wielkości tensorowej będącej miarą odstępstwa rozkładu ładunku w jądrze od rozkładu kulistego), a stąd na wnioskowanie o kształcie jądra.

efekty izotopowe

Na nadsubtelną strukturę linii widmowych wpływa skład izotopowy pierwiastka. Na przykład rtęć w postaci naturalnej jest mieszaniną sześciu izotopów. Dwa z nich (^{199}Hg i ^{201}Hg) obdarzone są spinem jądrowym, pozostałe (^{198}Hg, ^{200}Hg, ^{202}Hg i ^{204}Hg) mają spin jądrowy równy zeru.

Analiza zaobserwowanej struktury linii wykazała istnienie przesunięcia izotopowego, przesunięcia o innym charakterze niż wywołane przez różnice w wiel-

efekt objętościowy

kości masy zredukowanej m układu jądro–elektron. Powodem przesunięcia izotopowego jest tzw. efekt objętościowy.

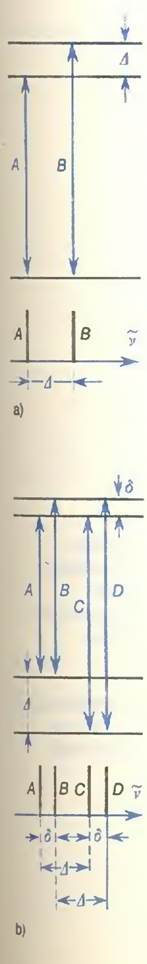
•Różne izotopy tego samego pierwiastka mają w jądrze różne liczby neutronów. Ze wzrostem liczby neutronów wzrasta objętość jądra, a zatem ulega zmianie rozkład pola kulombowskiego w jego bezpośrednim otoczeniu. Wynikają stąd niewielkie modyfikacje energii stanów elektronowych, szczególnie widoczne wówczas, gdy dodając neutron, przechodzimy od wypełnionej powłoki neutronowej w jądrze do obsadzania następnej.

Informacje uzyskane z omówionych efektów nie stanowią wszystkich informacji, jakie można uzyskać z badań nadsubtelnej struktury linii. Osiągnięto w tej dziedzinie taką dokładność pomiarów, że przy interpretacji zaobserwowanego rozszczepienia zachodzi np. konieczność uwzględnienia skończonych rozmiarów dipola magnetycznego, jakim jest jądro, a przy dyskusji efektu oddziaływania kwadrupolowego trzeba wprowadzić poprawkę uwzględniającą to, że skończone rozmiary ładunku jądra i jego odstępstwo od rozkładu kulistego wywołują ze swej strony pewne deformacje w powłoce elektronowej, częściowo kompensujące oddziaływanie kwadrupolowe.

badanie nadsubtelnej struktury krótkożyjących pierwiastków

Rozwinięcie metod pompowania optycznego i spektroskopii laserowej (zob. niżej) pozwoliło rozszerzyć badania nadsubtelnej struktury linii widmowych na krótkożyjące (czas życia rzędu 1 min) izotopy różnych pierwiastków dalekie od tzw. ścieżki trwałości (→ Jądra atomowe i ich wzbudzenia), zawierające znaczny nadmiar lub niedobór neutronów, dostępne w bardzo małych ilościach.

W wielu doświadczeniach, które objęły zbadanie izotopów baru (A 125–135), metali alkalicznych: sodu (A 21–31), rubidu (A 76–98), cezu (A 118–145) i fransu (A 208–213), wreszcie rtęci (A 181–206), stosowano bardzo wyspecjalizowane techniki spektroskopii laserowej nieraz połączone np. z metodami pompowania optycznego, separacji stanów przez odchylanie wiązek atomowych w niejednorodnym polu magnetycznym lub zestrajania częstości wzbudzenia optycznego atomu z częstością lasera przez skierowanie wiązki atomowej i światła współliniowo i dobór przesunięcia dopplerowskiego przez nadanie atomom odpowiedniej prędkości. Z badań tych można wnioskować o kształcie jąder i o nośnikach momentów jądrowych.



Rys. 10. Nadsubtelna struktura linii widmowych. Schematy przejścia i odpowiadające im struktury: a) jeden stan rozszczepiony, b) oba stany rozszczepione

Metoda pompowania optycznego

Przy omawianiu nadsubtelnej struktury linii widmowych zwróciliśmy uwagę na ograniczenie możliwości badania wąskich struktur spowodowane rozszerzeniem dopplerowskim linii widmowych. Metodami tradycyjnymi wyznaczamy rozszczepienie poziomów pośrednio — jako różnicę liczb falowych składowych rozszczepionych linii $\Delta = \tilde{\nu}_A - \tilde{\nu}_B$ (rys. 10a), co stanowi dalsze ograniczenie precyzji pomiarów. Gdy w dodatku rozszczepiają się oba stany zaangażowane w przejściu, wówczas zaobserwowana struktura zależy od rozszczepienia obu stanów, a to komplikuje procedurę analizy rozszczepienia (rys. 10b).

Metoda pompowania optycznego, zainicjowana i rozwinięta przez A. Kastlera i J. Brossela, jest sposobem przygotowania układu atomowego do pomiarów pozwalających m.in. na bezpośrednie mierzenie niewielkich rozszczepień określonego poziomu. Przygotowanie to polega na tym, że za pomocą wzbudzenia optycznego (promieniowaniem o odpowiednio dobranym składzie spektralnym, kierunku rozchodzenia się wiązki i rodzaju jej polaryzacji; rys. 7, 8, 9) możemy otrzymać selektywne obsadzenie wybranych podpoziomów badanej struktury. Różni się ono znacznie od obsadzenia, które odpowiada stanowi równowagi termodynamicznej. Tak przygotowany układ możemy poddać działaniu pól elektromagnetycznych o częstości rezonansowej dla przejścia między podpoziomami zeemanowskimi lub nadsubtelnymi i obserwować rezonans magnetyczny, tj. przejścia wymuszone przez pole magnetyczne, albo też przerwać pompowanie i obserwować samorzutny powrót układu do stanu równowagi termodynamicznej (relaksację układu). W obu rodzajach doświadczeń sygnał pomiarowy jest niesiony przez promieniowanie świetlne, co zapewnia bardzo wysoką czułość detekcji. Częstości przejść między rozszczepionymi podpoziomami zawierają się w obszarze od niewielu kHz (rezonanse jądrowe) do kilkunastu GHz (przejścia między podpoziomami nadsubtelnej struktury), czyli w zakresie częstości radiowych i mikrofalowych.

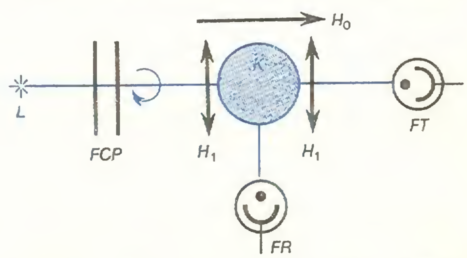
selektywne obsadzenie podpoziomów

W omawianym zakresie częstości szerokość dopplerowska sygnału jest tak mała, że można jej zupełnie nie brać pod uwagę. W ten sposób zostaje połączona bardzo wielka precyzja wyznaczania częstości w obszarze pól radiowych z ogromną czułością detekcji optycznej.

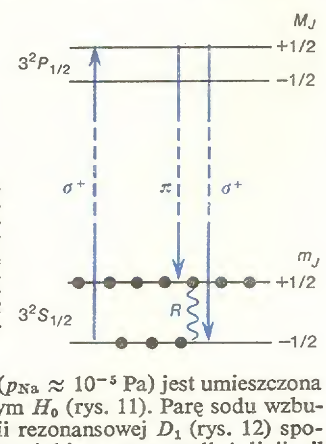
Pompowanie optyczne par atomowych w stanie podstawowym

pompowanie optyczne par sodu

Zasadę pompowania optycznego w stanie podstawowym rozpatrzymy na przykładzie sodu. Dla uproszczenia pominiemy na razie obecność spinu jądrowego. Komórka szklana K zawierająca parę sodu pod niewielkim ciśnieniem ($p_{Na} \approx 10^{-5}$ Pa) jest umieszczona w polu magnetycznym H_0 (rys. 11). Parę sodu wzbudzamy światłem linii rezonansowej D_1 (rys. 12) spolaryzowanym kołowo i biegnącym wzdłuż linii sił pola. Badamy bądź promieniowanie przepuszczone przez komórkę, bądź też promieniowanie fluorescencji rezonansowej obserwowane w kierunku prostopadłym do wiązki wzbudzającej. Z obrazu struktury zeemanowskiej linii D_1 sodu widzimy, że wzbudzenie promieniowaniem o polaryzacji kołowej σ^+ (odpowia-

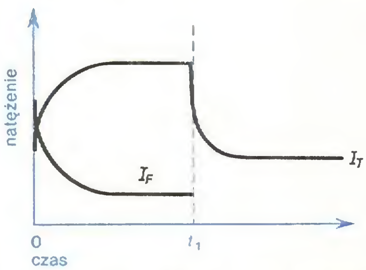


Rys. 11. Układ do badania pompowania optycznego w stanie podstawowym: K komórka z parami sodu, L lampa spektralna, F i CP filtry wydzielające linię D_1 i polaryzujące jej światło kołowo, FT i FR fotopowielacze elektronowe mierzące sygnał światła przepuszczonego lub fluorescencji rezonansowej; H_0 stałe pole magnetyczne, H_1 pole magnetyczne oscylujące



Rys. 12. Schemat przebiegu pompowania. Rozszczepienie zeemanowskie jest bardzo małe w porównaniu z odległością poziomów. Skuteczne dla pompowania jest przejście π. Relaksacje R przeciwdziałają pompowaniu

dającej $\Delta M_J = +1$) działa tylko na atomy w stanie podstawowym o $m_J = -1/2$ i prowadzi do stanu $3^2P_{1/2}$ o $M_J = +1/2$. Jest to w rozpatrywanych warunkach jedyne przejście dozwolone przez reguły wyboru. W ten sposób, chociaż pewna część atomów powróci do stanu wyjściowego, pozostałe znajdą się w stanie o $m_J = +1/2$ i pozostaną w nim (nie ma z niego absorpcji). Gdy będziemy pompowali światłem o dostatecznie dużym natężeniu, po upływie krótkiego czasu przeważająca część atomów zostanie przeniesiona do stanu o $m_J = +1/2$. Oczywiście będą występowały równocześnie procesy relaksacyjne R, dążące do przywrócenia stanu równowagi termodynamicznej, umiemy jednak zmniejszyć je do tego stopnia, by w czasie doświadczenia utrzymać znaczną przewagę obsadzenia podpoziomu $m_J = +1/2$ nad obsadzeniem $m_J = -1/2$. Zmiany w obsadzeniu podpoziomu dolnego są sygnalizowane przez wzrost przezroczystości komórki (przepompowanie większości atomów z podpoziomu $m_J = -1/2$ do $m_J = +1/2$ eliminuje ze zbioru te, które mogą pochłaniać padające promieniowanie o użytej polaryzacji). Zmniejszeniu absorpcji oczywiście towarzyszy spadek natężenia fluorescencji rezonansowej. Oba przebiegi ilustruje rys. 13. Przypuśćmy, że wskutek pompowania w ogólnej liczbie N

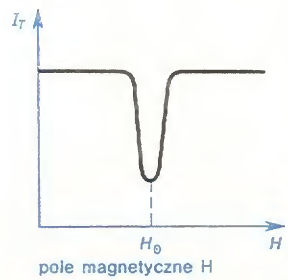


Rys. 13. Zmiany natężenia wiązki przechodzącej I_T oraz fluorescencji rezonansowej I_F w czasie pompowania ($0 \leq t \leq t_1$). Przebieg I_T znajdujemy stosując niezależną od pompowania słabą wiązkę sondującą. Pompowanie przerywamy w chwili t_1 i dla $t > t_1$ przebiegi ilustrują relaksację w ciemności

atomów sodu w cm^3 ustala się taki rozkład, iż na poziomie $m_J = 1/2$ znajduje się n^+ atomów, na poziomie $m_J = -1/2$ znajduje się n^- atomów (w stanie równowagi w każdym z tych podpoziomów jest $N/2$ atomów w cm^3). Stopniem orientacji (lub polaryzacją) zbioru atomów nazywamy wielkość $P = (n^+ - n^-)/N$. Zbiór zorientowany ma makroskopowy moment pędu o gęstości $\mathcal{J} = (\hbar/2)NP$ ($J = 1/2$) i związany z nimi makroskopowy moment magnetyczny o gęstości $\mathcal{M} = \gamma(\hbar/2)NP$ (tutaj γ — współczynnik giromagnetyczny).

rezonans magnetyczny w oscylującym polu

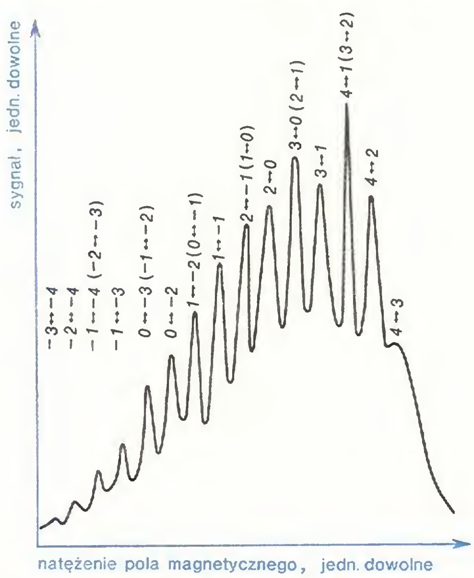
W słabym polu magnetycznym H_0 rozszczepienie zeemanowskie f_0 sąsiednich (tj. o m_J różniącym się o 1) podpoziomów wynosi $f_0 = g\mu_B H_0/h$. Jeżeli teraz, przy ustalonym natężeniu wiązki pompującej, wytworzymy w obszarze komórki oscylujące pole magnetyczne H_1 (prostopadłe do H_0) o częstości f przestrajanej w obszarze f_0, to przy $f = f_0$ zaobserwujemy raptowne zmniejszenie natężenia wiązki przechodzącej przez komórkę. Odpowiada to rezonansowi częstości pola oscylującego z częstością odpowiadającą rozszczepieniu podpoziomów zeemanowskich. Wymusza ono przejście między podpoziomami o $m_J = 1/2$ i $m_J = -1/2$, wyrównując ich obsadzenia. Ostrość sygnału przy przejściu przez rezonans pozwala na bardzo dokładne zmierzenie f_0, a stąd — rozszczepienia zeemanowskiego. Częściej (ze względów technicznych) stosuje się pewną modyfikację opisanego układu: częstość generatora pola oscylującego f_0 jest ustalona, a przez obszar rezonansu przechodzimy zmieniając powoli natężenie rozszczepiającego pola magnetycznego (rys. 14).



Rys. 14. Sygnał rezonansu magnetycznego. Częstość generatora pola oscylującego jest ustalona, a pole rozszczepiające jest zmieniane stopniowo

badanie rozszczepienia zeemanowskiego

W atomie rzeczywistym, którego jądro ma spin różny od zera, obraz pompowania nie różni się istotnie od opisanego powyżej, jakkolwiek jego przebieg i opis ilościowy są bardziej skomplikowane. Obecność spinu jądrowego powoduje pojawienie się nadsubtelnej struktury poziomów, ich rozszczepienie zeemanowskie jest więc bardziej złożone; słabość oddziaływania nadsubtelnego sprawia przy tym, że już w niewielkich polach (rzędu 10^{-4} T) sprzężenie między $\vec{I}$ i $\vec{J}$ zostaje naruszone, podpoziomy zeemanowskie przestają być równoodległe i przy ustalonej częstości f_0 rezonanse między sąsiednimi podpoziomami zeemanowskimi pojawiają się przy różnych wartościach H. Badanie widma rezonansowego (za przykład może



Rys. 15. Widmo rezonansowe cezu. Liczby przy maksimach oznaczają zmiany kwantowej liczby magnetycznej ($M_F \to M_F'$). Oprócz przejść jednokwantowych ($\Delta M_F = 1$) widoczne są ostre maksima przejść wielokwantowych ($\Delta M_F = 2, 3$)

służyć widmo Cs, rys. 15) przy różnych wielkościach pola rozszczepiającego i różnych f_0 jest ważnym źródłem informacji o przebiegu rozszczepienia zeemanowskiego w polach pośrednich.

pomiar natężenia pola magnetycznego

Gdy komórkę wypełnimy substancją, której stałe rozszczepienia zeemanowskiego są doskonale znane, wyznaczenie częstości rezonansowej f_0 pozwoli wyznaczyć natężenie pola magnetycznego, w którym się komórka znajduje. Na tej zasadzie zbudowano bardzo precyzyjne magnetometry do mierzenia słabych pól (m.in. do badań przestrzeni kosmicznej).

zastosowanie w fizyce jądrowej

Pompowanie atomów o różnym od zera spinie jądrowym prowadzi za pośrednictwem sprzężenia $\vec{I}\,\vec{J}$ do orientacji spinów jądrowych. W ten sposób można otrzymać układy zorientowanych jąder promieniotwórczych i obserwować np. anizotropię rozkładu przestrzennego wysyłanego promieniowania jądrowego β lub γ. Niepromieniotwórcze jądra mogą być użyte jako zorientowane (spolaryzowane) tarcze lub źródła zorientowanych wiązek w badaniach rozpraszania.

Procesy relaksacji

Wspomnieliśmy o tym, że wytworzeniu różnicy obsadzeń towarzyszy występowanie procesów relaksacyjnych, dążących do sprowadzenia układu do stanu równowagi termodynamicznej. Przy stałym pompowaniu, procesy relaksacyjne prowadzą do ustalenia się stanu stacjonarnego, w którym uzyskany stopień orientacji zależy od szybkości obu procesów. Gdy pompowanie zostaje wyłączone, relaksacja sprowadza układ do stanu równowagi termodynamicznej (rys. 13, przebiegi krzywych przy $t > t_1$). Z czasowego przebiegu zaniku orientacji (mamy do czynienia zazwyczaj z zanikiem wykładniczym lub kombinacją zaników wykładniczych o różnych stałych czasu) możemy wnioskować o naturze procesów relaksacyjnych. Są one związane ze zderzeniami, jednak zależnie od sytuacji fizycznej — natura mechanizmów dezorientujących może być różna. Przy pompowaniu par metali alkalicznych potężnym czynnikiem relaksacyjnym są zderzenia atomów zorientowanych ze ściankami szklanymi lub kwarcowymi komórki. Zderzeniu ze ścianką towarzyszą oddziaływania chemiczne wiążące na niej atom metalu. W ten sposób orientacja wywołana przez pompowanie zostaje całkowicie utracona. Aby zredukować wpływ ścianek, napełnia się komórki rezonansowe gazem buforującym (jest to najczęściej gaz szlachetny pod ciśnieniem od dużego ułamka hPa do kilkuset hPa) albo też pokrywa się ścianki komórki cienką warstwą nieaktywnej chemicznie substancji (ciężkie parafiny lub silany). W pierwszym wypadku droga swobodna zorientowanych atomów ulega znacznemu skróceniu i atomy wolniej dyfundują ku ściankom, w drugim — zderzenie ze ściankami wywiera tak słabe działanie relaksacyjne, że można uzyskać wysoki stopień orientacji. Dzięki pokryciu ścianek warstwą ochronną lub użyciu gazu buforującego dezorientacja następuje dopiero po wielu zderzeniach.

oddziaływania prowadzące do relaksacji

Oddziaływania międzyatomowe prowadzące do relaksacji są więc różnorodne. Na podstawie analizy procesów relaksacji stworzono liczne modele opisujące mechanizmy zderzeń. Badanie przebiegów relaksacji przyczyniło się do głębszego poznania tak podstawowego zjawiska fizycznego, jakim jest zderzenie. Odkryto m.in. tworzenie się chwilowych cząsteczek w zderzeniach potrójnych oraz zerwanie w czasie trwania zderzenia wewnątrzatomowego sprzężenia $\vec{I}\vec{J}$ (niezmiernie słabego w porównaniu z potężnymi siłami występującymi w zderzeniu); sprzężenie zostaje przywrócone po zakończeniu zderzenia (z inną, przypadkową wartością F) i dalsza ewolucja stanu atomu następuje przy jego zachowaniu — aż do następnego zderzenia.

natura procesów relaksacji

Oprócz analizy procesów zderzeniowych badanie relaksacji w obecności gazu buforującego pozwala na dokładne wyznaczenie współczynnika dyfuzji atomów zorientowanych w gazie, badanie zaś relaksacji na ściankach komórki pokrytych warstwą ochronną dostarcza wielu informacji o procesie adsorpcji na takiej warstwie.

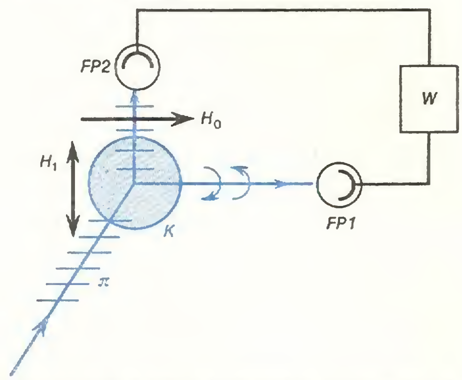
Relaksacje bada się obserwując przebiegi zmian natężenia światła wiązki sondującej w różnych procesach nieustalonych (zanik orientacji po przerwaniu pompowania lub czasowy przebieg narastania orientacji przy pompowaniu niezmiernie słabą wiązką). Skład spektralny i rodzaj polaryzacji wiązki sondującej jest taki sam jak wiązki pompującej, jednak jej natężenie winno być możliwie małe, by uniknąć efektów pompowania przez nią. Stałe czasowe relaksacji T zależą oczywiście od warunków fizycznych. W typowych warunkach doświadczalnych T przybiera wartości od kilku setnych do kilku dziesiątych sekundy.

Gaz buforujący może wnosić pewne dodatkowe zaburzenie. Gdy jego ciśnienie jest wystarczająco duże, by czas między kolejnymi zderzeniami był mniejszy od czasu życia atomów w stanie wzbudzonym, wówczas zderzenia wywołują wyrównanie obsadzeń podpoziomów zeemanowskich w tym samym stanie i wszystkie podpoziomy stanu podstawowego zostają wskutek reemisji równomiernie zapełnione. Do wytworzenia różnicy obsadzeń w stanie podstawowym można wtedy doprowadzić przez intensywne pompowanie, które wydajnie opróżnia jeden z podpoziomów.

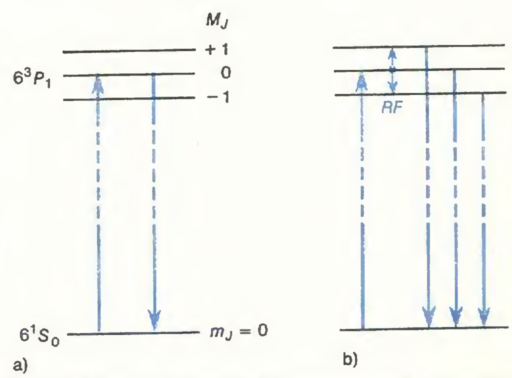
Pompowanie optyczne par atomowych w stanie wzbudzonym. Podwójny rezonans

pompowanie optyczne pary rtęci

Selektywne obsadzenie wybranych podpoziomów zeemanowskich stanu wzbudzonego można uzyskać przez odpowiednią polaryzację światła wzbudzającego. Rysunek 16 przedstawia w znacznym uproszczeniu układ doświadczenia Brossela nad pompowaniem rtęci w stanie rezonansowym, a rys. 17 — schemat pozio-



Rys. 16. Uproszczony schemat doświadczenia nad podwójnym rezonansem: K komórka z parą rtęci, H_0 stałe pole magnetyczne, H_1 oscylujące pole magnetyczne, $FP1$ i $FP2$ fotopowielacze, W wzmacniacz różnicowy



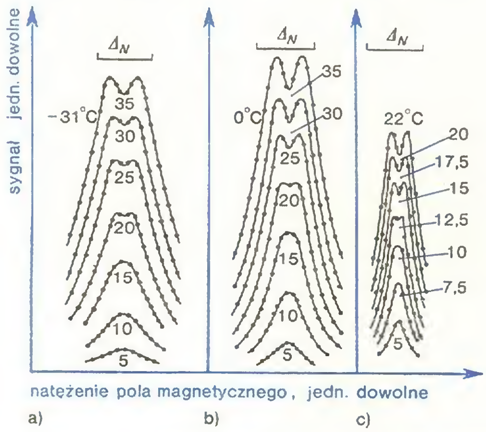
Rys. 17. Schemat poziomów rtęci zaangażowanych przy podwójnym rezonansie (wzbudzenie i reemisja): a) w nieobecności rezonansowego pola oscylującego, b) w wypadku rezonansu (RF częstość radiowa)

mów jednego z jej parzystych izotopów. Komórka kwarcowa wypełniona rtęcią umieszczona jest w stałym polu magnetycznym H_0. Wzbudzenie linią rezonansową o polaryzacji π następuje prostopadle do kierunku pola. Wzbudzona fluorescencja rezonansowa jest wykrywana w dwu kierunkach obserwacji. Wzdłuż pola H_0 wydzielamy składowe o polaryzacji σ, prostopadle do niego składową π. W nieobecności zaburzeń (pola rezonansowego częstości radiowej, zderzeń) z obsadzonego przez wzbudzenie podpoziomu $M_J = 0$ następuje bezpośredni powrót do stanu podstawowego — wysłane promieniowanie ma polaryzację π (równoległą do H_0) i tylko fotodetektor $FP2$ może je wykryć. Zgodnie z prawami przestrzennego rozkładu promieniowania π emisja wzdłuż pola H_0 nie zachodzi (rys. 17a). Jeśli jednak atomy w komórce zostaną poddane działaniu pola oscylującego H_1 o częstości rezonansowej dla przejścia między podpoziomami $M_J = 0 \rightarrow M_J = \pm 1$, wówczas w emisji pojawią się

własności wzbudzonej fluorescencji rezonansowej

podwójny rezonans

składowe σ struktury zeemanowskiej (rys. 17b), co zostanie zarejestrowane przez fotodetektor *FP1*. Gdy ustalimy częstość pola radiowego i będziemy powoli przestrajać H_0 przez obszar rezonansu, to ze wzmacniacza różnicowego otrzymamy sygnał odpowiadający rezonansowi magnetycznemu. Kształt krzywej rezonansu magnetycznego (amplitudę w funkcji H_0) możemy przewidzieć teoretycznie analizując procesy, od których jest on zależny, a więc: wzbudzenie atomu rtęci ze stanu podstawowego 6^1S_0 do stanu rezonansowego 6^3P_1, oddziaływanie w stanie wzbudzonym z polem radiowym wymuszającym przejście $M_J = 0$ do $M_J = \pm 1$, emisję samorzutną po średnim czasie τ i powrót atomu do stanu podstawowego. Funkcja zatem opisująca kształt linii rezonansowej będzie zawierała: zależność od odstrojenia $\omega - \omega_0$ częstości pola radiowego od częstości rezonansowej ω_0, zależność od amplitudy H_1 tegoż pola i zależność od czasu życia τ stanu wzbudzonego. Kształt otrzymanych krzywych znakomicie się zgadza z przewidywanym teoretycznie (rys. 18). Prosty rachunek pozwolił powiązać szerokość połówkową krzywej rezonansowej $\Delta\omega_{1/2}$ w granicznym wypadku dążących do zera amplitud H_1



Rys. 18. Krzywe rezonansowe otrzymane w doświadczeniu Brossela. Punkty — wyniki pomiarów, linie ciągłe — krzywe teoretyczne (jeden punkt rodziny krzywych dopasowano do wyniku doświadczenia). Układy krzywych (a, b, c) odpowiadają różnym gęstościom nasyconej pary rtęci; krzywe każdego układu odpowiadają różnym amplitudom pola częstości radiowej; Δ_N szerokość naturalna linii

z czasem życia stanu wzbudzonego ($\Delta\omega_{1/2} = 2/\tau$). Można więc zmierzyć bezpośrednio szerokość naturalną linii (jak wiemy, w obszarze częstości radiowych rozszerzenie dopplerowskie nie odgrywa żadnej roli) oraz rozszczepienie zeemanowskie, a stąd z kolei — stałą Landégo (stałą g). Oba pomiary doprowadziły do bardzo ciekawych wyników. Okazało się mianowicie, że rozszczepienie zeemanowskie stanu wzbudzonego różni się nieco od tego, które odpowiada doskonałemu sprzężeniu $\vec{L}\vec{S}$:

$$g_{\text{exp}} = 1{,}4383 \pm 0{,}0004; \quad g_{\text{teor}}^{LS} = 1{,}5.$$

Różnica przekracza wielkość błędu pomiarowego i pozwala wnioskować o pewnej niedoskonałości sprzężenia $\vec{L}\vec{S}$ w atomie rtęci w stanie 6^3P_1. Na ten sam fakt wskazuje równocześnie zachodzenie przejścia interkombinacyjnego, łączącego singletowy stan podstawowy z trypletowym stanem rezonansowym, które w sprzężeniu $\vec{L}\vec{S}$ jest silnie wzbronione. Wpływ tego wzbronienia zaznacza się wyraźnie w stosunkowo długim czasie życia stanu 6^3P_1 ($\tau = 1{,}18 \cdot 10^{-7}$ s).

Pomiary rozszczepienia zeemanowskiego i wyznaczenie stałej Landégo z taką dokładnością, na jaką pozwala metoda pompowania optycznego, dostarczają na ogół bardzo dobrych informacji o naturze sprzężenia.

Spójna dyfuzja promieniowania rezonansowego

Przy opisywaniu stanu atomu (ogólniej — elementarnego układu kwantowego, czyli mikroukładu), za pomocą funkcji falowej należy sprecyzować, jakimi wielkościami (obserwablami) jesteśmy zainteresowani. Musimy pamiętać przy tym, że zgodnie z podstawowymi prawami mechaniki kwantowej nie wszystkie obserwable mogą być wyznaczone w jednym pomiarze. W zagadnieniach spektroskopii najczęściej interesuje nas energia stanu i związane z nią wielkości kwadratu całkowitego momentu pędu oraz jednej z jego składowych. Odpowiada to tzw. reprezentacji energetycznej i sprowadza się do rozwiązania równania Schrödingera dla stanu stacjonarnego $\hat{H}\Psi = E\Psi$ (gdzie $\hat{H}$ jest operatorem energii układu). Rozwiązanie tego równania polega na znalezieniu takiego zbioru funkcji stanu ψ_i, by zachodziło: $\hat{H}\psi_i = E_i\psi_i$ (E_i jest liczbą). Funkcje ψ_i nazywamy funkcjami własnymi układu, liczby zaś E_i są to wartości własne energii (przedstawiające jedyne wyniki, jakie możemy otrzymać z pomiaru energii naszego mikroukładu). Można teraz rozróżnić dwie sytuacje w odniesieniu do stanu naszego mikroukładu. Mówimy, że jest on w stanie własnym ψ_i, jeżeli pomiar energii przeprowadzony na nim daje z pewnością wielkość E_i. Jeśli natomiast pomiar energii daje nam wielkość E_i z amplitudą prawdopodobieństwa a_i, wielkość E_j z amplitudą a_j itd., to stan rozważanego układu nazywamy stanem superpozycji i przedstawiamy jako liniową kombinację stanów własnych:

superpozycja stanów kwantowych

$$\Psi = a_1\psi_1 + a_2\psi_2 + \ldots + a_i\psi_i = \sum_i a_i\psi_i,$$

przy tym $\Sigma|a_i|^2 = 1$. Zbiór funkcji ψ_i musi spełniać pewne warunki kompletności (winien tworzyć bazę). Gdy widmo energii własnych układu jest ciągłe, a nie punktowe, definicje stanu superpozycji ulegają prostemu uogólnieniu. Widać wreszcie, że możemy uważać stan własny za szczególny przypadek stanu superpozycji, gdy tylko jedna z wartości $a_i = 1$, pozostałe zaś są równe zeru. Jeśli mikroukład poddany jest zewnętrznemu zaburzeniu zależnemu od czasu, wówczas stan ulega ewolucji i każde $a_i = a_i(t)$. W ogólnym wypadku ewolucje przebiegają w każdym mikroukładzie niezależnie. Istnieją jednak pewne specjalne sytuacje (i te właśnie nas obecnie interesują), gdy pod wpływem zaburzenia zewnętrznego wszystkie mikroukłady należące do określonego zbioru makroskopowego (ansamblu) dokonują ewolucji w pewien skoordynowany sposób. Tego typu ewolucje obserwujemy w zbiorze atomów pompowanych optycznie i poddanych rezonansowemu polu częstości radiowej. Długość fali radiowej wielokrotnie przewyższa rozmiary komórki rezonansowej i można przyjąć, że oddziaływanie między polem i atomem odpowiada w całym obszarze komórki tej samej fazie drgań pola. Wówczas ewolucja $a_i(t)$ we wszystkich atomach zachodzi jednakowo, czyli w uzgodnieniu fazowym (identyczne mikroukłady poddane synchronicznie periodycznemu zaburzeniu). Tego typu zjawiskiem jest np. uzgodniona w fazie precesja atomowych momentów magnetycznych dokoła kierunku pola H_0, prowadząca do wytworzenia wirującej składowej poprzecznej makroskopowego momentu magnetycznego. Ponieważ w opisanym przypadku ewolucja mikroukładów zachodzi w zbiorze przy uzgodnieniu fazowym, będziemy mówili, że w tych warunkach występuje spójność stanów kwantowych.

spójność stanów kwantowych

Omówiony obraz teoretyczny superpozycji stanów kwantowych znajduje piękne potwierdzenie w doświadczeniach nad pompowaniem optycznym w stanach wzbudzonych, w zjawisku zwężenia linii rezonansu magnetycznego w stanie 6^3P_1 atomu rtęci ze wzrostem ciśnienia pary (nawet poniżej szerokości naturalnej).

spójna dyfuzja promieniowania rezonansowego

Od dawna było znane zjawisko uwięzienia promieniowania rezonansowego. Przy podwyższeniu ciśnienia pary foton fluorescencji rezonansowej podlega reabsorpcji (jednorazowej lub kilkakrotnej), co wydłuża czas jego przebywania w komórce. Nie jest to rzeczywiste przedłużenie czasu życia atomu w stanie wzbudzonym i nie może prowadzić do zwężenia linii. Atomy poddane radiowemu polu rezonansowemu znajdują się jednak nie w stanie własnym, lecz w stanie superpozycji

$$\Psi = a_{+1}(t)\,\psi_1 + a_0(t)\,\psi_0 + a_{-1}(t)\,\psi_{-1},$$

gdzie $a_i(t)$ i ψ_i ($i = +1, 0, -1$) oznaczają odpowiednio amplitudy i funkcje własne stanów o $M_J = +1, 0, -1$.

Z ogólnych rozważań wynika, że kwant wysłany przez atom w stanie superpozycji po pochłonięciu go przez inny, identyczny atom wzbudzi go również do stanu superpozycji, a ten z kolei pod wpływem pola radiowego będzie kontynuował ewolucję poprzedniego itd. W tych warunkach, zgodnie z kwantową zasadą nierozróżnialności atomów, nie możemy wskazać, czy to pierwszy atom wykonywał ewolucję w stanie wzbudzonym przez czas odpowiednio dłuższy,

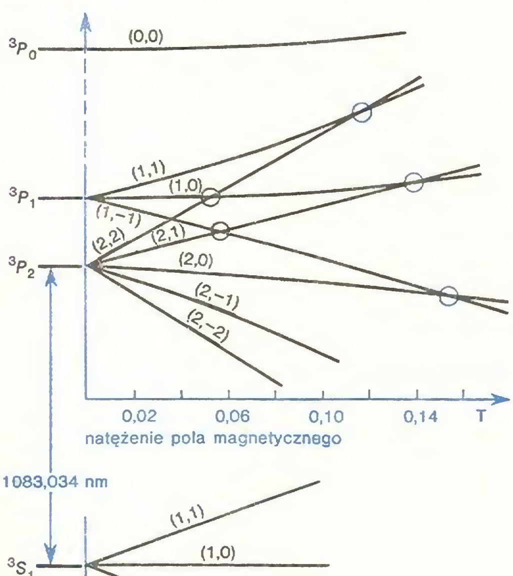
czy też foton przechodził z atomu do atomu, tylko że ich ewolucje były utrzymywane przez rezonansowe pole radiowe w ustalonej fazie. Taka sytuacja w istocie prowadzi do zwężenia linii. Jeśli się teraz przeprowadzi serie pomiarów szerokości linii rezonansu magnetycznego w funkcji ciśnienia pary, to się okaże, że wraz z malejącym ciśnieniem pary szerokość linii wzrasta (!), dążąc do wartości odpowiadającej rzeczywistemu czasowi życia atomu w stanie wzbudzonym (widać to wyraźnie na wykresach przedstawionych na rys. 18).

efekty modulacyjne

Dalsza analiza wyrażeń na natężenie promieniowania wysłanego przez zbiór atomów w takich warunkach prowadzi do wniosku, że gdy emisja następuje w obecności rezonansowego pola częstości radiowej, natężenie promieniowania winno być zmodulowane z częstością Larmora lub podwójną. Te modulacje w istocie były zaobserwowane, zgodnie z przewidywaniami teoretycznymi.

Spójne wzbudzenie stanów można uzyskać nie tylko w obecności radiowego pola rezonansowego. Dobre wyniki dało wzbudzanie bardzo krótkimi (najwyżej nanosekundowymi) błyskami. Szerokość spektralna błysku (impulsu) jest wtedy wystarczająca, by objąć całą badaną strukturę (zeemanowską, nadsubtelną lub nawet subtelną). Wzbudzenie wszystkich atomów następuje właściwie w tym samym momencie (stałe czasowe ewolucji są znacznie dłuższe niż czas wzbudzenia) i zanika wykładniczo z nałożoną modulacją o częstości równej rozszczepieniu stanów (zob. str. 297).

W doświadczeniach nad pompowaniem optycznym w stanie podstawowym rezonansowe pole radiowe wprowadza spójność między podpoziomami zeemanowskimi. Można ją wykryć za pośrednictwem modulacji wiązki poprzecznej i wykorzystać do powiększenia dokładności pomiarów (m.in. magnetometrycznych).



Rys. 19. Rozszczepienia zeemanowskie poziomów energetycznych He należących do $n = 2$ w układzie trypletowym. Stan 2^3S_1 jest metatrwały i może stanowić bazę przejść w układzie trypletowym. Linia $\lambda = 1083{,}034$ nm jest w nim linią rezonansową. Przecięcia poziomów zaznaczone czarnymi okręgami dają sygnały, na podstawie których można obliczyć rozszczepienie stanów $^3P_1 \rightarrow {}^3P_2$, wynoszące 2291,56±0,09 MHz. Odległość między $^3P_1$ i $^3P_0$ obliczona na podstawie analizy teoretycznej rozszczepienia subtelnego stanu ^3P helu, wynosi 29650±280 MHz. Czas życia stanu 2^3P obliczony z szerokości sygnału był również zgodny ze zmierzonym bezpośrednio. Przecięcia poziomów oznaczone okręgami niebieskimi nie spełniają warunku $\Delta m_J = 2$ i takie sygnały nie były obserwowane przy użytej polaryzacji światła

Metoda „przecinania poziomów"

Wzbudzenie atomów do spójnej superpozycji stanów może być osiągnięte wówczas, gdy znajdują się one w polu magnetycznym o stosownie dobranym natężeniu

Rysunek 19 przedstawia diagram energii stanów $2^3P_{0,1,2}$ helu. Każda gałąź wykresu odpowiada stanowi o określonych liczbach kwantowych (J, M_J) i obrazuje zależność energii tego stanu od natężenia pola magnetycznego. Weźmy pod uwagę przecięcie gałęzi $(2, 1)$ i $(1, -1)$ przy $H_c \approx 0{,}057$ T. W tych warunkach dwóm różnym stanom odpowiadać będzie ta sama energia. Jeśli teraz wzbudzimy fluorescencję rezonansową atomów znajdujących się w polu magnetycznym i przejdziemy przez obszar H_c (zmieniając stopniowo natężenie pola), to zaobserwujemy zmianę natężenia fluorescencji w obszarze „przecięcia poziomów". Na podstawie ogólnej teorii zjawiska Zeemana możemy z dobrym przybliżeniem obliczyć zmianę energii stanu wywołaną przez przejście od $H = 0$ do $H = H_c$. Z rys. 19 widać, że suma zmian energii poziomów $(2, 1)$ i $(1, -1)$ daje wprost rozszczepienie poziomów $^3P_1$ i $^3P_2$. Szerokość sygnału (w skali H, którą łatwo przeliczyć na MHz) zależy tylko od szerokości naturalnej poziomów, daje więc informację o czasie życia stanu wzbudzonego. Dzięki swej prostocie metoda wyznaczania tych dwóch ważnych parametrów stanów atomowych jest ostatnio bardzo szeroko stosowana, szczególnie przy badaniach wysokich stanów wzbudzenia atomów trudno dostępnych innymi metodami. Z rys. 19 widać również, że gdy $H = 0$, przecinają się więcej niż dwa poziomy. Sygnał „przecięcia poziomów" w polu zerowym może być również wykorzystany do obliczenia czasu życia stanu wzbudzonego. Ostatnio stosowanie tej metody rozciągnięto również na stany wzbudzone prostych cząsteczek.

badanie stanów wysoko wzbudzonych

Pompowanie nadsubtelne

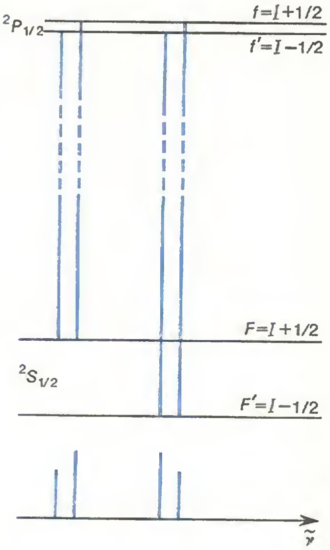
Przy pompowaniu optycznym w stanie podstawowym atomów o niezerowym spinie jądrowym wytwarzamy różnicę obsadzeń nie tylko między podpoziomami zeemanowskimi, lecz również między poszczególnymi składowymi hipermultipletu. Można to wykryć przez zmianę sygnału optycznego po przyłożeniu rezonansowego pola elektromagnetycznego. W metalach alkalicznych częstości rezonansowe $\Delta\nu$ przejść między podpoziomami struktury nadsubtelnej odpowiadają już jednak obszarowi mikrofalowemu (np. $\Delta\nu = 1772$ MHz dla ^{23}Na, $\Delta\nu = 6834{,}7$ MHz dla ^{87}Rb i $\Delta\nu = 9192{,}6$ MHz dla ^{133}Cs). Szczególnie interesujące są wypadki, w których jest możliwe przeprowadzenie pompowania w nieobecności pola zewnętrznego, wówczas bowiem częstość przejścia (tzw. przejścia rezonansowego nadsubtelnego) może być wyznaczona z olbrzymią dokładnością (dzięki niezależności od ewentualnych fluktuacji pól zewnętrznych). Można więc ją przyjąć jako dogodny wzorzec częstości. Po bardzo licznych badaniach i dyskusjach jako definicję sekundy przyjęto w 1967 r. czas równy 9 192 631 770 okresów promieniowania odpowiadającego przejściu między podpoziomami nadsubtelnej struktury $F = 4 \rightarrow F' = 3$ $(M_F = M_{F'} = 0)$ w stanie podstawowym $6^2S_{1/2}$ atomu izotopu cezu ^{133}Cs. Wzorcem pierwotnym jest układ, w którym atomy cezu biegną w postaci promienia atomowego (smukłej, równoległej wiązki atomów) i separacja atomów w stanach F i F' następuje w polu magnetycznym (→ Spektroskopia mikrofalowego rezonansu rotacyjnego), a nadsubtelne pompowanie optyczne wykorzystuje się w precyzyjnych wzorcach wtórnych, w których się stosuje optyczną detekcję rezonansu mikrofalowego. Pompowanie odbywa się przez wzbudzanie jedną z dwu składowych nadsubtelnej struktury linii D_1. Linia D_1 w widmie metali alkalicznych składa się w rzeczywistości

pompowanie w nieobecności pola

wzorce sekundy

z czterech składowych, jednak rozszczepienie nadsubtelne stanu wzbudzonego jest wielokrotnie mniejsze od rozszczepienia stanu podstawowego; w przyrządzie o niezbyt wielkiej zdolności rozdzielczej wydzielamy jedną z dwu par składowych i pompując nią, opróżniamy jeden z podpoziomów nadsubtelnych stanu podstawowego (rys. 20).

struktura nadsubtelna



Rys. 20. Struktura nadsubtelna linii D_1 metalu alkalicznego o spinie jądrowym I

Wielkie możliwości wydajnego pompowania nadsubtelnego otwiera stosowanie strojonych laserów barwnikowych o działaniu ciągłym, a zastosowanie gazu buforującego i wykorzystanie wywołanego przezeń specyficznego efektu zwężenia linii rezonansu nadsubtelnego pozwala uzyskać dokładność przy wyznaczaniu rozszczepienia nadsubtelnego do ułamka Hz (przy częstości $9 \cdot 10^9$ Hz).

powiązanie zegara z wzorcem

Ogólna zasada powiązania częstości z wzorca atomowego z zegarem (w tym wypadku z wysoce stabilnym generatorem elektronicznym, umożliwiającym zsyntetyzowanie częstości wzorcowej i zaopatrzonym w urządzenie do zliczania okresów) polega na porównaniu obu częstości drgań metodą interferencji. Przy bardzo bliskich częstościach interferujących fal otrzymujemy dudnienia, których liczba w sekundzie daje wprost odstrojenie zegara od wzorca atomowego.

Ostatnio podjęto dalsze próby zwiększenia dokładności wzorca częstości. Wykorzystuje się układy laserów o działaniu ciągłym i o częstości stabilizowanej na wybranych liniach widma molekularnego. Uzyskiwanie częstości kombinacyjnych przez zdudnianie wiązek laserowych (dzięki wykorzystaniu nieliniowych zjawisk optycznych) pozwala na nawiązanie częstości optycznych do częstości mikrofalowych i uzyskanie dokładności rzędu 10^{-15}, a stabilności nawet do 10^{-17} (N.P. Chebotaev). Ta dokładność nasuwa poważne problemy związane ze znacznie mniej dokładną definicją innych podstawowych stałych fizycznych.

Zderzenia wymienne

W mieszaninie dwu rodzajów atomów paramagnetycznych A i B (np. Rb i Cs) orientacja układu atomów A uzyskana przez pompowanie optyczne może być w zderzeniach przekazana atomom B. Wynik zderzenia wymiennego jest więc taki, jak gdyby zderzające się atomy wymieniały wzajemnie swe elektrony walencyjne (stąd nazwa: zderzenia wymienne). Symbolicznie można to przedstawić w postaci reakcji $A\uparrow + B\downarrow \rightarrow A\downarrow + B\uparrow$, gdzie strzałki wskazują orientację wektorów spinu elektronu walencyjnego. Ten proces jest, jak widać, pompowaniem atomów B oraz równocześnie jednym z czynników relaksacji atomów A. W ten sposób można uzyskiwać orientację atomów, których bezpośrednie pompowanie jest ogromnie niewygodne, a czasem wręcz niemożliwe. Za pomocą zderzeń wymiennych uzyskano orientację wszystkich izotopów wodoru (^{1}H, ^{2}H, ^{3}H), azotu (^{14}N, ^{15}N), fosforu ^{32}P a także elektronów swobodnych e^-.

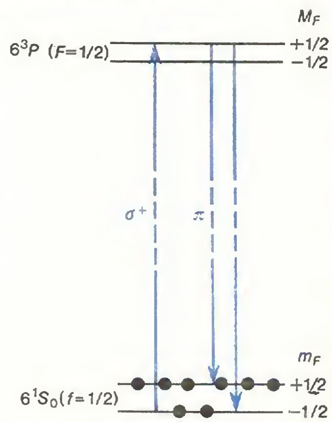
mechanizm wymiany

W rzeczywistości zderzenie wymienne prowadzi do utworzenia układu przejściowego, złożonego ze składników A i B, mającego charakter chwilowej cząsteczki. Jej czas życia, ok. 10^{-12} s, jest wystarczająco długi, by powstało sprzężenie w powłoce elektronowej, i jednocześnie dość krótki na to, by sprzężenia między powłoką elektronową i jądrami atomów zostały zerwane.

Czysto jądrowa orientacja atomów

Przy pompowaniu optycznym (lub przez zderzenia wymienne) atomów obdarzonych spinem jądrowym i mających w stanie podstawowym wypadkowy moment magnetyczny powłoki elektronowej równy zeru mamy do czynienia z ciekawym przypadkiem czysto jądrowej orientacji. Przykładem mogą tu służyć ^{3}He ($I = {}^1/_2$, stan podstawowy 1^1S_0 o zerowym momencie orbitalnym i zerowym wypadkowym spinie elektronowym) oraz ^{199}Hg ($I = {}^1/_2$, stan podstawowy 6^1S_0). Rtęć można pompować czysto optyczną metodą (rys. 21), wykorzystując przy tym koincydencję nadsubtelnej składowej A ($f = {}^1/_2 \rightarrow F = {}^1/_2$) i linii parzystego izotopu ^{204}Hg. Pompowanie ^{3}He jest znacznie bardziej złożone: za pomocą słabego wyładowania wytwarzamy w He atomy w stanie metatrwałym 2^3S_1 i pompujemy je optycznie linią 1083 nm (rys. 19). Przez sprzężenie między powłoką elektronową i spinem jądrowym pompowanie wywołuje częściową orientację

pompowanie jądrowe



Rys. 21. Pompowanie jądrowe izotopu ^{199}Hg. Na rys. zaznaczono strukturę zeemanowską tylko składowej A

jądrową. W zderzeniach wymiennych z atomami helu w stanie podstawowym atomy zorientowane przekazują partnerom zderzenia energię wzbudzenia i orientację powłoki elektronowej, natomiast orientacja jądrowa pozostaje niezaburzona. Wypełniona powłoka elektronowa o konfiguracji $1s^2$ stanowi znakomitą ochronę jądra przed oddziaływaniami zaburzającymi, atomy ^{3}He mogą więc zachowywać orientację jądrową przez długie godziny, stanowiąc w ten sposób zorientowany zbiór przygotowany do doświadczeń fizyki jądrowej.

Atomowa spektroskopia laserowa

Omówimy tu kilka zasadniczych kierunków, w jakich się rozwija zastosowanie laserów w spektroskopii atomowej.

użycie przestrajanych laserów

1) Użycie lasera jako przestrajanego źródła promieniowania w badaniach absorpcyjnych. Szerokość spektralna źródła laserowego może być bardzo mała (do kilku MHz). Wiązka z lasera zostaje podzielona za pomocą półprzezroczystej płytki na dwie, z których jedna przechodzi przez badany ośrodek, druga zaś jest wiązką odniesienia. Odpowiednie wywzorcowanie pozwala nałożyć na widmo absorpcji skalę częstości.

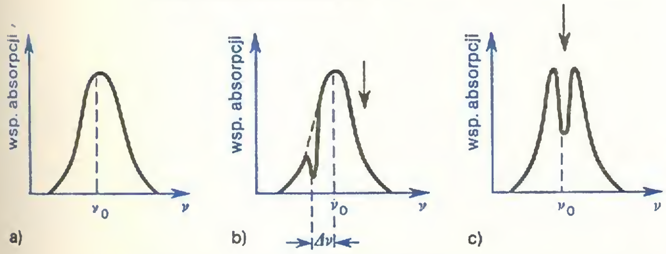
2) Użycie lasera impulsowego do obserwacji zaniku fluorescencji i dudnień kwantowych, powstających przy spójnym wzbudzeniu nierozszczepionej struktury, i wnioskowaniu stąd o wielkości rozszczepienia poziomów i o czasach ich życia.

atomowa spektroskopia błyskowa

Rozszczepienie subtelne stanów atomowych szybko maleje ze wzrostem głównej liczby kwantowej n i liczby kwantowej momentu orbitalnego L, toteż nie można go obserwować w badaniach wyższych stanów prowadzonych zwykłymi metodami spektroskopowymi (rozszczepienie staje się znacznie mniejsze od szerokości dopplerowskiej linii, przeto struktury subtelnej w ogóle nie można zaobserwować). Wykorzystanie zjawiska dudnień kwantowych pozwoliło na przeprowadzenie bardzo dokładnych pomiarów rozszczepień subtelnych (np. S. Haroche zmierzył tą metodą rozszczepienia stanów $D_{3/2}$–$D_{5/2}$ w sodzie i w cezie, od stanu $9D$ do $16D$). Idea doświadczenia (prosta w schematycznym przedstawieniu, lecz bardzo trudna i skomplikowana w realizacji) polegała na tym, że atomy badanej pary wzbudzano ze stanu podstawowego krótkimi (najlepiej nanosekundowymi) błyskami ze strojonego lasera barwnikowego. Stan podstawowy w metalach alkalicznych jest stanem S, a więc przejście bezpośrednie do stanu D jest wzbronione przez reguły wyboru dla liczby kwantowej L. Jednak (i to jest bardzo wygodne) wiązka laserowa o dużej mocy może wywołać przejście dwufotonowe — każdy z fotonów niesie kwantową jednostkę momentu pędu $\pm\hbar$ i przy jednoczesnej absorpcji dwu fotonów możliwe są przejścia ze zmianą liczby kwantowej L o dwa lub bez jej zmiany ($\Delta L_{2fot} = 0$ lub 2). Musi być spełniony również warunek zachowania energii $h\nu_{S-D} = h\nu_1 + h\nu_2$, przy czym w najprostszym wypadku $\nu_{S-D} = 2\nu_L$ (wszystkie fotony z lasera mają częstość ν_L). Krótki błysk wzbudza atomy do stanu superpozycji i fluorescencja obserwowana prostopadle do wzbudzenia gaśnie wykładniczo, co daje bezpośrednią informację o czasie życia τ stanu wzbudzonego, a krzywa zaniku fluorescencji jest modulowana z częstością odpowiadającą częstości dudnień. Gdy na sygnał modulowanego zaniku nałożymy sygnały czasu, z częstości modulacji wywnioskujemy bezpośrednio o częstości rozszczepienia stanów.

wykorzystanie dudnień kwantowych

Zauważymy wreszcie, że jeśli struktura jest bardziej złożona (np. wysokie stany wzbudzenia układu trypletowego), to modulacja ma postać złożoną i widmo częstości znajdujemy jako transformatę fourierowską czasowego przebiegu modulacji. Zarówno oddzielenie modulacji od wykładniczego zaniku, jak też jej fourierowska analiza są dokonywane przez urządzenia elektroniczne.



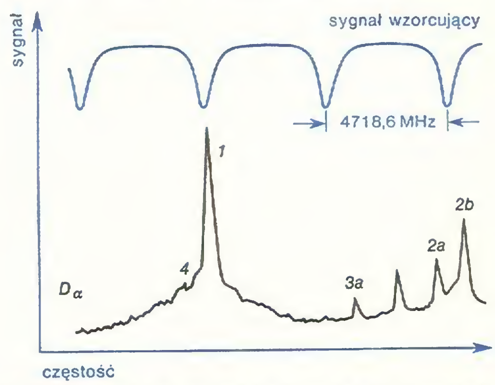
Rys. 22. Zasada metody spektroskopii nasyceniowej: a) linia absorpcyjna o rozkładzie dopplerowskim; b) zmiana absorpcji wskutek nasycenia w odległości $\Delta\nu$ od środka linii ν_0; c) zmiana absorpcji w środku linii wskutek nasycenia; strzałki oznaczają te częstości wiązki sondującej, przy których następowałyby efekty nasycenia (wiązka sondująca biegnie w kierunku przeciwnym do wiązki nasycającej)

3) Badanie wąskich struktur metodami spektroskopii nasyceniowej; eliminują one rozszerzenie dopplerowskie i umożliwiają rejestrację widma szerokości takiej jak szerokość spektralna wiązki sondującej.

spektroskopia nasyceniowa

Idea metody jest objaśniona na rys. 22. Przy prześwietlaniu ośrodka pochłaniającego wiązką z lasera o bardzo małej szerokości spektralnej (np. rzędu kilku lub kilkunastu MHz) padające promieniowanie jest pochłaniane tylko przez te atomy, których ruch wywołuje przestrojenie dopplerowskie częstości przejścia ν_0 atomu w spoczynku do częstości ν wiązki z lasera. Przepuszczenie przez ośrodek wiązki z lasera o dużym natężeniu może wywołać znaczne zmniejszenie liczby atomów zdolnych do jej pochłaniania (następuje nasycenie absorpcji, rys. 22b). Wykorzystajmy teraz do sondowania osłabioną wiązkę laserową. Przesyłamy ją przez kuwetę absorpcyjną w kierunku przeciwnym do wiązki nasycającej. Jeżeli częstość wiązki nasycającej nie przypadnie akurat w środku badanej linii (co odpowiada jej absorpcji przez atomy o zerowej składowej prędkości ruchu termicznego wzdłuż kierunku wiązki laserowej), to nasycenie absorpcji nie będzie odpowiadało częstości powracającej wiązki sondującej i nie wywrze żadnego wpływu na jej osłabienie przy przejściu przez kuwetę. Dopiero dokładne dostrojenie wiązki nasycającej do środka linii badanej będzie sygnalizowane przez ostre maksimum wiązki sondującej (rys. 22c).

Nawet względne pomiary ilościowe rozszczepień struktury tą metodą są trudne i wymagają wzorcowa-



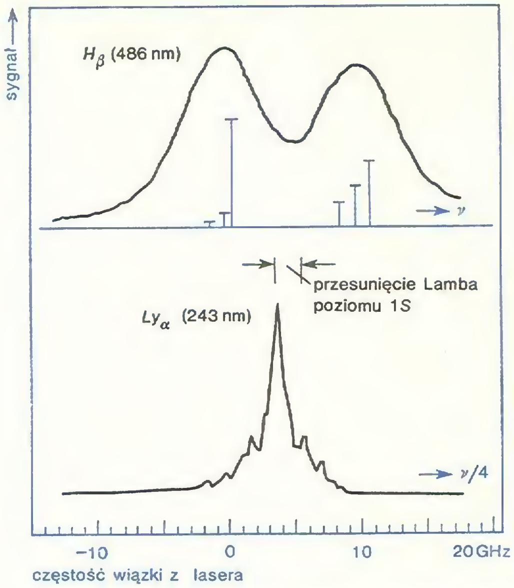
Rys. 23. Struktura linii D_α serii Balmera deuteru. Oznaczenia składowych odpowiadają przejściom wskazanym na rys. 5; sygnał wzorcujący pochodzi ze stabilizowanego lasera He–Ne (wg T. W. Hänscha, A. L. Schawlowa i in.)

nia układu za pomocą linii o długości fali nawiązanej bezpośrednio do kryptonowego wzorca długości. Pokonanie tych trudności daje jednak w rekompensacie pomiar z dokładnością nigdy przedtem nie osiąganą. Strukturę linii D_α serii Balmera deuteru przedstawia rys. 23.

przesunięcie Lamba w wodorze

Jednym ze szczytowych osiągnięć metody spektroskopii nasyceniowej (dwufotonowej) był bezwzględny pomiar przesunięcia Lamba stanu podstawowego w wodorze. Wiązka z lasera barwnikowego po podwojeniu częstości do $\lambda \approx 243{,}0$ nm w nieliniowym krysztale (→ Optyka nieliniowa) wywoływała dwufotonową absorpcję w wodorze, powodując przejście $1S \rightarrow$ $\rightarrow 2S$, które można obserwować pośrednio przez śledzenie indukowanej w zderzeniach ($2S \rightarrow 2P$) linii Ly_α o $\lambda = 121{,}5$ nm. Przez wzbudzenie dwufotonowe wiązkami przeciwbieżnymi zwężono spektralny obszar wzbudzenia poniżej 2% szerokości dopplerowskiej linii. Równocześnie podstawowa częstość lasera wzbudza linię H_β ($2S, P \rightarrow 4S, P, D$; $\tilde{\nu}_{Ly_\alpha} = 4\tilde{\nu}_{H_\beta}$). T. W. Hänsch otrzymał niezwykle dokładne wyznaczenie częstości przejścia Ly_α rejestrując w czasie przestrajania lasera przez obszar absorpcji dwufotonowej równocześnie przebieg zmian absorpcji linii H_β (wzorcując zatem bardzo dokładnie częstość podstawową ν_0). Jest ona o $8{,}6 \pm 0{,}8$ GHz dla wodoru

i 8,3±0,3 GHz dla deuteru mniejsza od przewidywanej teoretycznie częstości linii Ly_α, bez uwzględnienia przesunięcia Lamba poziomu $1S$ (rys. 24).



Rys. 24. Wyznaczenie przesunięcia Lamba poziomu $1S$ w deuterze. Wykres górny — zapis struktury linii β serii Balmera; wykres dolny — krzywa dwufotonowego wzbudzenia linii α serii Lymana; odcinki — teoretyczna struktura linii H_β. Skala dolnego wykresu, a więc i zaznaczonego przesunięcia Lamba poziomu $1S$, w $\nu/4$ (wg T. W. Hänscha, A. L. Schawlowa i in.)

dwufotonowe wzbudzenie bez rozszerzenia dopplerowskiego

4) Dwufotonowe wzbudzenie z eliminacją rozszerzenia dopplerowskiego (B. Cagnac) — metoda o bardzo dużej zdolności rozdzielczej. Jeśli na komórkę K, zawierającą np. parę sodu, pada wzdłuż jej osi silna i bardzo wąska spektralnie wiązka z przestrajanego lasera o częstości ν bliskiej ν_0 (przy tym $2\nu_0$ jest częstością przejścia np. $3^2S_{1/2}$–$3^2D_{1/2}$), to staje się możliwe dwufotonowe wzbudzenie atomów sodu wprost do stanu $3^2D_{1/2}$ (ogólnie do n^2D_J lub $n^2S_{1/2}$). Jednak nawet przy dostrojeniu lasera dokładnie do częstości ν_0 wzbudzenie jest słabe, bowiem pochłaniać mogą tylko te atomy, których składowe prędkości bezładnego ruchu cieplnego $v_x = 0$. Gdy częstość wiązki $\nu' \neq \nu_0$, pochłaniać mogą atomy dla których

$$\nu' = \nu_0(1+v_x/c).$$

W układzie odniesienia związanym z tymi atomami „widziana" przez nie częstość wiązki odbitej od zwierciadła Z, ustawionego za kamórką, wynosi

$$\nu'' = \nu_0(1-v_x/c).$$

Przy jednoczesnym oddziaływaniu obu wiązek (rezonans) energia pochłonięta (równa $h\nu_i - h\nu_f$) wyniesie

$$h\nu_0\left(1+\frac{v_x}{c}\right)+h\nu_0\left(1-\frac{v_x}{c}\right) = 2h\nu_0.$$

Warunek zgodności energii przejścia dwufotonowego jest jak widać spełniony dla wszystkich atomów niezależnie od ich prędkości, przy czym prawdopodobieństwo wzbudzenia przez dwa fotony biegnące w kierunkach przeciwnych jest czterokrotnie większe niż gdy fotony poruszają się w tym samym kierunku. Nastąpi więc bardzo silne wzbudzenie dwufotonowe; szerokość spektralna sygnału jest określona przez szerokość spektralną wiązki laserowej.

badanie struktury subtelnej i nadsubtelnej stanów rydbergowskich

Ważne zastosowanie tej metody to: a) Badania struktury subtelnej bardzo wysoko wzbudzonych stanów atomowych (tzw. stanów rydbergowskich) przy n rzędu kilkudziesięciu. Atom ma wówczas osobliwą stiukturę: na zewnątrz rdzenia o rozmiarach 10^{-8} cm rozciąga się chmura elektronu walencyjnego na obszar tysiące razy większy. B. Stoicheff zbadał w ten sposób rozszczepienie poziomów struktury subtelnej stanów D w rubidzie aż do $n = 65$ ($\Delta 65\ D_{3/2,\ 5/2}) = 35$ MHz, zaś położenie tych stanów aż do $n = 85$ (0,16 cm^{-1} poniżej granicy jonizacji). Czasy życia atomów w stanach rydbergowskich są bardzo długie (proporcjonalne do n^3), nie jest więc możliwa bezpośrednia obserwacja fluorescencji z tych stanów. Podstawową metodą badania jest wyznaczanie energii jonizacji przez badanie elektronów wyzwolonych w tym procesie. b) Badania struktury nadsubtelnej wysoko wzbudzonych stanów (rozszczepiono struktury odległe o kilkadziesiąt MHz i wyznaczono przesunięcie izotopowe linii). c) Badania rozszerzenia ciśnieniowego linii widmowych sodu przy bardzo małych ciśnieniach gazu zaburzającego (wyniki bardzo ważne przy analizie procesów oddziaływania w zderzeniach). d) Uzyskanie możliwości selektywnego obsadzania wybranego poziomu w układzie złożonym z wielu bardzo blisko leżących poziomów. Ma to znaczenie przy badaniach mechanizmów zderzeniowych przekazywania wzbudzenia, a także przy realizacji metody rozdzielania izotopów przez selektywne wzbudzanie optyczne.

rozszerzenie ciśnieniowe linii widmowych

Spektroskopia wiązka–tarcza

Omówione w poprzednich rozdziałach metody stosuje się do badań stanów podstawowych i wzbudzonych atomów neutralnych. Badania widm jonów są znacznie bardziej skomplikowane (szczególnie przy wielokrotnej jonizacji) wskutek trudności w wytworzeniu i zmagazynowaniu dostatecznej liczby jonów w obszarze obserwacji oraz w uzyskaniu wielokrotnej jonizacji. Inne trudności w badaniach atomowych pochodziły z ograniczonej rozdzielczości czasowej fotodetektorów i układów elektronicznych. Wiele z nich można przezwyciężyć dzięki wprowadzonej przed kilku laty metodzie spektroskopowej: wiązka–tarcza (S. Bashkin, H.J. Andrä).

W tej metodzie jony badanych atomów zostają przyspieszone w akceleratorze (np. van de Graaffa lub liniowym) i zależnie od rodzaju doświadczenia uzyskują energię 100 keV–kilku MeV (nabywają one prędkość 1–10 mm/ns). Przyspieszone jony przechodzą następnie przez niesłychanie cienką folię węglową (ok. 10 µg/cm²). W czasie przejścia trwającego $\Delta t \approx$ $\approx 10^{-2}$ ps zachodzi oddziaływanie z folią i dalsza jonizacja, prowadząca niejednokrotnie do całkowitego niemal odarcia atomu z elektronów (wytworzone zostają jony dziesięcio-, a nawet kilkunastowartościowe, np. Ar^{17+}).

oddziaływanie przyspieszonych jonów z folią

Spektroskopia wiązka–tarcza dostarczyła niezwykle interesujących wyników. Okazało się możliwe połączenie z nią specjalnych metod omówionych poprzednio i uzyskanie niezmiernie wysokiej precyzji pomiarów. Ponadto możliwe jest osiągnięcie wzbudzenia bardzo wysokich stanów (praktycznie dowolnych). Przy pomiarach tą metodą nie ma zaburzeń wprowadzonych przez zderzenia. Badania można przeprowadzać w ogromnie szerokim obszarze widma (od promieni rentgenowskich do czerwonego krańca widma widzialnego). Gdy wzbudzamy stany o liczbie kwantowej $J \geqslant 1$, występują zależne od $|M_J|$ różnice obsadzeń, co daje możliwość prowadzenia doświadczeń z rezonansem magnetycznym i jego optyczną detekcją. Ponieważ czas wzbudzenia Δt jest rzędu 10^{-2} ps, a nawet 10^{-3} ps, otrzymujemy spójne wzbudzenie blisko leżących stanów i modulowany zanik wysyłanego promieniowania. Szybkość wiązki atomowej wynosi ok. 10^8 cm/s, co zapewnia rozdzielczość pomiarów czasowych lepszą niż 0,1 ns.

Pierwsze badania widm metodą wiązka–tarcza przyniosły ogromny materiał, w którym oprócz linii znanych z pomiarów metodami tradycyjnymi znaleziono ogromną liczbę linii nowych.

badania widm atomów zjonizowanych

Wychodzące z folii wiązki jonów o różnej krotności rozdzielano w separatorach. Badając świecenie rozdzielonych grup, można było przypisać nowo odkryte linie określonym jonom. Niejednokrotnie dodatkową informację przynosiło wyznaczenie czasu gaśnięcia różnych linii. Z dużą dozą prawdopodobieństwa można przypuszczać, że linie o tym samym czasie gaśnięcia wychodzą z tych samych stanów.

Badaniom poddawano systematycznie pewne grupy jonów, a więc np. widma jonów o strukturze izoelektronowej: LiI, BeII, BIII, CIV itd. (widmo atomu obojętnego oznacza się cyfrą rzymską I, jonu jednowartościowego — cyfrą II; tak więc np. CV oznacza widmo, lub układ termów, węgla czterokrotnie zjonizowanego, ArXVIII — widmo argonu zjonizowanego siedemnastokrotnie, Ar^{17+}), o strukturze homologicznej: NeII, ArII, KrII, XeII. Ważnym elementem tych badań było wyznaczanie prawdopodobieństw przejść dla sekwencji jonów o strukturze izoelektronowej i porównywanie wyników doświadczalnych z przewidywaniami teoretycznymi, opartymi na przybliżonych obliczeniach funkcji falowych. Wyniki mogły ocenić dobroć przybliżenia.

badanie czasów życia wzbudzonych jonów

Metoda pomiaru czasu życia w wiązce wzbudzonych atomów jest stosunkowo prosta i polega na pomiarze gaśnięcia badanej linii wzdłuż wiązki, w funkcji odległości od folii. Układ spektrometryczny i pole obserwacji są stałe, a folię przesuwa się skokami, (ich wielkość waha się od 0,05–0,01 mm zależnie od warunków. Te skoki połączone są z przestrajaniem kanałów analizatora układu elektronicznego gromadzącego dane pomiarowe.

Pomiar przesunięcia Lamba przy dużych Z

Wyrażenie na rozszczepienie Lamba, o którym była mowa na str. 287, jest przybliżone. W rzeczywistości jest ono opisane przez podwójne rozwinięcie względem α i (αZ). W wodorze wyrazem znaczącym (mającym udział przekraczający 99% wielkości rozszczepienia) jest wyraz zawierający $\alpha\,(\alpha Z)^4$. Wyrazy wyższego rzędu można więc pominąć. Tymczasem w jonach o większym Z (np. O^{7+}), w stanie o $n = 2$, udział wyrazów wyższego rzędu wzrasta niemal do 10%; zbadanie ich poprawności i zgodności z doświadczeniem staje się rzeczą bardzo interesującą. Bezpośredni pomiar rozszczepienia stanów przy użyciu metody Lamba i Retherforda w jonach wodoropodobnych o większym Z napotyka jednak ogromne trudności. Jeżeli nawet pominąć trudności związane z ogniskowaniem szybkich jonów w polach magnetycznych, to i tak w wypadku ciężkich jonów będziemy mieli do czynienia z bardzo znacznymi częstościami rezonansowymi między rozszczepionymi stanami, leżącymi w bardzo niedogodnym obszarze widma (przesunięcie S wynosi 783 GHz dla C^{5+}, a 2200 GHz dla O^{7+}; jest to obszar dalekiej podczerwieni). Do wyznaczenia S potrzebna więc będzie inna metoda, która wykorzystuje zjawisko mieszania się stanów $^2S_{1/2}$ i $^2P_{1/2}$, gdy atomy wiązki adiabatycznie wchodzą w pole elektryczne. (Termin „przejście adiabatyczne" oznacza tu, że czas narastania pola elektrycznego jest bardzo długi w porównaniu z okresem przejścia $2^2S_{1/2} \to 2^2P_{1/2}$). Zmieszanie stanów wywołuje skrócenie czasu życia poprzednio metatrwałego stanu $2^2S_{1/2}$. To skrócenie zależy od natężenia pola elektrycznego E, a także od wielkości przesunięcia Lamba S i wyraża się wzorem:

$$\frac{1}{\tau_{2S'}} = \frac{1}{\tau_{2P}}\,\frac{|V(E)|^2}{\hbar^2(S^2+\frac{1}{4}\tau_{2P}^2)},$$

gdzie $\tau_{2S'}$ jest czasem zaniku stanu $2S$ po wprowadzeniu atomu w pole elektryczne, τ_{2P} — czasem zaniku promienistego stanu $2P$, $V(E)$ — funkcją natężenia pola elektrycznego wyrażającą zaburzenie stanu $^2S_{1/2}$. Przesunięcie S wyznacza się z pomiaru czasu zaniku $\tau_{2S'}$ linii Ly_α, wysłanej ze stanu $2S'$, w funkcji natężenia pola elektrycznego E. Czas τ_{2P} jest znany z licznych doświadczeń. Wyniki prowadzą do doskonałej zgodności z przewidywaniami teoretycznymi:

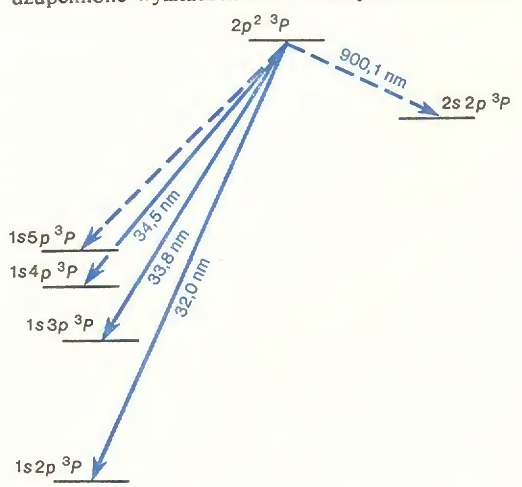
$$S(C^{5+})_{teor} = 783{,}68 \pm 0{,}25 \text{ GHz},$$
$$S(C^{5+})_{exp} = 780{,}1 \pm 8{,}0 \text{ GHz},$$
$$S(O^{7+})_{teor} = 2205{,}17 \pm 1{,}56 \text{ GHz},$$
$$S(O^{7+})_{exp} = 2202{,}7 \pm 11 \text{ GHz}.$$

badanie wzbudzeń wieloelektronowych

Zastosowanie metody wiązka–tarcza umożliwiło badanie wzbudzeń wieloelektronowych (wzbudzone dwa — lub więcej — elektrony w stosunku do konfiguracji odpowiadającej stanowi podstawowemu atomu). Systematyczne badania wzbudzenia dwuelektronowego rozpoczęto w 1969 r. (W.S. Bickel i M. Dufay). W wiązce poza folią znaczna część atomów znajdowała się w stanie dwuelektronowego wzbudzenia. Widma wysyłane przez te atomy zawierały ogromną liczbę linii nie znanych uprzednio. Dotychczasowe badania obejmują zaledwie kilka najlżejszych pierwiastków i jonów, np. H^-, HeI i układy izoelektronowe LiII, BIV oraz LiI i układy izoelektronowe BeII, CIV. Napotyka się wiele różnych trudności. Jedną z najpoważniejszych stanowi proces autojonizacji. Gdy energia stanu wzbudzenia dwuelektronowego przewyższa energię jonizacji atomu przy wzbudzeniu jednoelektronowym, wówczas w wyniku oddziaływania prowadzącego do mieszania stanów o różnych konfiguracjach (tu dyskretnego stanu wzbudzonego dwuelektronowo i kontinuum jonizacji) — bardzo się zwiększa prawdopodobieństwo przejścia bezpromienistego do stanu kontinuum powyżej granicy jednoelektronowej jonizacji, połączone z emisją elektronu. Proces ten może wywołać skrócenie czasu życia stanu wzbudzenia dwuelektronowego tak znacznie (do 10^{-14} s), że inne przejścia z niego będą zupełnie nieobserwowalne. Pomiary utrudnia fakt, że znaczna część przejść ze wzbudzenia dwuelektronowego leży w obszarze bardzo dalekiego, próżniowego nadfioletu, graniczącego z obszarem miękkich promieni rentgenowskich, a nawet zachodzącego na ten obszar. Przejścia autojonizacyjne są jednak częściowo ograniczone przez reguły wyboru i to sprawia, że wiele stanów ma czasy życia już uchwytne w pomiarach (10^{-10}–10^{-9} s).

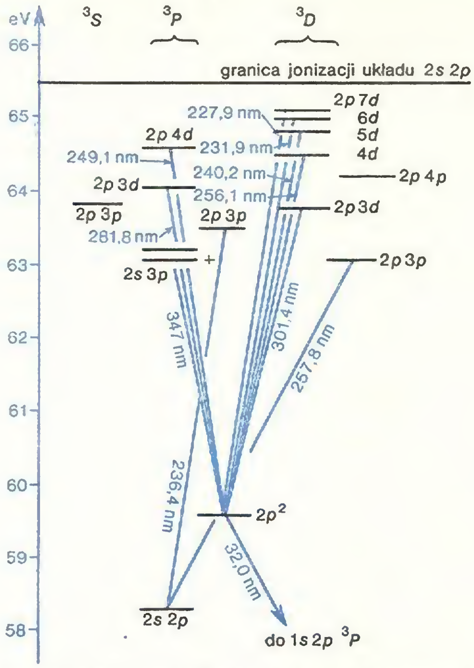
badanie wzbudzeń dwuelektronowych

W badaniu wzbudzeń dwuelektronowych otrzymano dwa typy widm: jedne z nich odpowiadają przejściom od konfiguracji wzbudzenia dwuelektronowego do poziomów dyskretnych wzbudzenia jednoelektronowego (rys. 25), inne — przejściom między różnymi, nie podlegającymi autojonizacji stanami układu wzbudzenia dwuelektronowego (rys. 26). Pomiary położeń linii widmowych oczywiście były uzupełnione wyznaczeniem ich natężeń i czasów ży-



Rys. 25. Przejścia w HeI między stanami wzbudzenia dwu- i jednoelektronowego: linią przerywaną oznaczono przejścia dozwolone, lecz nie obserwowane (wg H.G. Berry'ego)

cia, co w metodzie spektroskopii wiązka–tarcza nie nastręcza szczególnych trudności, aż do czasów rzędu 10 ps. Przyporządkowanie zaobserwowanych



Rys. 26. Przejścia w układzie trypletowym między stanami wzbudzenia dwuelektronowego w HeI zaobserwowane w doświadczeniach wiązka–tarcza

linii przejściom między określonymi stanami wymagało trudnej i wnikliwej analizy teoretycznej. Trzeba było przyjąć model opisujący konfigurację i oddziaływania w powłoce elektronowej, przeprowadzić przybliżone obliczenia funkcji falowych i policzyć stąd energie stanów i czasy życia. Na ogół otrzymano dobrą zgodność otrzymanych wyników dla HeI, LiII i LiI, których diagramami poziomów przy różnych konfiguracjach wzbudzenia dwuelektronowego już dysponujemy. Obliczenia dotyczące układów o większej liczbie elektronów dają gorszą zgodność z doświadczeniem.

widma elektronów z autojonizacji

Bardzo ważnym uzupełnieniem badań spektroskopowych jest badanie widma energetycznego elektronów pochodzących z procesu autojonizacji. Ich energia składa się z dwóch części: energii związanej z ruchem atomów w wiązce i energii związanej z procesem autojonizacji z określonego stanu wzbudzenia dwuelektronowego. Po oddzieleniu pierwszej części energii otrzymujemy widmo energetyczne elektronów z procesu autojonizacji. Jego kolejne maksima odpowiadają energii kolejnych stanów wzbudzenia dwuelektronowego.

Wzbudzenie wiązka–laser

Mimo wielu niewątpliwych zalet metoda wzbudzania atomów w folii ma pewne wady: np. nieselektywne jednoczesne wzbudzanie atomów do różnych stanów, a nawet produkcja wielu różnych stopni jonizacji, lub efekty kaskadowe (atom osiąga kaskadowo badany stan) w zaniku stanów, utrudniające analizę wyników pomiarów. Często można uniknąć tych efektów wzbudzając przyspieszone w akceleratorze jony wiązką laserową. Metoda ta pozwala na celowy wybór szerokości spektralnej wiązki z lasera, na precyzyjne dostrojenie częstości przez regulację kąta nachylenia wiązki laserowej względem wiązki atomów (dostrojenie dopplerowskie). Omówione poprzednio pomiary można w ten sposób przeprowadzić ze znacznie zwiększoną dokładnością; dotyczy to np. pomiaru czasu zaniku fluorescencji z określonych stanów prostych, badania efektów modulacyjnych (przy użyciu do wzbudzenia wiązki laserowej o takiej szerokości spektralnej, która by wystarczyła do objęcia wszystkich poziomów badanej struktury, pomiaru nadsubtelnej struktury dolnego stanu przez stopniowe przestrajanie częstości wzbudzającej zmianą nachylenia wiązki laserowej względem wiązki atomów i obserwację efektywności wzbudzenia.

stała Rydberga

Dokładność otrzymanych tą metodą wyników należy do szczytowych osiągnięć nie tylko spektroskopii, lecz badań fizycznych w ogóle. Pomiary czasu życia są obarczone błędem rzędu 0,02 ns; rozszczepienie poziomów w stanie wzbudzonym można mierzyć z dokładnością do ułamka MHz. Prowadzi to do dokładniejszego wyznaczenia wielkości stałych powszechnych; ostatnie pomiary stałej Rydberga dały wartość $R = 109\ 737{,}3143(4)\ \mathrm{cm}^{-1}$.

H.A. Enge, M.R. Wehr, J.A. Richards *Wstęp do fizyki atomowej*, Warszawa 1981; W. Hanle, H. Kleinpoppen *Progress in Atomic Spectroscopy*, A i B Plenum Press... 1979; D. Kunisz *Fizyczne podstawy emisyjnej analizy widmowej*, Warszawa 1973; Z. Leś *Wstęp do spektroskopii atomowej*, Kraków 1969; artykuły w Postępach Fizyki: 9, 495 (1958); 11, 379 (1960); 12, 533 (1961); 13, 27 (1962); 13, 41 (1962); 19, 557 (1968); 18, 287 (1968); 21 209 (1970); 21, 511 (1970); 21, 543 (1970); 25, 497 (1974); 26, 303 (1975); 27, 297 (1976); 28, 29 (1977); 28, 167 (1977); 29, 3 (1978); 29, 419 (1978).

Spektroskopia molekularna

Podstawowe pojęcia spektroskopii molekularnej

Jerzy Prochorow

Zanim przejdziemy do szczegółowego omówienia zakresu i przedmiotu badań spektroskopii molekularnej, poświęcimy nieco uwagi ogólnym cechom cząsteczek i promieniowania elektromagnetycznego, oraz metodom badania ich wzajemnego oddziaływania.

Energia cząsteczek

Jeżeli izolowana cząsteczka ma pewną energię, związaną np. z ruchem jaki wykonuje, lub z określonym oddziaływaniem pomiędzy jej elementami, to mówimy, że cząsteczka znajduje się w określonym stanie energetycznym, a ilość energii, jaką ma, określa ten stan energetyczny.

Próbując określić energię cząsteczki stosujemy zazwyczaj uproszczenie polegające na założeniu, że poszczególne ruchy i oddziaływania w obrębie cząsteczki są od siebie niezależne, a suma ich energii stanowi całkowitą energię cząsteczki. Można wyróżnić cztery podstawowe źródła energii cząsteczki.

translacja cząsteczki

Pierwszym źródłem energii jest ruch translacyjny (postępowy) cząsteczki jako całości. Cząsteczka o masie M poruszająca się z prędkością v w danym kierunku ma energię kinetyczną $E_{trans} = Mv^2/2$. Prędkość, a więc i energia kinetyczna cząsteczki w gazie zależy od temperatury gazu, przy czym z kinetyczno-molekularnej teorii gazów wynika, że średnia energia kinetyczna ruchu translacyjnego jest równa $E_{trans} = {}^3/_2 kT$ (k — stała Boltzmanna, T — temperatura).

Energia ta, w odróżnieniu od innych rodzajów energii cząsteczki, zmieniać się może w sposób ciągły (nie przybiera dyskretnych wartości, albo jak mówimy, nie jest skwantowana).

rotacja cząsteczki

Oprócz ruchu postępowego cząsteczka może rotować, tzn. obracać się jako całość wokół określonej w przestrzeni osi, co jest kolejnym źródłem energii cząsteczki. Jeżeli moment bezwładności względem wybranej osi cząsteczki izolowanej jest równy I, to energia ruchu rotacyjnego

$$E_{rot} = \frac{h^2}{8\pi^2 I} J(J+1).$$

Liczba J nazywa się kwantową liczbą rotacji i może przybierać wartości $J = 0, 1, 2, 3, \ldots$. A zatem energia rotacyjna może przybierać tylko pewne (nie wszystkie możliwe) wartości. Dlatego właśnie poziomy rotacyjne (stany energetyczne rotacji) są dyskretne (energia rotacji jest skwantowana).

ruchy oscylacyjne cząsteczki

Innym źródłem energii cząsteczki są ruchy oscylacyjne (drgania) atomów w cząsteczce względem ich położeń równowagi. Oscylacje takie można rozpatrywać, tak jakby to były drgania harmoniczne, a wtedy każdemu drganiu atomów, zachodzącemu z częstością ν, można przypisać energię $E_{osc} = h\nu(v+\frac{1}{2})$; tutaj h jest stałą Plancka, v — tzw. kwantową liczbą oscylacyjną, która podobnie jak kwantowa liczba rotacyjna, może przybierać tylko pewne wartości, a mianowicie $v = 0, 1, 2, 3, \ldots$. Tym samym również i poziomy oscylacyjne (stany energetyczne oscylacji) są poziomami dyskretnymi (energia oscylacji jest skwantowana).

Dwa pierwsze źródła energii związane są z cząsteczką traktowaną jako całość, trzecie wiąże się z ruchem elementów składowych cząsteczki, jakimi są atomy (bądź ściślej jądra atomowe).

energia elektronowa cząsteczki

Czwarte źródło energii wiąże się z elektronami w cząsteczce. Elektrony mogą być w cząsteczce rozłożone w różny sposób, a z każdym takim możliwym rozkładem ładunku elektronowego związana jest inna energia. Energię tę nazywamy energią elektronową cząsteczki, a stany energetyczne odpowiadające energii elektronowej nazywamy stanami elektronowymi cząsteczki. W zasadzie, nie można podać prostego, ogólnego wyrażenia opisującego energię elektronową cząsteczki (→ Chemia kwantowa), takiego jak w przypadku energii rotacyjnej i oscylacyjnej. Ale również i energia elektronowa cząsteczki jest skwantowana i może przybierać tylko pewne dyskretne wartości. Zmiana rozkładu ładunku elektronowego w cząsteczce, a więc zmiana stanu elektronowego cząsteczki, może pociągać za sobą również zmiany (czasem bardzo znaczne) momentu bezwładności i częstości oscylacji. Tym samym ze zmianą stanu elektronowego cząsteczki wiąże się na ogół zmiana energii E_{rot} i E_{osc}.

Te cztery źródła energii są najważniejszymi, ale nie jedynymi źródłami energii cząsteczki. I tak np. jeżeli jądra atomowe cząsteczki mają spin jądrowy, to z różną orientacją jąder wiąże się różna energia. Podobnie dodatkowa energia pochodzić może od orientacji spinów elektronowych i ruchów orbitalnych. Są to jednak na ogół już tylko efekty dodatkowe (i często bez większej wagi) w stosunku do wymienionych poprzednio rodzajów energii cząsteczki.

stan podstawowy cząsteczki

Mówiąc o różnych rodzajach energii cząsteczki, nie mówiliśmy nic o tym, jak duże są te energie i czy są one jednakowo istotne z punktu widzenia spektroskopii molekularnej. Aby móc o tym nieco więcej powiedzieć, musimy przede wszystkim sprecyzować skalę, a więc i zero tej skali, oraz jednostki, w jakich będziemy mierzyli energię cząsteczki. W spektroskopii molekularnej definiujemy zazwyczaj zero energii jako tzw. stan podstawowy cząsteczki. Jest to umowna definicja i przez stan podstawowy należy rozumieć taki stan, do osiągnięcia którego zmierzać będzie coraz więcej cząsteczek układu, wtedy gdy (w warunkach równowagi termicznej) temperatura układu zmierza do temperatury zera bezwzględnego. Taki stan jest najniższym możliwym stanem energetycznym cząsteczki, albo mówiąc inaczej, jest to najniższa możliwa energia, jaką może osiągać cząsteczka zachowując swoje cechy indywidualne, każde dalsze obniżenie jej energii związane będzie z reakcją chemiczną, a więc i ze zmianą samej cząsteczki.

jednostki energii

W spektroskopii molekularnej nie posługujemy się jednolitą skalą energii. Ze względu na dość znaczne różnice energii pochodzących z różnych źródeł, a także ze względu na różny często charakter problemów, w których rozwiązaniu pomaga spektroskopia, stosuje się różne jednostki do określania energii samych cząsteczek i energii procesów, w których one uczestniczą. Podstawową jednostką mógłby być oczywiście J/cząsteczkę, jest ona jednak bardzo niewygodna i dlatego podajemy inne jednostki, częściej używane w spektroskopii molekularnej, jak również związki między nimi:

1 eV/cząsteczkę = $1{,}6021 \cdot 10^{-19}$ J/cząsteczkę = $9{,}6481 \cdot 10^4$ J/mol
1 cm^{-1} ≙ $1{,}9862 \cdot 10^{-23}$ J/cząsteczkę = 11,9612 J/mol = $1{,}2398 \cdot 10^{-4}$ eV/cząsteczkę.

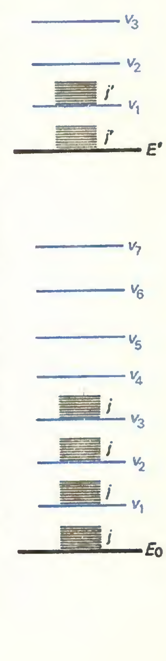
Operując tymi jednostkami możemy teraz porównać różne energie cząsteczki. I tak średnia energia translacyjna, w temperaturze pokojowej, jest rzędu 4 kJ/ /mol. Jednak tylko w nielicznych procesach, które bada spektroskopia, w grę wchodzi zmiana energii translacyjnej. Dlatego też w dalszych naszych rozważaniach nie będziemy brali pod uwagę energii translacyjnej.

Typowa wielkość zmiany energii rotacyjnej cząsteczki (energia przejścia rotacyjnego — odstęp energetyczny między sąsiednimi poziomami rotacyjnymi cząsteczki) jest rzędu 10^{-2} – 10^{-3} kJ/mol. Typowe zmiany energii oscylacyjnej są rzędu 4–40 kJ/mol. Natomiast odstępy energetyczne pomiędzy stanami elektronowymi są rzędu 400 kJ/mol.

poziomy energetyczne cząsteczki 2-atomowej

Widać więc, że poszczególne rodzaje energii składające się na całkowitą energię cząsteczki są bardzo różne. Gdybyśmy chcieli zobrazować ten wniosek graficznie, mogłoby to wyglądać tak jak na rys. 1. Taki schemat poziomów odpowiada zasadniczo cząsteczce dwuatomowej. Dla cząsteczek wieloatomowych liczba możliwych oscylacji, jakie mogą wykonywać atomy cząsteczki, jest tym większa im więcej atomów w cząsteczce (nie jest to zupełnie ścisłe i będziemy o tym jeszcze dalej mówili) i nasz rysunek musielibyśmy uzupełnić o wiele innych poziomów oscylacyjnych (z innym odstępem niż ten na rysunku) i każdemu z tych poziomów przyporządkować odpowiednie poziomy rotacyjne. Nietrudno sobie wyobrazić, że im większa byłaby cząsteczka, tym mniej czytelny byłby nasz rysunek.

Stany energetyczne cząsteczki charakteryzują pewne określone, chociaż czasami (jak np. dla stanów elektronowych) bardzo różne wartości energii. Ale to stwierdzenie nie mówi nam wcale, jakie warunki muszą być spełnione, ażeby cząsteczka przebywała w tym czy innym stanie energetycznym. Czy można w ogóle przewidywać, znając energię poszczególnych stanów, jaka część cząsteczek tworzących układ (np. gaz) będzie w określonych warunkach znajdowała się w tym czy innym stanie energetycznym? Otóż takie przewidywania są w pewnych warunkach możliwe. Przypuśćmy, że mamy pewien zbiór cząsteczek w równowadze termicznej (tzn., że temperatura jest jednakowa w całym układzie). Intuicja podpowie nam zapewne, że na to, by jakaś cząsteczka takiego układu mogła zmienić swój stan energetyczny, np. przejść ze stanu o niższej energii do stanu o wyższej energii, musi istnieć jakiś czynnik zewnętrzny, który dostarczy



Rys. 1. Schemat poziomów energetycznych cząsteczki: E_0, E' stany elektronowe — podstawowy i wzbudzony, v i v' stany oscylacyjne, j i j' stany rotacyjne (dla tych ostatnich skala energii na rysunku została zwiększona). Na rysunku brak jest wyższych wzbudzonych stanów elektronowych

jej energii potrzebnej do takiego przejścia. Czynniki takie mogą być różnej natury i o niektórych z nich będziemy jeszcze dalej mówili. Na razie niech nasz układ będzie odizolowany od wszelkich zewnętrznych czynników doprowadzających do niego energię. Czy w takich warunkach cząsteczki mogą czerpać skądś energię potrzebną do przejścia do wyższego stanu energetycznego? Tak, takim źródłem energii może być energia ruchu translacyjnego, którą cząsteczki mogą wymieniać między sobą wtedy, gdy się zderzają. Średnia energia ruchów translacyjnych zależy od temperatury układu i jest przez ten układ zachowana, jeżeli niemożliwa jest wymiana energii z otoczeniem układu. Oznacza to, że jeżeli w jakimś zderzeniu jedna cząsteczka uzyskuje energię i przechodzi do wyższego stanu energetycznego, to druga musi tę energię tracić, a w następnym zderzeniu może być odwrotnie. W sumie jednak w układzie takim musi istnieć równowaga (dynamiczna), tzn. w każdym określonym stanie energetycznym musi znajdować się stale, średnio rzecz biorąc, tyle samo cząsteczek. I trzeba tylko umieć określić — ile cząsteczek w jakim stanie? Albo inaczej, jakie jest obsadzenie (populacja) danego stanu? Ilościowej odpowiedzi na to pytanie dostarcza wyrażenie znane pod nazwą rozkładu Boltzmanna:

rozkład obsadzeń

$$\frac{n_i}{n_0} = e^{-(E_i-E_0)/kT}$$

Tutaj n_i jest liczbą cząsteczek w stanie i o energii E_i, albo inaczej obsadzeniem i-tego stanu, n_0 — obsadzeniem najniższego stanu (stanu o najniższej możliwej energii). Sens tego wyrażenia jest prosty. Im większa jest różnica energii $E_i - E_0$, tzn. im wyższy jest dany stan energetyczny i, tym mniej cząsteczek może w nim w danej temperaturze przebywać, tzn. tym słabiej jest on obsadzony. Na przykład dla stanów elektronowych, dla których $E_i - E_0$ jest rzędu 10^2 kJ/mol i więcej, stosunek n_i/n_0 jest w temperaturze pokojowej rzędu 10^{-10} i w praktyce poza stanem podstawowym żadne wyższe stany elektronowe nie są w takich warunkach w ogóle obsadzone. Natomiast wyższe stany oscylacyjne i wyższe stany rotacyjne podstawowego stanu elektronowego mogą być w tej temperaturze w pewnym stopniu obsadzone, szczególnie te ostatnie, dla których różnica energii jest rzędu zaledwie 10^{-2}–10^{-3} kJ/mol. Znajomość obsadzenia poszczególnych stanów energetycznych cząsteczki jest ważnym elementem, gdyż, jak przekonamy się jeszcze niejednokrotnie, obsadzenie stanów decyduje o rozkładzie natężeń w widmach cząsteczek.

Przejścia do wyższych stanów energetycznych wymagają doprowadzenia do cząsteczki energii z zewnątrz. Taką energię niesie ze sobą promieniowanie elektromagnetyczne. Oddziaływanie promieniowania elektromagnetycznego z cząsteczką może zatem zmienić energię cząsteczki. I to właśnie, jak powiedzieliśmy na samym początku, stanowi fundament spektroskopii. Zanim jednak przejdziemy do omawiania tego oddziaływania i obserwacji jego konsekwencji, przypomnimy krótko pewne podstawowe wiadomości dotyczące promieniowania elektromagnetycznego.

Promieniowanie elektromagnetyczne

Zacznijmy od przypomnienia, że promieniowanie elektromagnetyczne można opisywać dwojako: jako falę i jako strumień fotonów. Fala elektromagnetyczna — to rozchodząca się w przestrzeni i w czasie spójna zmiana pola elektrycznego i magnetycznego. Fali takiej, jak każdej fali, można przyporządkować długość λ i częstość ν; obie te wielkości są ze sobą związane zależnością:

własności falowe

$$\lambda = c/\nu,$$

gdzie c jest prędkością rozchodzenia się fali.

Widmo promieniowania elektromagnetycznego, tzn. zakres długości fal obejmujący promieniowanie elektromagnetyczne jest olbrzymi. Jednak z punktu widzenia spektroskopii molekularnej zakres ten jest stosunkowo wąski, bo obejmuje fale o długościach od ok. 10^{-7} m do ok. 10^{-3} m. W tym obszarze mieści się tzw. nadfiolet i promieniowanie widzialne (światło) oraz podczerwień i daleka podczerwień (granicząca z mikrofalami). Zamiast długością fali można się posługiwać jej odwrotnością $1/\lambda = \tilde{\nu}$, nazywaną liczbą falową; jednostką liczby falowej jest cm^{-1}.

Obszar widma	Długość fali λ nm	Liczba falowa $\tilde{\nu}$ cm^{-1}
Nadfiolet (bliski)	200–380	50000–26300
Widzialny	380–780	26300–12800
Podczerwień	780–$3 \cdot 10^4$	12800–333
Podczerwień (daleka)	$3 \cdot 10^4$–$3 \cdot 10^5$	333–33,3

fotony

Inny sposób opisu promieniowania elektromagnetycznego polega na traktowaniu go jako strumienia cząstek — fotonów, pozbawionych wprawdzie masy spoczynkowej, ale niosących ze sobą ściśle określoną energię,

$$E = h\nu,$$

gdzie ν jest wspomnianą wyżej częstością, a h — stałą Plancka.

Promieniowaniu o danej długości fali można przyporządkować ściśle określoną energię. Energię tę można obliczyć korzystając z przytoczonych do tej pory związków (np. promieniowaniu o długości fali λ = 400 nm, odpowiada energia 293,2 kJ/mol).

Cząsteczka, promieniowanie, widma

warunek Bohra

Kiedy kwant promieniowania elektromagnetycznego — foton, pada na cząsteczkę, może być przez nią pochłonięty. Warunek, który muszą spełniać cząsteczka i foton jest prosty:

$$\Delta E_{nm} = E_n - E_m = h\nu,$$

tzn. energia jaką niesie ze sobą foton musi pasować do różnicy energii ΔE_{nm} pomiędzy stanami energetycznymi m i n cząsteczki. Warunek ten nosi nazwę warunku Bohra.

absorpcja i emisja

Jeżeli warunek Bohra jest spełniony, to promieniowanie może zostać pochłonięte — mamy do czynienia z procesem absorpcji promieniowania. Cząsteczka przechodzi wtedy do stanu o wyższej energii, lub jak mówimy — zostaje wzbudzona. Możliwy jest również proces odwrotny. Wzbudzona cząsteczka może powrócić do stanu niższego, a nadmiar energii zostaje wysyłany przez nią w postaci kwantu promieniowania, o częstości (i długości fali) określonej warunkiem Bohra. Taki proces nazywa się emisją.

Warunek Bohra określa, jakie promieniowanie (o jakiej energii, długości fali) może być absorbowane lub emitowane przez cząsteczkę, nie jest to jednak jedyny warunek określający prawdopodobieństwo zachodzenia aktów absorpcji i emisji. Istnieją jeszcze inne warunki, które musi spełniać cząsteczka i promieniowanie i o nich będziemy dokładniej mówili w następnych rozdziałach.

(Z tego, co powiedziano wyżej, wynika, że jeżeli padające na cząsteczkę promieniowanie nie spełnia warunku Bohra, to nie może zostać przez nią zaabsorbowane. Nie jest to zupełnie prawdziwe, gdyż znane są procesy, w których takie promieniowanie może być absorbowane. Należy do nich dwu- i wielofotonowa absorpcja promieniowania; → Optyka nieliniowa).

Jeżeli na układ cząsteczek pada promieniowanie o różnych długościach fal, to może się zdarzyć, że dla niektórych z tych długości będzie spełniony warunek Bohra (i inne dodatkowe warunki). Promieniowanie o takich długościach fali będzie przez cząsteczki układu absorbowane, przy czym z reguły absorpcja ma różne natężenie dla promieniowania

widmo absorpcyjne

o różnych długościach fali (np. dlatego, że obsadzenie różnych stanów jest różne). Jeżeli potrafimy prześledzić i zarejestrować zmiany natężenia absorpcji w funkcji długości fali absorbowanego promieniowania, to uzyskany przez nas w ten sposób obraz jest tzw. widmem absorpcyjnym badanych cząsteczek. Podobnie, jeżeli padające promieniowanie wzbudza cząsteczki i wzbudza je do różnych stanów, to cząsteczki pozbywając się energii wzbudzenia mogą ją emitować w postaci promieniowania, przy czym natężenie tej emisji może być dla różnych długości fal różne (o tym, co decyduje o natężeniu emitowanego promieniowania, będziemy jeszcze szczegółowo mówili). Jeżeli teraz potrafimy zmierzyć natężenie emitowanego promieniowania w funkcji długości fali, to otrzymamy widmo emisyjne.

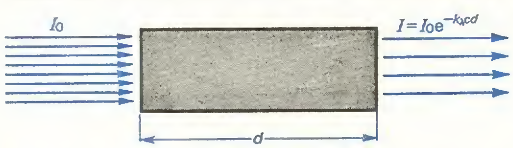
widmo emisyjne

Widma mogą być obserwowane dlatego, że energia promieniowania jest pochłaniana (bądź wysyłana) przez cząsteczkę. Absorpcja promieniowania wywołuje przejścia pomiędzy odpowiednimi stanami cząsteczki. Jeżeli z absorpcją taką wiążą się przejścia pomiędzy różnymi stanami rotacyjnymi cząsteczki, to odpowiednie widmo absorpcyjne nazywa się widmem rotacyjnym. Jeżeli są to przejścia pomiędzy stanami oscylacyjnymi, to odpowiednie widmo nosi nazwę widma oscylacyjnego. I wreszcie w wypadku przejść pomiędzy stanami elektronowymi — obserwujemy widmo elektronowe.

Warunek Bohra łączy ze sobą energię promieniowania i energię stanów cząsteczki, a ściślej różnicę energii pomiędzy różnymi stanami — zwaną często energią przejścia. Wiemy już, jakiego rzędu są różnice energii dla różnych rodzajów stanów energetycznych cząsteczki, a więc jakiego rzędu jest odpowiednia energia przejść, a tym samym, jaka musi być energia promieniowania, żeby mogło ono spowodować przejścia pomiędzy odpowiednimi stanami (znaleźć ją możemy korzystając z danych przytoczonej poprzednio tabeli i pamiętając, że $E = h\nu = hc/\lambda$). Jeżeli porównamy energię przejść z energią promieniowania elektromagnetycznego, to stwierdzimy, że: widma rotacyjne leżą w dalekiej podczerwieni, widma oscylacyjne — w obszarze podczerwieni, a widma elektronowe — w obszarze widzialnym i nadfiolecie.

prawo Beera

Jak w praktyce przebiega określanie natężenia absorpcji, a więc i badanie widma absorpcyjnego? Pochłanianie promieniowania przez układ cząsteczek (nazywa się taki układ po prostu próbką badaną) opisuje prawo Beera. Z prawa tego wynika, że padające na próbkę promieniowanie, określonej długo-



Rys. 2. Osłabienie wiązki światła przechodzącej przez ośrodek pochłaniający światło

ści fali λ, ulega w miarę wnikania w głąb próbki (np. na odległość d) stopniowemu osłabieniu (rys. 2), zgodnie z równaniem

$$I = I_0 e^{-ik_\lambda cd},$$

gdzie I jest natężeniem promieniowania po przejściu przez próbkę o grubości d, I_0 — natężeniem promieniowania padającego na próbkę, c — stężeniem cząsteczek pochłaniających promieniowanie w próbce, a k_λ — tzw. współczynnikiem absorpcji, charakterystycznym dla badanych cząsteczek. Logarytmując obie strony tego równania można je przedstawić w dogodniejszej postaci:

$$D = \log\frac{I_0}{I} = k'_\lambda cd, \; k'_\lambda = 0{,}4343\, k_\lambda;$$

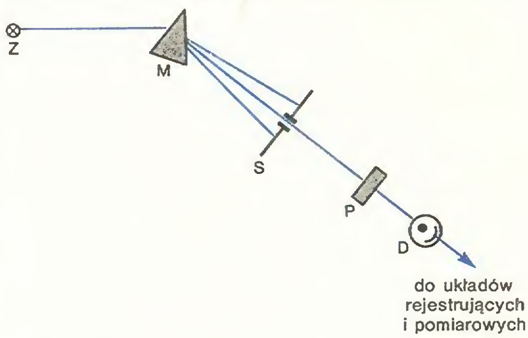
D nosi nazwę gęstości optycznej. Jak widać, jest to wielkość proporcjonalna do stężenia cząsteczek i grubości próbki. Jeżeli stężenie próbki podawane jest w molach na litr, a grubość w centymetrach, to wówczas prawo Beera można zapisać następująco:

$$D = \varepsilon_\lambda \cdot cd.$$

Nowy współczynnik absorpcji ε_λ nazywa się teraz molowym współczynnikiem absorpcji, albo współczynnikiem ekstynkcji (lub po prostu ekstynkcją). W praktyce więc badanie widma absorpcyjnego polega na wyznaczaniu gęstości optycznej lub współczynnika ekstynkcji ε w funkcji długości fali (lub liczby falowej $\tilde{\nu} = 1/\lambda$).

metody pomiaru widm

Sama idea takiego pomiaru jest prosta — ilustruje ją rys. 3. Podstawowe elementy układu pomiarowego to źródło promieniowania, układ dyspersyjny (monochromator) i detektor promieniowania. Źródło promieniowania wytwarza promieniowanie o różnych



Rys. 3. Schemat układu do badania widm absorpcyjnych: Z źródło promieniowania, M element dyspersyjny (pryzmat, siatka), P — badana próbka D detektor (fotopowielacz, fotoopór, klisza fotograficzna, itp.), S pomocnicze elementy optyczne (soczewki, lustra itp.). Promieniowanie przetworzone na sygnał (np. elektryczny, zaczernienie kliszy fotograficznej itp.) przekazywany jest z detektora do dalszej obróbki, której celem jest jego pomiar i ewentualnie rejestracja

długościach fal (na ogół tylko z pewnego obszaru widma). Z promieniowania tego układ dyspersyjny wydziela promieniowanie o odpowiedniej długości fali. Promieniowanie to przechodzi przez badaną próbkę i jest w niej pochłaniane, w mniejszym lub większym stopniu, bądź też przechodzi przez nią bez pochłaniania. Trafia ono następnie do detektora promieniowania, którego zadaniem jest wykryć je i przetworzyć w taki sposób, żeby można było określić jego natężenie, a tym samym wyznaczyć, jaka ilość promieniowania została w próbce zaabsorbowana.

Idea układu pomiarowego jest prosta, ale przy jego praktycznej realizacji występują pewne komplikacje. Wiążą się one z tym, że różnym rodzajom widm (rotacyjnym, oscylacyjnym i elektronowym) odpowiada inny obszar widma (daleka podczerwień, podczerwień, obszar widzialny i nadfiolet). A każdy z tych obszarów wymaga stosowania innych źródeł, elementów dyspersyjnych i detektorów. Nie ma np. takich źródeł, które dawałyby promieniowanie o dostatecznie dużym natężeniu dla celów praktycznych w całym zakresie widma. Trzeba wobec tego stosować różne źródła przy badaniu różnych obszarów widmowych. Do badań w nadfiolecie używa się np. lamp wodorowych, w obszarze widzialnym — lamp żarowych, a w podczerwieni — źródeł cieplnych (globarów). Podobnie elementy dyspersyjne — przede wszystkim pryzmaty, właściwe dla różnych obszarów widma, muszą być wykonywane z różnych materiałów, np. z kwarcu do badań w nadfiolecie, z soli kamiennej do badań w podczerwieni. Wreszcie detektory promieniowania, dobre do odbioru promieniowania z jednego obszaru widma, mają z reguły zbyt małą czułość do odbioru promieniowania z innego obszaru, np. dla nadfioletu i obszaru widzialnego stosuje się jako detektory fotopowielacze, a do badań w podczerwieni — termoelementy.

źródła i detektory promieniowania

Różnice w stosowanych materiałach optycznych i metodach wytwarzania i detekcji promieniowania sprawiają, że układy pomiarowe do badania widm absorpcyjnych (tzw. spektrometry i spektrofotometry) są w zasadzie budowane z przeznaczeniem do pracy w określonym obszarze widma, np. w podczerwieni. Tym samym nadają się one przede wszystkim do badania jednego, określonego typu widm, np. widm oscylacyjnych. Nie będziemy głębiej wnikali w różnego rodzaju rozwiązania konstrukcyjne, właściwe dla spektrofotometrów przeznaczonych do pracy w określonych obszarach widmowych. Czytelnika zainteresowanego problemami aparaturowymi odsyłamy do podręczników spektroskopii i monografii.

Widma oscylacyjne i rotacyjne cząsteczek

Krystyna Szczepaniak

Oscylacje atomów wchodzących w skład cząsteczek oraz rotacja całej cząsteczki powodują absorpcję, emisję i rozproszenie promieniowania elektromagnetycznego, a więc pojawienie się widm absorpcji, emisji i rozproszenia. W praktyce spektroskopowej badane są zwykle oscylacyjne i rotacyjne widma absorpcji i widma rozproszenia ramanowskiego.

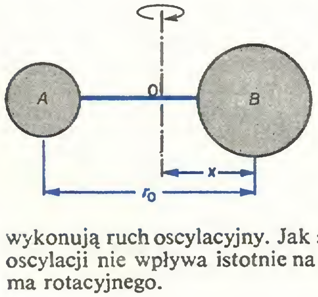
Widma rotacyjne

warunek trwałego momentu dipolowego

Badania widm rotacyjnych przyczyniają się skutecznie do wyznaczania rozmiarów i kształtu cząsteczek w fazie gazowej. Podstawowym warunkiem powstania widma rotacyjnego absorpcyjnego (inaczej mówiąc główną regułą wyboru) jest posiadanie przez cząsteczkę trwałego dipolowego momentu elektrycznego, tj. takiego rozkładu ładunków, przy którym środki ciężkości ładunków dodatnich i ujemnych nie pokrywają się. Miarą momentu dipolowego μ jest iloczyn ładunku q i odległości r między środkami ciężkości ładunków dodatnich i ujemnych: $\mu = q \cdot r$. Trwały moment dipolowy mają dwuatomowe cząsteczki złożone z różnych atomów, np. HCl, CO, NO, oraz wiele cząsteczek wieloatomowych, jak np. HCN, H_2O, $HCCl_3$. Cząsteczki nie mające trwałego momentu dipolowego, takie jak np. H_2, N_2, Cl_2, CO_2, CH_4, rotując nie pochłaniają promieniowania elektromagnetycznego, a więc nie dają one absorpcyjnego widma rotacyjnego. Cząsteczki takie mogą natomiast rozpraszać promieniowanie elektromagnetyczne dając rotacyjne widmo rozpraszania ramanowskiego. Powstawanie tego widma związane jest z nietrwałym momentem dipolowym indukowanym w cząsteczce przez falo elektromagnetyczne.

cząsteczka dwuatomowa

Omawianie ruchów rotacyjnych cząsteczek rozpoczniemy od najprostszych cząsteczek — cząsteczek dwuatomowych. Pozwoli nam to zrozumieć podstawowe cechy widma rotacyjnego bez wprowadzania dodatkowych komplikujących elementów, które pojawiają się przy omawianiu bardziej złożonych cząsteczek. Początkowo zajmiemy się rotacją, nie uwzględniając faktu, że atomy w cząsteczce jednocześnie wykonują ruch oscylacyjny. Jak się okaże, pominięcie oscylacji nie wpływa istotnie na większość cech widma rotacyjnego.



Rys. 4. Model cząsteczki dwuatomowej AB obracającej się wokół osi prostopadłej do wiązania

cząsteczka jako sztywny rotator

Najprostszym modelem, który pomaga w zrozumieniu i wyjaśnieniu podstawowych cech widma rotacyjnego, jest cząsteczka złożona z dwóch atomów A i B o masach m_A i m_B znajdujących się w stałej odległości r_0 od siebie (rys. 4). Model taki nosi nazwę sztywnego rotatora. Gdy cząsteczka obraca się wokół osi prostopadłej do linii łączącej oba atomy i przechodzącej przez środek ciężkości O, jej energia kinetyczna określona jest przez moment bezwładności I i prędkość kątową ω:

$$E_{\text{kin}} = \tfrac{1}{2} I\omega^2.$$

Moment bezwładności

$$I = m_A(r_0-x)^2+m_B x^2,$$

gdzie x jest odległością atomu B od środka ciężkości. Położenie środka ciężkości można wyznaczyć z równania $m_A(r_0-x) = m_B x$, skąd $x = m_A r_0/(m_A+m_B)$. Podstawiając wartość x do wyrażenia określającego moment bezwładności I otrzymuje się, po prostych przekształceniach,

$$I = \frac{m_A m_B}{m_A+m_B} r_0^2 = M r_0^2.$$

Wyrażenie $M = m_A m_B/(m_A+m_B)$ nazywamy masą zredukowaną. Omawiany ruch rotacyjny jest zagadnieniem dobrze znanym w mechanice. Potrafimy również rozwiązać równanie falowe Schrödingera dla takiego ruchu (→ Chemia kwantowa). Otrzymane w ten sposób wartości energii układu wyraża się wzorem

$$E_{\text{rot}} = \frac{h^2}{8\pi^2 M r_0^2} J(J+1) = BJ(J+1),$$

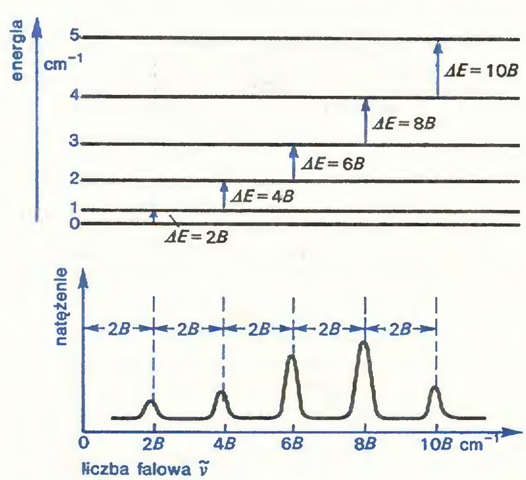
kwantowa liczba rotacyjna

gdzie J jest kwantową liczbą rotacyjną, która może przyjmować tylko całkowite wartości lub 0 ($J = 0, 1, 2, 3, \ldots$).

Wyrażenie $h^2/8\pi^2 M r_0^2$ zwane jest stałą rotacyjną i oznaczone literą B. Przejściami między poziomami rotacyjnymi rządzi reguła wyboru, która mówi, że przy absorpcji dozwolone są tylko te przejścia, w wyniku których kwantowa liczba rotacyjna zmienia się o ± 1 ($\Delta J = \pm 1$). Zgodnie z tą regułą zmiana energii cząsteczki przy absorpcji wynosi

$$\Delta E_{\text{rot}} = E_{J+1} - E_J = 2B(J+1).$$

Jak wynika z tego wzoru w miarę wzrostu J odległości między poziomami są coraz większe. Schemat poziomów energii rotacyjnej cząsteczki dwuatomowej



Rys. 5. Schemat poziomów energetycznych i widmo rotacyjne cząsteczki dwuatomowej (rotatora sztywnego)

przedstawia rys. 5. W dolnej części rysunku pokazane są linie rotacyjne odpowiadające przejściom między kolejnymi poziomami rotacyjnymi. Liczby falowe ko-

lejnych linii są $\tilde{\nu}_{01} = 2B$, $\tilde{\nu}_{12} = 4B$, $\tilde{\nu}_{23} = 6B$ itd., a odległość między liniami $\Delta\tilde{\nu}$ jes stała i równa $2B$. Mierząc tę odległość możemy wyznaczyć stałą B, a następnie moment bezwładności i odległość r_0.

cząstka jako rotator niesztywny

Z tak prostym sposobem ustalania rozmiarów cząsteczek na podstawie pomiarów spektroskopowych mamy do czynienia tylko w przypadku cząsteczek dwuatomowych. I dla nich zresztą przyjęliśmy niezupełnie prawidłowy model sztywnego rotatora.

Dokładne badania linii widma rotacyjnego wykazują, że odległości między liniami maleją dla większych wartości J. Przyczyną tego jest fakt, że cząsteczka pobudzona do wyższych stanów rotacyjnych obraca się prędzej i wiązanie nieco się rozciąga pod wpływem siły odśrodkowej. Wskutek tego moment bezwładności wzrasta, a odległości między liniami maleją. Konieczne jest zatem wprowadzenie pewnej poprawki do wyrażenia na energię:

$$E_{rot} = BJ(J+1) - DJ(J+1).$$

Stałą $D = 4B^3/\omega^2$ nazywa się stałą odkształcenia odśrodkowego. Jest ona około dziesięciu tysięcy razy mniejsza niż B; dla małych wartości J możemy więc pominąć drugi składnik wyrażenia.

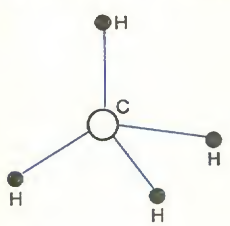
cząsteczki wieloatomowe liniowe

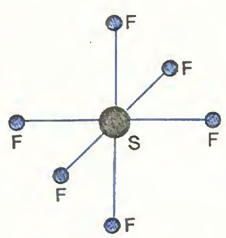
Podane wyżej wzory określające stany energii rotacyjnej cząsteczek dwuatomowych stosują się również do liniowych cząsteczek wieloatomowych. Od razu jednak napotykamy pewne trudności, które ilustruje przykład cząsteczki trójatomowej O=C=S. Cząsteczka ta ma dwa wiązania, lecz tylko jeden moment bezwładności. Problem możemy rozwiązać badając dwie cząsteczki, w skład których wchodzą różne izotopy siarki: $^{16}O^{12}C^{32}S$ oraz $^{16}O^{12}C^{34}S$. Zakłada się, że dla obu odmian cząsteczki długości wiązań są takie same. Każdej z odmian cząsteczki odpowiada nieco inne widmo rotacyjne, możemy więc wyznaczyć dwa momenty bezwładności i znaleźć długości obu wiązań cząsteczki.

cząsteczki wieloatomowe nieliniowe

Sytuacja jest bardziej złożona, gdy mamy do czynienia z cząsteczkami wieloatomowymi nieliniowymi. Rozróżniamy wówczas trzy typy ruchu rotacyjnego w zależności od stosunków trzech głównych momentów bezwładności I_A, I_B, I_C względem trzech wzajemnie prostopadłych osi związanych z cząsteczką:

a) Do pierwszego typu należą cząsteczki o nejwyższej symetrii, np. cząsteczki o tetraedrycznym lub oktaedrycznym rozkładzie atomów, takie jak np.:

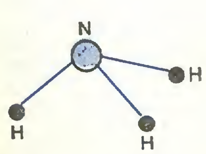


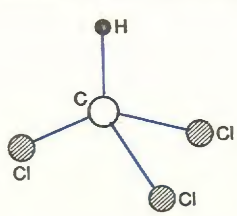


bąk sferyczny

W cząsteczkach takich wszystkie trzy momenty bezwładności są sobie równe $I_A = I_B = I_C$. Ten typ cząsteczki nazywa się bąkiem sferycznym. Cząsteczki tego typu nie mają trwałego momentu dipolowego, $\mu = 0$, zatem model ten jest nieinteresujący dla absorpcyjnej spektroskopii rotacyjnej.

b) Do drugiej grupy należą cząsteczki, dla których dwa spośród trzech momentów bezwładności są sobie równe $I_A = I_C \neq I_B$. Przykładem takich cząsteczek mogą być NH_3 i $CHCl_3$:





bąk symetryczny

Rozwiązując równanie falowe dla sztywnego bąka symetrycznego otrzymuje się

$$E_{rot} = AJ(J+1) - (A-B)K^2,$$

gdzie:

$$A = \frac{h^2}{8\pi^2 I_A}, \quad B = \frac{h^2}{8\pi^2 I_B}.$$

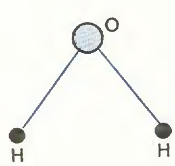
Trzeba było wprowadzić dwie liczby kwantowe J i K, ponieważ dla symetrycznego bąka występują dwa rodzaje obrotów. Liczba kwantowa J opisuje obrót wokół osi prostopadłej do jedynej osi symetrii, a liczba kwantowa K — obrót cząsteczki wokół osi symetrii. Obrót cząsteczki typu bąka symetrycznego wokół osi symetrii nie powoduje zmiany momentu dipolowego, a tym samym pojawienia się widma absorpcyjnego. Nic więc dziwnego, że reguła wyboru dla liczby kwantowej K brzmi $\Delta K = 0$. Reguła wyboru dla liczby kwantowej J jest taka sama, jak dla cząsteczek dwuatomowych i liniowych $\Delta J = \pm 1$. Ze względu na te reguły wyboru różnica energii między dwoma kolejnymi poziomami rotacyjnymi cząsteczki tego typu wyrażona jest wzorem:

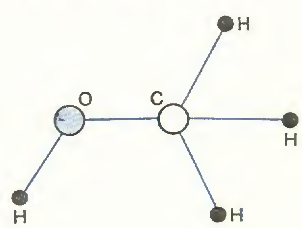
$$\Delta E_{rot} = 2B(J+1).$$

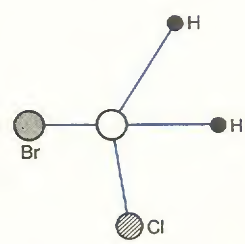
Wyrażenie to jest identyczne z wyrażeniem dla cząsteczek liniowych. Podobnie jak w wypadku cząsteczek liniowych, stosując podstawienie izotopowe wyznacza się na podstawie widma momenty bezwładności, a następnie parametry określające strukturę cząsteczki. W tabeli na następnej stronie podane są długości wiązań i kąty między wiązaniami dla kilku cząsteczek typu bąka symetrycznego wyznaczone za pomocą widm rotacyjnych.

bąk niesymetryczny

c) Cząsteczki, które mają trzy różne momenty bezwładności $I_A \neq I_B \neq I_C$, należą do trzeciej grupy, zwanej bąkiem niesymetrycznym. Do grupy tej należą cząsteczki o najniższej symetrii lub całkowicie niesymetryczne np. H_2O, CH_3OH, CH_2BrCl:



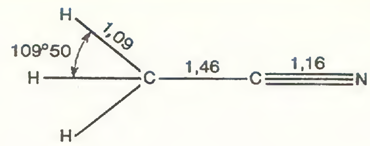
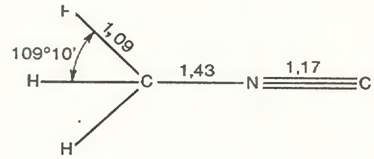
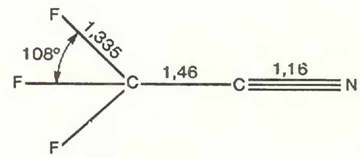
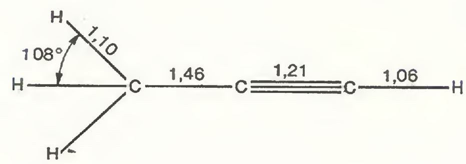
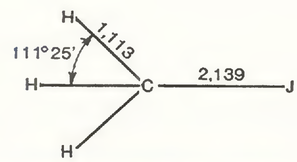




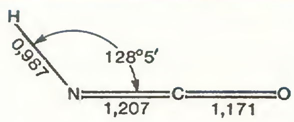
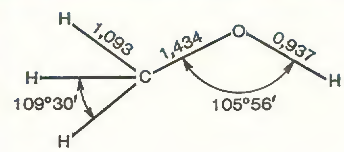
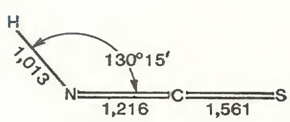
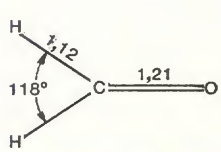
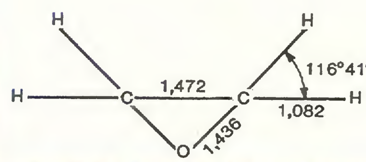
Poziomy energii rotacji takich cząsteczek nie mogą być wyrażone za pomocą jednego ogólnego równania, tak jak poziomy energii cząsteczek opisanych wyżej. Każdą cząsteczkę należy traktować oddzielnie. Utrudnia to bardzo obliczenia, mimo tego zrobiono duże postępy w interpretacji wielu złożonych widm rotacyjnych niektórych prostych cząsteczek typu bąka niesymetrycznego, szczególnie tych, dla których dwa momenty bezwładności mają zbliżone wartości. Dla niektórych płaskich cząsteczek, jak np. H_2O, wyznaczono z widm trzy momenty bezwładności. Tabela przedstawia wyniki analizy widma rotacyjnego kilku małych cząsteczek typu bąka niesymetrycznego.

Wszystko co dotychczas powiedzieliśmy o rotacji, dotyczyło cząsteczek swobodnych, z jakimi mamy do

Struktura cząsteczek typu bąka symetrycznego wyznaczona metodami spektroskopii rotacyjnej

Cząsteczka	B, MHz	Struktura (długość wiązań, Å)
CH_3CN	9198,83	
CH_3NC	10052,90	
CF_3CN	2945,54	
CH_3CCH	8545,84	
CH_3J	7501,31	

Struktura cząsteczek typu bąka niesymetrycznego wyznaczona metodami spektroskopii rotacyjnej

Cząsteczka	Struktura (długości wiązań, Å)	Cząsteczka	Struktura (długości wązań, Å)
HNCO		CH_3OH	
HNCS		CH_3SH	
HCHO		C_2H_4O	

rotacja w fazach skondensowanych

czynienia w gazie przy niezbyt wysokich ciśnieniach. Przy wysokich ciśnieniach w gazach, a zwłaszcza w fazach skondensowanych (cieczach i ciałach stałych), rotacja cząsteczek jest prawie zawsze mniej lub bardziej hamowana w wyniku oddziaływań międzycząsteczkowych. Energia większości oddziaływań międzycząsteczkowych jest bowiem tego rzędu wielkości (4–40 kJ/mol) co energia rotacji. Z tego powodu ciecze i ciała stałe nie mają widma rotacyjnego. Wyjątek stanowią cząsteczki zamrożone lub rozpuszczone w gazach szlachetnych, takich jak argon, krypton czy ksenon. W gazach szlachetnych bowiem oddzia-

ływania międzycząsteczkowe są tak słabe, że cząsteczki znajdujące się w stałym i ciekłym roztworze gazu szlachetnego w temperaturze 4 K i 20 K mogą wykonywać prawie swobodną rotację. Badaniami tego typu zamrożonych roztworów zajmuje się na świecie wiele laboratoriów naukowych.

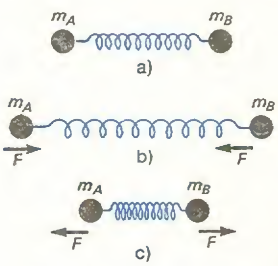
Widma oscylacyjne

Zajmiemy się teraz oddziaływaniem cząsteczek z promieniowaniem elektromagnetycznym o większej energii, a mianowicie promieniowaniem podczerwonym, wzbudzającym oscylacje atomów w cząsteczce. Aby cząsteczka mogła mieć absorpcyjne widmo oscylacyjne, konieczne jest by drganie wywoływało zmianę jej momentu dipolowego. W przypadku dwuatomowych cząsteczek, od których zaczniemy nasze rozważania, wymaganie to sprowadza się do tego, by cząsteczka składała się z dwu różnych atomów. Gdybyśmy chcieli się zająć, tak jak w przypadku rotacji, cząsteczkami w fazie gazowej, to napotykalibyśmy dodatkowe trudności. Otóż wzbudzając oscylacje promieniowaniem podczerwonym powoduje się zmiany nie tylko energii oscylacyjnej, ale również mniejszej od niej energii rotacyjnej cząsteczki. Tej kombinacji równoczesnego wzbudzenia oscylacji i rotacji można uniknąć badając cząsteczki w cieczach lub roztworach, w których jak już mówiliśmy, oddziaływania międzycząsteczkowe hamują rotację. Oczywiście oddziaływania te zmieniają również oscylacje, ale zmiany te nie są duże, gdyż energia większości oddziaływań międzycząsteczkowych jest znacznie mniejsza od energii oscylacji.

warunek zmiany momentu dipolowego

Omawianie ruchu oscylacyjnego zaczniemy od najprostszej cząsteczki dwuatomowej, którą można traktować, jak dwie masy połączone sprężyną (rys. 6).

cząsteczka dwuatomowa



Rys. 6. Model cząsteczki dwuatomowej

Gdy sprężyna nie jest rozciągnięta ani ściśnięta, jej długość wynosi r_e (rys. 6a). Przy tej odległości układ jest w równowadze. Podczas rozciągania sprężyny pojawia się siła sprężysta F działająca w kierunku przeciwnym do rozciągania (rys. 6b), która kieruje masy m_A i m_B z powrotem do stanu równowagi. Atomy mijają jednak stan równowagi i przekraczają odległość r_e, powodując ściśnięcie sprężyny i pojawienie się siły F skierowanej przeciwnie do kierunku ruchu (rys. 6c). Pod wpływem tej siły następuje ponownie powrót do stanu równowagi. Ruch taki powtarza się tak długo, aż energia włożona w pierwsze rozciągnięcie sprężyny zamieni się na skutek tarcia w energię cieplną. Zgodnie z prawem Hooke'a siła F jest proporcjonalna do wychylenia mas z położenia równowagi:

$$F = -k(r-r_e);$$

r jest odległością drgających mas w danej chwili, a r_e ich odległością w stanie równowagi. Współczynnik proporcjonalności k nazywa się stałą siłową oscylatora. Znak minus pochodzi stąd, że siła F jest skierowana przeciwnie do wychylenia $r - r_e$. Ruch pod wpływem siły proporcjonalnej do wychylenia jest ruchem harmonicznym, a potencjał dla takiego ruchu ma kształt paraboliczny: $U = kq^2$, gdzie q jest wychyleniem z położenia równowagi ($q = r-r_e$). Rozwiązując równania ruchu takiego układu w przybliżeniu klasycznym:

cząsteczka jako oscylator harmoniczny

$$\frac{d}{dt}\frac{dT}{d\dot{q}} + \frac{dU}{dq} = 0,$$

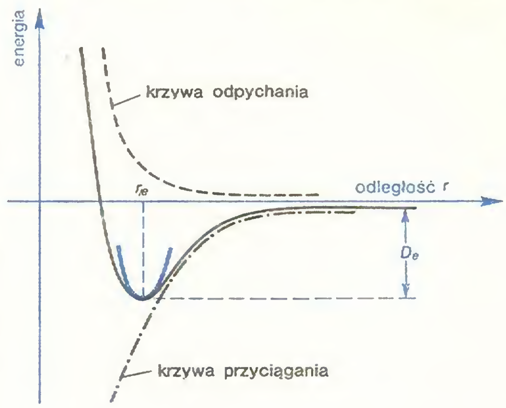
gdzie T jest energią kinetyczną $^1/_2M\dot{q}^2$, $\dot{q} = dq/dt$, U — energią potencjalną ($U = kq^2$), otrzymuje się częstość drgań oscylatora harmonicznego

$$\nu_0 = \frac{1}{2\pi}\sqrt{\frac{k}{M}},$$

gdzie M jest znaną nam już masą zredukowaną oscylatora, zaś k — stałą siłową.

Harmoniczny charakter ruchu atomów w cząsteczce wynika również z dokładniejszych rozważań kwantowomechanicznych (→ Chemia kwantowa). Energia potencjalna cząsteczki, w najprostszym przypadku cząsteczki dwuatomowej, zależy od sił przyciągania

energia potencjalna



Rys. 7. Krzywa energii potencjalnej cząsteczki dwuatomowej. Grubą linią niebieską oznaczono przybliżenie paraboliczne w pobliżu minimum. D_e energia dysocjacji względem minimum krzywej energii potencjalnej; r_e odległość w stanie równowagi

ładunków o różnych znakach i od sił odpychania ładunków o jednakowych znakach. Przyjmuje ona postać taką, jak na rys. 7. Krzywą taką można najogólniej opisać za pomocą szeregu potęgowego

$$U = U_0+U_1q+U_2q^2+ \ldots +U_nq^n,$$

gdzie $U_i = (1/n!)\,(d^nU/dq^n)_{q=0}$ są zależne od kolejnych pochodnych funkcji energii potencjalnej w punkcie $q = 0$. Gdy wychylenia atomów z ich położeń równowagi są małe, tzn. gdy znajdujemy się w pobliżu minimum krzywej energii potencjalnej, możemy w przybliżeniu przyjąć, że krzywa potencjalna ma postać $U = U_0+U_1q+U_2q^2$. Przyjmując U_0 za zero układu oraz biorąc pod uwagę fakt, że $(dU/dq)_{q=0} = 0$ w punkcie odpowiadającym minimum, otrzymamy na energię potencjalną wyrażenie $U = \frac{1}{2}(d^2U/dq^2)_{q=0}q^2$. W tym przybliżeniu krzywa energii potencjalnej jest więc parabolą.

Rozwiązanie równania falowego Schrödingera dla oscylatora harmonicznego daje energię skwantowanych poziomów

$$E = h\nu_0(v+\tfrac{1}{2}),$$

gdzie ν_0 jest tą samą częstością, jaką otrzymuje się z rozwiązań klasycznych, $\nu_0 = (1/2\pi)\sqrt{k/M}$, v jest oscylacyjną liczbą kwantową, która może przybierać wartości 0, 1, 2,... Jak wynika ze wzoru, energia oscylacji na poziomie $v = 0$ jest różna od zera i wynosi $\frac{1}{2}h\nu_0$. Oznacza to, że w żadnych warunkach, nawet w temperaturze zera bezwzględnego, oscylacje atomów nie ustają. Ten zasób energii oscylacyjnej, której cząsteczka nie może się pozbyć, nazywamy zerowa energią oscylacji.

oscylacyjna liczba kwantowa

Porównując wyniki przeprowadzonych poprzednio rozważań klasycznych dotyczących sił działających na atomy w cząsteczce z wynikami rozważań kwantowomechanicznych widzimy, że wprowadzona poprzednio stała siłowa k wiąże się z drugą pochodną przybliżonej funkcji opisującej energię potencjalną,

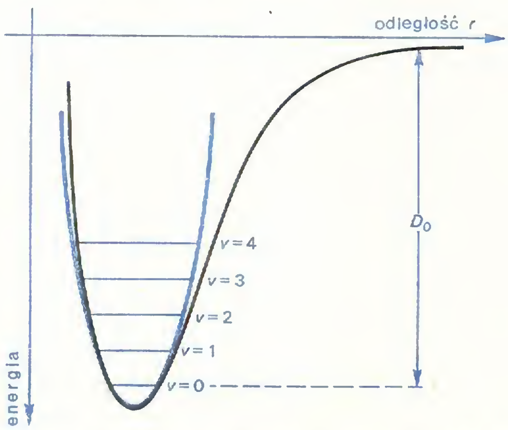
$$k = \left(\frac{d^2U}{dq^2} \right)_{q=0}$$

Jak wiadomo, energia ruchu elektronów i jąder w cząsteczce jest skwantowana. Przejściami między różnymi stanami energii oscylacyjnej rządzi reguła wyboru $\Delta v = \pm 1$. Odległość między sąsiednimi poziomami, określająca częstość linii w widmie oscylacyjnym, wyraża się wzorem

$$\Delta E = E_{v+1} - E_v = h\nu_0.$$

poziomy oscylacyjne

Rys. 8 przedstawia schemat poziomów oscylacyjnych cząsteczki HCl w przybliżeniu harmonicznym. Wyznaczając częstość oscylacji, możemy więc wyznaczyć stałą siłową. W tabeli zestawiono dane charakteryzujące niektóre cząsteczki dwuatomowe.



Rys. 8. Krzywa energii potencjalnej i schemat poziomów energetycznych cząsteczki dwuatomowej w przybliżeniu oscylatora harmonicznego (linia niebieska). Linią czarną zaznaczono krzywą energii bez uwzględnienia przybliżenia oscylatora harmonicznego; D_0 spektroskopowa energia dysocjacji

Dane charakteryzujące oscylacje cząsteczek dwuatomowych

Cząsteczka	Częstość oscylacji cm^{-1}	Stała siłowa N/cm
HF	3958	8,8
HCl	2886	4,84
HBr	2558	3,78
HJ	2233	2,89
CO	2155	18,6
NO	1877	15,4
NaCl	378	1,2

Przybliżenie oscylatora harmonicznego stosuje się dla niższych poziomów energetycznych. Gdy wzbudzone są wyższe poziomy oscylacyjne (patrz rys. 8), pojawia się czynnik nieharmoniczny w potencjale w postaci $U = U_2q^2 + U_3q^3$. Taka zmodyfikowana postać funkcji energii potencjalnej prowadzi do następującego wyrażenia na energię stanów oscylacyjnych:

$$E = h\nu_0(v+\tfrac{1}{2}) - xh\nu_0(v+\tfrac{1}{2})^2,$$

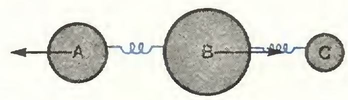
gdzie x jest stałą anharmoniczności.

anharmoniczność

Anharmoniczność pociąga za sobą pewne ważne następstwa. Po pierwsze odległości między poziomami oscylacyjnymi stają się coraz mniejsze, w miarę jak rośnie oscylacyjna liczba kwantowa. Mierząc stopniowe zmniejszanie się tych odległości, możemy znaleźć wartość x, a stąd wartość spektroskopowej energii dysocjacji $D_0 = \frac{1}{4}\nu_0(1-\frac{1}{2}x)$ (rys. 8.) Po drugie nie obowiązuje już reguła wyboru $\Delta v = \pm 1$, lecz dozwolone są również przejścia, dla których $\Delta v = \pm 2$, $\pm 3, \ldots$ Widmo składa się więc nie z jednego pasma podstawowego, jak to ma miejsce w przypadku oscylatora harmonicznego, lecz z pasma podstawowego i z szeregu pasm harmonicznych, tzw. nadtonów. Natężenie nadtonów jest jednak o kilka rzędów wielkości mniejsze od natężenia pasma podstawowego.

cząsteczka wieloatomowa

Dwuatomowa cząsteczka ma tylko jedną możliwość oscylacji lub — jak to się często mówi — ma jeden stopień swobody. Stanowi ona jeden oscylator, którego widmo jest, jak widzieliśmy, bardzo proste i łatwe do interpretacji. Znacznie bardziej skomplikowana jest sytuacja w cząsteczce zawierającej więcej atomów. Cząsteczkę wieloatomową możemy traktować jako pewną liczbę punktów materialnych połączonych ze sobą sprężynami. W tym przypadku nie można już traktować oscylacji jako zmian długości pojedynczych wiązań, lecz jako równoczesny ruch wszystkich atomów w cząsteczce. Kiedy w cząsteczce wzbudzone zostaną drgania jednego wiązania, ich energia jest przekazywana do drugiego za pośrednictwem wspólnego atomu:



Rozciągnięcie wiązania AB powoduje ściśnięcie wiązania BC. Energia potencjalna cząsteczki będzie zależała od wychyleń wszystkich atomów z ich położeń równowagi. Gdy cząsteczka składa się z N atomów, to istnieje $3N$ współrzędnych w trzech kierunkach prostopadłych x, y, z opisujących te wychylenia. We współrzędnych kartezjańskich xyz, opis ruchu, jaki wykonują wszystkie atomy w cząsteczce, jest bardzo skomplikowany. Niewiele upraszcza sprawę wprowadzenie tzw. współrzędnych naturalnych, tj. zmian długości wiązań i zmian kątów między wiązaniami. Okazuje się jednak, że można znaleźć takie współrzędne, zwane współrzędnymi normalnymi, których zastosowanie uprości bardzo opis ruchu cząsteczki, utożsamiając go z opisem ruchu atomów w cząsteczce dwuatomowej. Te niezwykle ułatwiające obliczenia współrzędne normalne są po prostu kombinacjami liniowymi współrzędnych naturalnych.

rozkład na drgania normalne

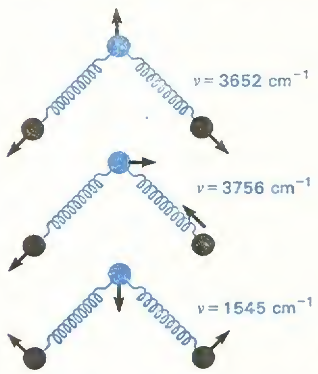
Złożony ruch cząsteczki wieloatomowej można rozłożyć na szereg ruchów harmonicznych, tzw. drgań normalnych, opisywanych tak, jak ruch cząsteczki dwuatomowej. Każdemu drganiu normalnemu o energii $E_i = h\nu_0^i(v+{}^1/_2)$ odpowiada częstość normalna $\nu_0^i = (1/2\pi)\sqrt{k_i/M}$. Energia stanu podstawowego równa $E_0 = \frac{1}{2}h\sum_{i=1}^{n}\nu_0^i$ może być bardzo duża, gdy cząsteczka zawiera wiele atomów. Do każdego drgania normalnego można zastosować regułę wyboru $\Delta v = \pm 1$, obowiązującą dla oscylatora harmonicznego.

Wiemy, że w cząsteczkach dwuatomowych drganie może być wzbudzone tylko wówczas, gdy moment dipolowy cząsteczki zmienia się przy drganiu. Uogólniając ten wynik na przypadek drgań normalnych, możemy powiedzieć, że wzbudzone będą tylko te drgania normalne, które prowadzą do zmian momentu dipolowego cząsteczki.

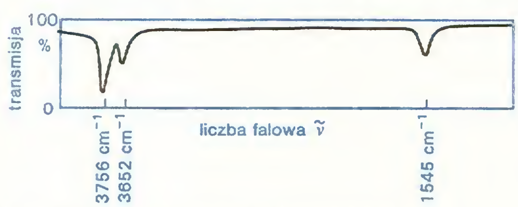
Zastanówmy się teraz, ile drgań normalnych może mieć cząsteczka wieloatomowa. Gdyby atomy tworzące cząsteczkę były zupełnie swobodne, każdy z nich miałby 3 stopnie swobody ruchu postępowego w kierunkach x, y, z, czyli N atomów miałoby $3N$ stopni swobody. Ponieważ atomy w cząsteczce są związane, ruch postępowy w przestrzeni wykonuje cała cząsteczka. Na ten ruch przypadają 3 stopnie swobody. Ponadto cząsteczka wykonuje rotację, na którą przypadają 3 stopnie swobody w cząsteczkach nieliniowych względem trzech prostopadłych osi i dwa stopnie swobody w cząsteczkach liniowych (gdzie trzecia oś jest osią cząsteczki). Reszta stopni swobody, tj. $3N-6$

dla cząsteczek nieliniowych oraz $3N-5$ dla cząsteczek liniowych przypada na oscylacje. Cząsteczkę N-atomową możemy rozpatrywać jako $3N-6$ (albo $3N-5$) oscylatorów wykonujących $3N-6$ (albo $3N-5$) drgań normalnych. Energia drgań każdego z oscylatorów jest skwantowana. Do każdego z oscylatorów stosuje się reguły wyboru opisane poprzednio dla oscylatorów dwuatomowych. Nie są to jednak oscylatory dwuatomowe, gdyż każdy atom w cząsteczce oddziałuje ze wszystkimi pozostałymi atomami. W oscylatorze dwuatomowym o energii i częstości oscylacji decyduje siła jednego istniejącego wiązania chemicznego. Miarą tej siły jest stała siłowa oscylatora.

liczba drgań normalnych



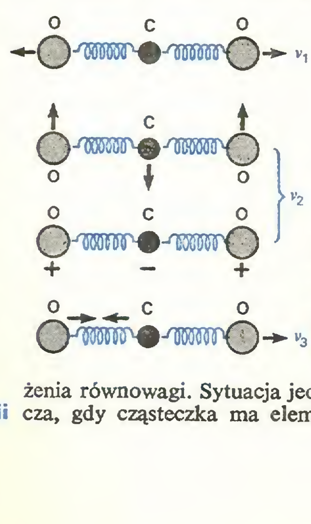
Rys. 9. Drgania normalne cząsteczki H_2O



Rys. 10. Widmo oscylacyjne H_2O w roztworze w CCl_4

W cząsteczce wieloatomowej występuje wiele takich sił różnie skierowanych. Prowadzą one do skomplikowanych ruchów oscylacyjnych atomów. Rysunek 9 pokazuje drgania normalne, a rys. 10 widmo oscylacyjne cząsteczki wody. Jak widać, liczba drgań normalnych zgodna jest z liczbą wewnętrznych stopni swobody $3N-6$ (dla H_2O $3 \cdot 3-6 = 3$). Można by zapytać, na jakiej podstawie zostały wprowadzone na rys. 9 strzałki oznaczające kierunek ruchu atomów. Przy dobieraniu strzałek kierujemy się następującą zasadą: równoczesne wychylenie atomów z położenia równowagi nie może powodować przemieszczenia środka ciężkości cząsteczki ani jej obrotu wokół jakiejś osi.

Współrzędne normalne są przeważnie złożonymi funkcjami wychyleń poszczególnych atomów z poło-



Rys. 11. Drgania normalne cząsteczki CO_2. Strzałki oznaczają kierunki wychyleń atomów w płaszczyźnie rysunku; znaki + i − oznaczają wychylenia atomów prostopadle do płaszczyzny rysunku

rola symetrii

żenia równowagi. Sytuacja jednak bardzo się upraszcza, gdy cząsteczka ma element symetrii. Rolę symetrii w układach cząsteczkowych zilustrujemy na podstawie współrzędnych normalnych liniowej cząsteczki CO_2 (rys. 11). Cząsteczka CO_2 ma płaszczyznę symetrii prostopadłą do jej osi. Oznacza to, że w stanie równowagi CO_2 ma identyczną konfigurację po obydwu stronach płaszczyzny symetrii. Innym elementem symetrii cząsteczki CO_2 jest jej oś. Obrót wokół tej osi o dowolny kąt nie zmienia konfiguracji układu. Typ symetrii z dwoma wymienionymi elementami symetrii oznaczamy symbolicznie $D_{\infty h}$, gdzie D_∞ odnosi się do osi symetrii nieskończonego rzędu, a wskaźnik h dotyczy płaszczyzny prostopadłej do tej osi symetrii.

drgania normalne cząsteczki CO_2

Liniowa cząsteczka CO_2 o symetrii $D_{\infty h}$ ma 4 drgania normalne ($3 \cdot 3-5 = 4$) przedstawione na rys. 11. Symetryczne drganie rozciągające ν_1 nie wywołuje zmiany momentu dipolowego cząsteczki, ponieważ równa się zeru, zarówno gdy cząsteczka jest rozciągnięta, jak też gdy jest ona ściśnięta. Drganie tego typu, symetryczne względem każdego elementu symetrii, nazywamy drganiem pełnosymetrycznym. Nie jest ono nigdy czynne w podczerwieni. Drganie ν_2 jest drganiem deformacyjnym (cząsteczka ulega wygięciu — pojawia się moment dipolowy). Atomy cząsteczki CO_2 mogą wykonywać pełne drgania w dwóch prostopadłych do siebie kierunkach, przy czym częstość obu tych drgań jest identyczna. Takie drgania o identycznej częstotliwości nazywa się drganiami podwójnie zdegenerowanymi. Moment dipolowy wygiętej cząsteczki CO_2 zmienia się podczas drgań deformacyjnych od pewnej skończonej wartości dodatniej poprzez zero do pewnej skończonej wartości ujemnej. Dzięki temu cząsteczka może oddziaływać z promieniowaniem elektromagnetycznym, a więc drganie ν_2 jest czynne w podczerwieni. Drganie ν_3 jest antysymetrycznym drganiem rozciągającym. Cząsteczka w końcowych fazach ruchu ma momenty dipolowe o przeciwnych znakach. Drganie ν_3 będzie więc również czynne w podczerwieni. Podczas drgania ν_3 zmieniający się moment dipolowy jest zawsze równoległy do osi cząsteczki.

drgania normalne cząsteczki H_2O

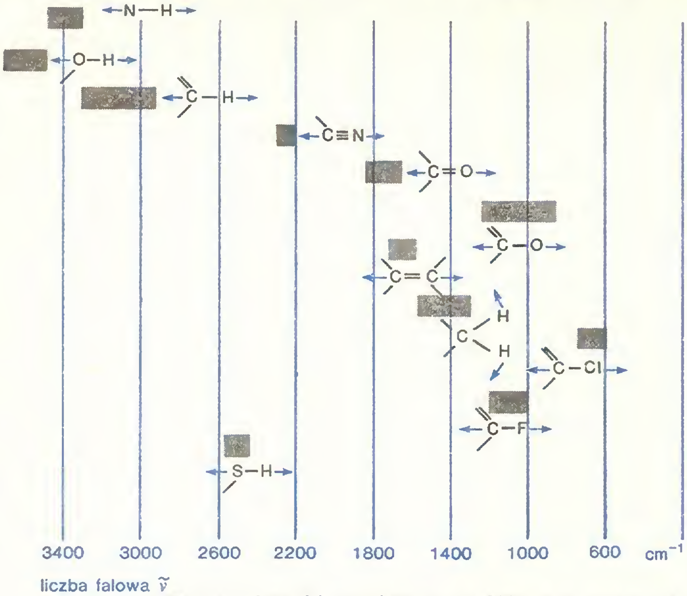
Nieliniowa cząsteczka H_2O należy do grupy symetrii C_{2v}. Oznacza to, że ma ona oś symetrii drugiego rzędu, tzn. obrót wokół tej osi o kąt 180° (360°: 2) nie zmienia konfiguracji cząsteczki, oraz dwie płaszczyzny symetrii (płaszczyznę cząsteczki i płaszczyznę do niej prostopadłą przechodzącą przez atom tlenu). Cząsteczka H_2O ma trzy drgania normalne ($3 \cdot 3-6 = 3$) pokazane na rys. 9. Zmieniają one moment dipolowy cząsteczki i w związku z tym są czynne w podczerwieni.

Widma oscylacyjne cząsteczek wieloatomowych są przeważnie bardziej złożone aniżeli omówione przez nas dwa przykłady CO_2 i H_2O. Aby znaleźć postać drgań normalnych, należy wziąć pod uwagę symetrię cząsteczki i zastosować metodę teorii grup. Przypisanie odpowiednim drganiom normalnym linii obserwowanych w widmie jest olbrzymią pracą, prowadzoną dzisiaj za pomocą najnowocześniejszych elektronicznych maszyn cyfrowych. Dokonano już analizy drgań normalnych wielu cząsteczek, ale jeszcze wielka liczba, zwłaszcza cząsteczek o niższej symetrii, pozostaje ciągle nie rozszyfrowana.

Chociaż oscylacje w cząsteczkach wieloatomowych są oscylacjami wszystkich atomów cząsteczki, zdarza się często, że drgania zlokalizowane są w pewnych grupach atomów wchodzących w skład cząsteczki. Gdy taka grupa atomów znajduje się w różnych cząsteczkach, w widmach występować będą pasma o tych samych prawie częstościach i natężeniach. Jest to efekt tak pewny i sprawdzony w tak wielu przykładach, że obecność lub nieobecność w widmie badanej cząsteczki pasma w danym zakresie częstości może być poważnym argumentem świadczącym o obecności lub braku odpowiedniego zgrupowania atomów w badanej cząsteczce. Częstości związane z danym ugrupowaniem atomów zwane są częstościami charakterystycznymi grup atomowych. Rys. 12 przed-

częstości charakterystyczne

stawia najbardziej rozpowszechnione i najlepiej zbadane częstości charakterystyczne grup atomów.



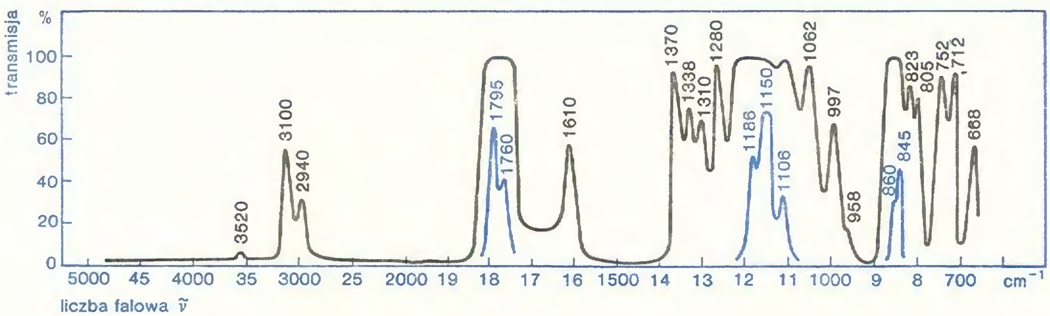
Rys. 12. Częstości charakterystyczne kilku grup atomowych. Strzałkami zaznaczono kierunek przesunięć atomów

Aby znaleźć najbardziej prawdopodobną strukturę i skład, badamy zakresy widma, w których występują charakterystyczne częstości grup OH i $>CH_2$. Brak silnego pasma w obszarze 3500–3600 cm^{-1} świadczy o niewystępowaniu grup OH, a więc przeciw strukturze I, III, V (bardzo słabe pasmo przy 3520 cm^{-1} jest najprawdopodobniej spowodowane przez zanieczyszczenia). Brak silniejszej absorpcji przy 2850 cm^{-1} pozwala odrzucić struktury II, III i V z kilkoma grupami $>CH_2$, którym ta absorpcja odpowiada. Występowanie słabego pasma przy 2940 cm^{-1} świadczy o obecności pojedynczej grupy $>CH_2$ dołączonej do grupy $>C{=}O$. Najbardziej prawdopodobna wydaje się więc struktura IV.

Widma oscylacyjno-rotacyjne

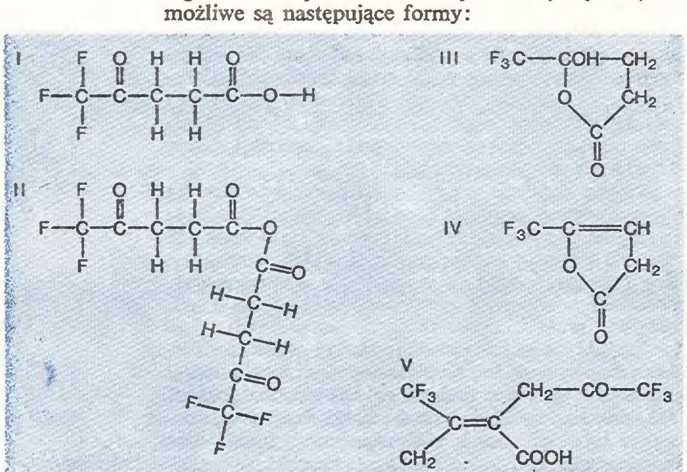
Dotychczas rozpatrywaliśmy oscylacje cząsteczek nie uwzględniając faktu, że wykonują one równocześnie ruch oscylacyjny i rotacyjny. Pominięcie rotacji jest, jak mówiliśmy, uzasadnione w odniesieniu do cząsteczek w cieczach, roztworach i ciałach stałych, gdzie siły międzycząsteczkowe powodują prawie całkowite zahamowanie obrotów. W fazie gazowej należy koniecznie uwzględnić równoczesne występowanie oscy-

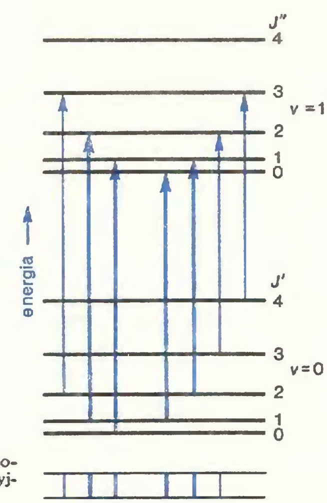
oscylacje i rotacje w gazach



Rys. 13. Widmo oscylacyjne nieznanej próbki. Krzywa czarna przedstawia widmo zarejestrowane przy większej grubości warstwy absorbującej, krzywa niebieska — przy grubości mniejszej

Aby zilustrować zastosowanie częstości charakterystycznych grup do analizy widma oscylacyjnego cząsteczki o nieznanym składzie i strukturze, rozpatrzymy następujący przykład. Weźmy pod uwagę widmo absorpcyjne w podczerwieni, takie jak np. na rys. 13. Ze sposobu otrzymywania związku, którego strukturę i skład mamy ustalić, wynika, że możliwe są następujące formy:





Rys. 14. Schemat poziomów oscylacyjno-rotacyjnych

lacji i rotacji. Korzystając z naszych dotychczasowych wiadomości możemy zapisać, że dla najprostszego przypadku dwuatomowych cząsteczek, energia równoczesnej rotacji i oscylacji cząsteczki w przybliżeniu rotatora sztywnego i oscylatora harmonicznego wynosi

cząsteczka dwuatomowa

$$E_{\text{osc rot}} = h\nu_0(v+\tfrac{1}{2})+BJ(J+1).$$

Gdy zachodzi przejście pomiędzy dwoma poziomami, z których wyższy oznaczamy pojedynczą kreseczką a niższy podwójną (rys. 14), otrzymujemy następujące wyrażenie na różnicę energii poziomów:

$$E'_{osc\,rot}-E''_{osc\,rot} = h\nu_0(v'-v'')+B'J'(J'+1)- -B''J''(J''+1),$$

gdzie B' i B'' są stałymi rotacyjnymi odpowiadającymi dwu stanom oscylacyjnym, a $J' = 0, 1, 2, 3, \ldots$ $J'' = 0, 1, 2, 3, \ldots$. Przejściami między poziomami rządzą reguły wyboru dla liczb kwantowych v i J, które mówią że $\Delta v = v'-v'' = 1$, a $\Delta J = J'-J'' = \pm 1$.

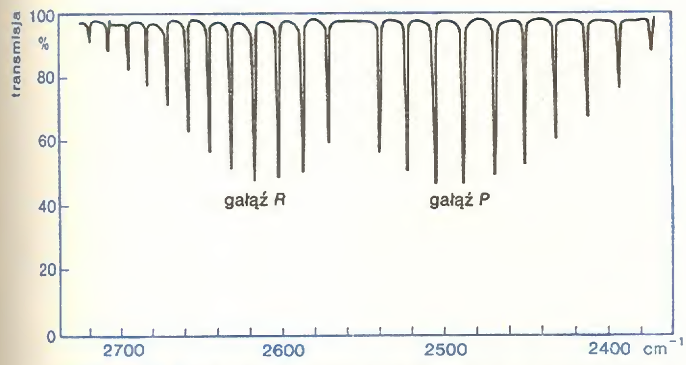
Musimy teraz rozważyć dwa możliwe sposoby zmian rotacyjnej liczby kwantowej J. Gdy $J'-J'' = +1$ otrzymujemy

$$\Delta E_{osc\,rot} = h\nu_0+(B'+B'')J'+(B'-B'')J'^2,$$

$J' = 1, 2, 3$, zaś dla $J'-J'' = -1$

$$\Delta E_{osc\,rot} = h\nu_0-(B'+B'')J'+(B'-B'')J'^2,\quad J' = 1, 2, 3.$$

Rysunek 15 przedstawia typowe pasmo oscylacyjno-rotacyjne cząsteczki dwuatomowej. Składa się ono z wielu prawie równoległych linii. Układ linii po stro-



Rys. 15. Struktura rotacyjna pasma oscylacyjnego HBr

gałęzie P i R nie niższych częstości $\Delta J = -1$ nazywa się zwykle gałęzią P, a układ linii po stronie wyższych częstości $\Delta J = +1$ — gałęzią R. W środku pasma przy częstości ν_0 występuje przeważnie luka odpowiadająca niedozwolonemu przejściu z $\Delta J = 0$. Wartości B' i B'' można bezpośrednio wyznaczyć z widma oscylacyjno-rotacyjnego. Dla składowych gałęzi P i R, w których zachodzi absorpcja z tego samego poziomu rotacyjnego J', otrzymuje się (rys. 14):

$$\nu_R(J')-\nu_P(J') = 2B''(2J'+1).$$

Z zależności tej można wyznaczyć B'' a następnie moment bezwładności I''. Dla składowych gałęzi P i R, w których zachodzi wzbudzenie do tego samego poziomu rotacyjnego J'' (rys. 14),

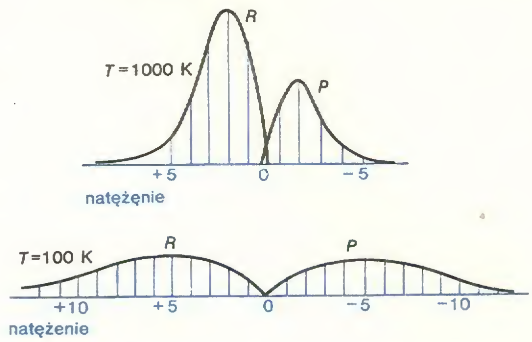
$$\nu_R(J')-\nu_P(J'+2) = 2B'(2J'+3).$$

Z tej różnicy można wyznaczyć B' oraz moment bezwładności I'.

Rozważania nasze prowadzą do wniosku, że struktura widm oscylacyjno-rotacyjnych zależy od momentu bezwładności, a więc od odległości międzyatomowych. Widma oscylacyjno-rotacyjne dają więc jeszcze jedną możliwość wyznaczania rozmiarów i kształtu cząsteczek.

względne natężenia składowych w widmie Omawiając widma oscylacyjno-rotacyjne cząsteczek dwuatomowych warto zwrócić uwagę na natężenia składowych pasma pokazanego na rys. 15. Względne natężenia tych składowych są w przybliżeniu proporcjonalne do obsadzania poziomów rotacyjnych będących poziomami wyjściowymi dla przejść kwantowych. Obsadzenie poziomu rotacyjnego, czyli liczba cząsteczek znajdujących się w stanie energetycznym opisanym przez liczbę kwantową J, wynosi:

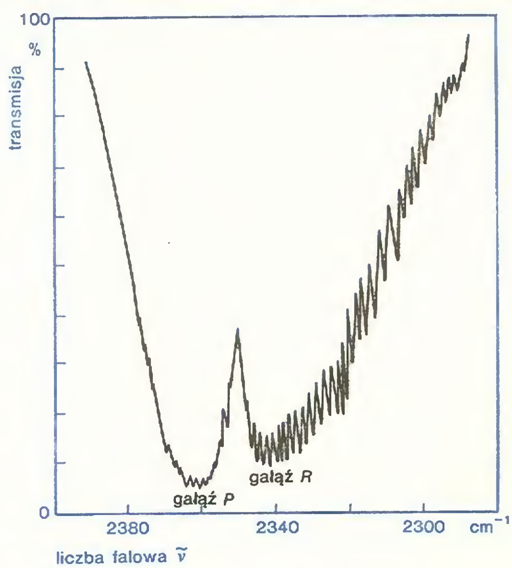
$$N_J = (2J+1)\,N_0 e^{-J(J+1)B/kT}$$



Rys. 16. Wpływ temperatury na rozkład natężenia pasma oscylacyjno-rotacyjnego HCl

Natężenia składowych odpowiadających przejściom z poziomów o różnym J będą się więc zmieniać wraz ze zmianą temperatury próbki (rys. 16).

cząsteczki wieloatomowe liniowe Energia równoczesnych oscylacji i rotacji wieloatomowych cząsteczek liniowych jest w przybliżeniu sumą $E_{osc}+E_{rot}$. Dla takich cząsteczek występują dwa podstawowe typy pasm oscylacyjno-rotacyjnych. Gdy momenty dipolowe związane z drganiami cząsteczki są równoległe do jej osi, otrzymuje się pasmo rotacyjno-oscylacyjne równoległe (rys. 17). Jeżeli na- **pasmo równoległe**



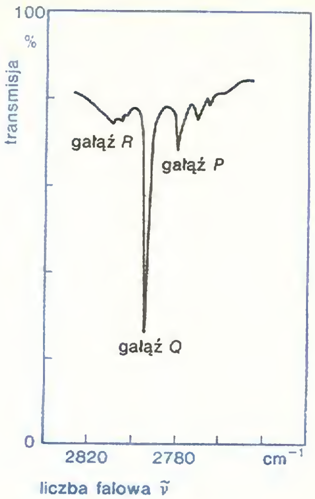
Rys. 17. Pasmo równoległe cząsteczki CO_2

pasmo prostopadłe tomiast oscylacyjne momenty dipolowe są prostopadłe do osi cząsteczki, otrzymuje się pasmo nazywane pasmem prostopadłym (rys. 18). Dla pasm równoległych ważne są reguły wyboru takie same, jak dla cząsteczek dwuatomowych, tj. $\Delta v = \pm 1$, $\Delta J = \pm 1$. Podobnie jak dla cząsteczek dwuatomowych, w widmie wystąpią tylko gałęzie P i R. Do tych pasm można zastosować rozważania dotyczące cząsteczek dwuatomowych i wyznaczyć z odległości między liniami rotacyjnymi pasma oscylacyjnego równoległego wartości B' i B''.

Przy drganiach prostopadłych do osi cząsteczki ważne są inne reguły wyboru, a mianowicie: $\Delta v = \pm 1$, $\Delta J = 0, \pm 1$. W związku z tym pasma prostopadłe cząsteczek liniowych mają oprócz gałęzi P i R także

gałąź Q

gałąź centralną Q odpowiadającą przejściu $\Delta J = 0$ (rys. 18). Jedną z głównych cech gałęzi Q jest to, że wszystkie przejścia $J' \rightarrow J''$ tej gałęzi mają prawie taką

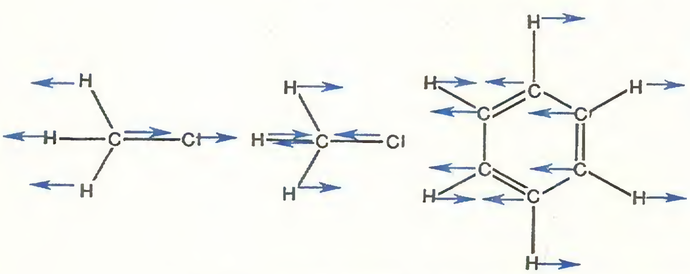


Rys. 18. Pasmo prostopadłe cząsteczki NO_2 odpowiadające drganiom zginającym cząsteczkę. Dwa satelity gałęzi Q należy przypisać cząsteczkom NO_2 we wzbudzonym stanie oscylacyjnym

samą częstość ν_0. Ze względu na nieznacznie mniejszą wielkość B'' w porównaniu do B', gałąź Q jest nieco asymetryczna od strony większych J.

cząsteczki typu bąka symetrycznego

Drgania cząsteczek typu bąka symetrycznego dające pasma oscylacyjno-rotacyjne są związane z wibracyjnym momentem dipolowym, ustawionym równolegle bądź prostopadle do osi cząsteczki. Podobnie jak dla cząsteczek liniowych, obydwu typom drgań odpowiadają różne rotacyjno-oscylacyjne reguły wyboru i stąd różny kształt pasm. Drgania dające pasmo równoległe przedstawia rys. 19. Wszystkie pokazane na rysunku drgania mają wibracyjny moment dipo-



Rys. 19. Drgania cząsteczek będących rotatorami symetrycznymi, prowadzące do pasm absorpcyjnych równoległych

lowy skierowany wzdłuż osi cząsteczki. Dla drgań tego typu regułami wyboru są:

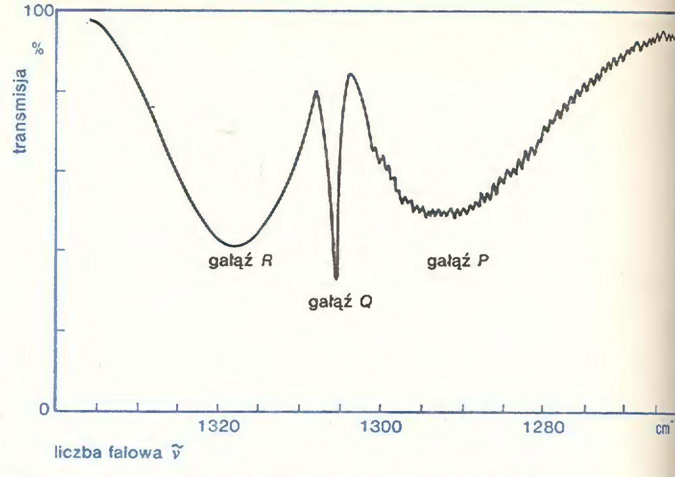
$$\Delta v = \pm 1, \quad \Delta K = 0, \quad \Delta J = 0, \pm 1 \quad \text{jeżeli } K \neq 0;$$
$$\Delta v = \pm 1, \quad \Delta K = 0, \quad \Delta J = \pm 1, \quad \text{jeżeli } K = 0,$$

gdzie liczby kwantowe K i J znane nam już są z rozważań widm czysto rotacyjnych. Przykład równoległego pasma oscylacyjno-rotacyjnego cząsteczki CH_3Br przedstawia rys. 20. Pasmo składa się z trzech gałęzi P, Q i R. Jak widać na rysunku tylko dla gałęzi P widoczne są rozdzielone linie rotacyjne. Dla gałęzi R i Q występują pasma bez zaznaczonej struktury rotacyjnej.

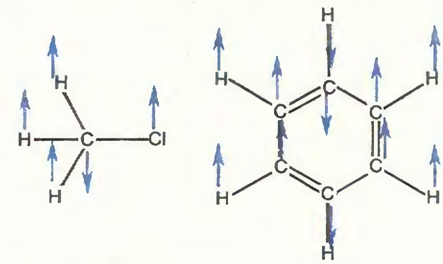
cząsteczki typu bąka niesymetrycznego

Drgania, przy których powstaje moment dipolowy prostopadły do osi cząsteczki, przedstawia schematycznie rys. 21. Do drgań tego typu stosują się następujące reguły wyboru dla liczb kwantowych oscylacyjnych i rotacyjnych $\Delta v = \pm 1$, $\Delta K = \pm 1$, $\Delta J = 0, \pm 1$. Typowe pasmo oscylacyjno-rotacyjne prostopadłe przedstawia rys. 22.

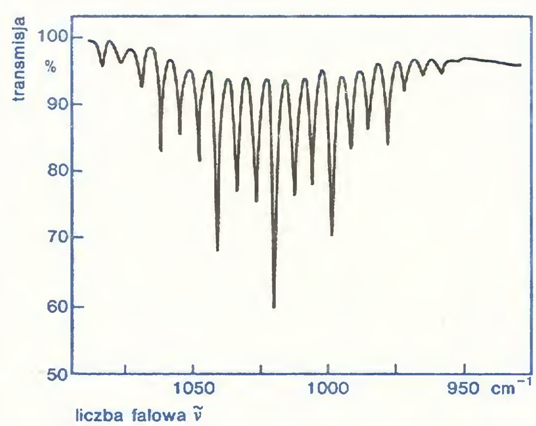
W przypadku cząsteczek typu bąka niesymetrycznego nie możemy dzielić pasm oscylacyjno-rotacyjnych na równoległe i prostopadłe. Zależnie od tego czy zmiana momentu dipolowego odbywa się w kierunku najmniejszego, pośredniego lub największego



Rys. 20. Pasmo równoległe cząsteczki CH_3Br (rotator symetryczny)



Rys. 21. Drgania cząsteczek będących rotatorami symetrycznymi prowadzące do pasm prostopadłych



Rys. 22. Pasmo prostopadłe cząsteczki CH_3Cl (rotator symetryczny)

pasma typu A, B i C

momentu bezwładności mówimy o pasmach typu A, B lub C. Każde z tych trzech pasm ma zwykle pewien charakterystyczny kształt, co pozwala je odróżnić od siebie.

Widma rozproszenia Ramana

Widma rozproszenia Ramana są jednym z wielu przykładów zjawisk przewidywanych przez teoretyków wcześniej, zanim zostały odkryte doświadczalnie. W 1923 r. A.G. Smekal zwrócił uwagę, że w promieniowaniu rozproszonym powinny się pojawić obok fotonów o częstości promieniowania padającego ν_0, fotony o częstościach $\nu_0 \pm \nu$. Teoretycy z niecierpliwo-

ścią oczekiwali doświadczalnego potwierdzenia swych przewidywań, które mogło ugruntować lub obalić podstawy mechaniki kwantowej. I oto w 1928 r. pojawiają się pierwsze prace doświadczalne fizyka hinduskiego Ch.V. Ramana dotyczące rozproszenia w cieczy (benzenie) oraz prace fizyków radzieckich G. S. Landsberga i L.I. Mandelsztama dotyczące rozproszenia w kryształach kwarcu.

mechanizm rozproszenia ramanowskiego

Mechanizm rozproszenia ramanowskiego jest następujący: padające promieniowanie elektromagnetyczne indukuje w cząsteczce moment dipolowy

$$\vec{\mu}_i = \alpha\vec{E},$$

polaryzowalność cząsteczki

gdzie α jest polaryzowalnością cząsteczki, czyli miarą zdolności do deformacji rozkładu jej ładunków w polu elektromagnetycznym lub elektrycznym. Polaryzowalność jest tym większa, im słabiej związane są elektrony zewnętrznych powłok z jądrami atomów. W czasie drgania cząsteczki, powodującego periodyczne zmiany jej struktury, zmienia się również polaryzowalność. Jest więc ona funkcją współrzędnych q opisujących drganie cząsteczki:

$$\alpha = \alpha(q).$$

Periodycznym zmianom współrzędnych q odpowiadają zmiany periodyczne $\alpha(q)$. W przybliżeniu harmonicznym

$$\alpha(q) = \alpha_0 \cos 2\pi\nu t,$$

gdzie ν jest częstością drgania, np. drgania normalnego w przypadku cząsteczek wieloatomowych. Promieniowanie elektromagnetyczne E też zmienia się periodycznie: $E = E_0 \cos 2\pi\nu_0 t$, a więc indukowany w cząsteczce przez falę elektromagnetyczną moment dipolowy μ_i wynosi

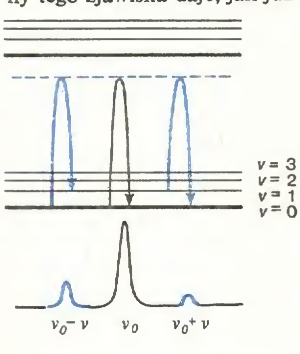
$$\mu_i = \alpha_0 E_0 \cos 2\pi\nu t \cos 2\pi\nu_0 t.$$

Stosując znany wzór trygonometryczny na iloczyn cosinusów otrzymamy

$$\mu_i = {}^1/_2\alpha_0 E_0 \; [\cos 2\pi(\nu_0-\nu)t + \cos 2\pi(\nu_0+\nu)t].$$

Drgający indukowany moment dipolowy ma więc składowe $\nu_0-\nu$ i $\nu_0+\nu$. Częstościami ν mogą być częstości oscylacji lub rotacji opisywane wzorami podanymi poprzednio przy omawianiu widm oscylacyjnych i rotacyjnych.

Przedstawiony mechanizm rozproszenia ramanowskiego jest bardzo uproszczony. Pełny opis teoretyczny tego zjawiska daje, jak już mówiliśmy, mechanika



Rys. 23. Schemat poziomów energetycznych i widmo Ramana cząsteczki dwuatomowej; linią przerywaną zaznaczono poziom niestacjonarny

kwantowa. Schemat poziomów energetycznych wynikających z rozważań mechaniki kwantowej przedstawia rys. 23. Fotony padającego promieniowania o częstości ν_0 ulegają rozproszeniu na cząsteczkach. Gdy po rozproszeniu promieniowania cząsteczka pozostaje w tym samym stanie energii — mamy do czynienia z rozproszeniem bez zmiany długości fali, któremu odpowiada środkowa linia ν_0 w dolnej części rys. 23. Zdarza się jednak, że cząsteczka po rozproszeniu znajdzie się na wyższym poziomie rotacyjnym lub oscylacyjnym i rozproszony foton ma częstość zmniejszoną o różnicę energii rotacyjnych lub oscylacyjnych poziomów energetycznych. Takiemu rozproszeniu odpowiada linia $\nu_0-\nu$ w dolnej części rys. 23, zwana linią stokesowską. Jeżeli przed rozproszeniem cząsteczka znajdowała się we wzbudzonym stanie rotacyjnym lub oscylacyjnym, to możliwe jest, że po rozproszeniu znajdzie się w stanie podstawowym. Rozproszony foton zwiększy swą częstość o różnicę energii $h\nu$ rotacyjnych lub oscylacyjnych poziomów energetycznych. Odpowiada mu linia $\nu_0+\nu$ pokazana w dolnej części rys. 23, zwana linią antystokesowską. Ponieważ we wzbudzonym stanie rotacyjnym, a zwłaszcza oscylacyjnym, jest znacznie mniej cząsteczek niż w stanie podstawowym, przejścia antystokesowskie są znacznie rzadsze, a linie im odpowiadające mają mniejsze natężenie. Częstości linii ramanowskich rotacyjnych i oscylacyjnych otrzymać można na podstawie rozważań przeprowadzonych poprzednio dla odpowiednich widm rotacyjnych i oscylacyjnych. W odniesieniu do tych częstości słuszne będzie to wszystko, co mówiliśmy o rotacji i oscylacji cząsteczek dwuatomowych i wieloatomowych, rozpatrywanych jako rotatory sztywne i niesztywne oraz oscylatory harmoniczne i anharmoniczne, drgania normalne itp.

linia stokesowska

linia antystokesowska

Na zakończenie warto jeszcze kilka słów powiedzieć na temat warunków pojawienia się linii czy pasma widma ramanowskiego. Podczas gdy warunkiem pojawienia się linii czy pasma w widmie absorpcji jest zmiana momentu dipolowego cząsteczki, warunkiem pojawienia się linii i pasma w widmie ramanowskim jest zmiana polaryzowalności $\alpha(q)$. To zróżnicowanie ogólnej reguły wyboru widma absorpcji i widma Ramana powoduje, że rotacja i niektóre drgania nieaktywne w absorpcji mogą być aktywne w widmie Ramana i na odwrót. Na przykład drganie i rotacja dwuatomowej cząsteczki homopolarnej jest nieaktywne w absorpcji, natomiast pojawia się w widmie Ramana, ponieważ rotacji i drganiu towarzyszy zmiana polaryzowalności. W szczególnym przypadku cząsteczek mających środek symetrii (np. CO_2) obowiązuje tzw. zakaz alternatywny, który mówi, że drgania nieaktywne w absorpcji w podczerwieni są aktywne w rozpraszaniu Ramana.

widmo Ramana a widmo absorpcji

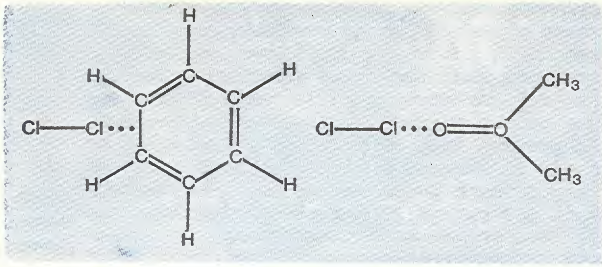
zakaz alternatywny

Wpływ oddziaływań międzycząsteczkowych na widma oscylacyjne i rotacyjne

Mówiliśmy już, że oddziaływania międzycząsteczkowe mają wpływ na obserwowane widma molekularne. W przypadku widm rotacyjnych, dla których energia oddziaływania z promieniowaniem elektromagnetycznym jest tego samego rzędu co energia oddziaływań międzycząsteczkowych, decydują one o istnieniu widma. Wpływ oddziaływań międzycząsteczkowych na widmo oscylacyjno-rotacyjne związany jest z tym, że oddziaływania te mogą w większym lub mniejszym stopniu zmieniać siły wiązań między atomami w cząsteczce, a więc zmieniać stałe siłowe, a tym samym częstości oscylacji. Z drugiej strony mogą one powodować zmiany rozkładu ładunków w cząsteczce, a więc zmiany momentu dipolowego, co znajduje wyraz w zmianach natężeń pasm. Oddziaływania międzycząsteczkowe wywołują często pojawienie się momentu dipolowego i wystąpienie w widmie odpowiednich pasm, które odpowiadają zmianom tego momentu podczas drgania. Występuje to w cząsteczkach takich, jak Cl_2, J_2, Br_2. Ze względu na zerowy moment dipolowy, cząsteczki te w fazie gazowej nie absorbują promieniowania podczerwonego i nie mają absorpcyjnego widma oscylacyjnego i rotacyjnego. Sytuacja zmienia się radykalnie w roztworach, gdy jako rozpuszczalnika użyjemy np. benzenu, acetonu czy pirydyny. W roztworach takich występują bowiem silne oddziaływania pomiędzy cząsteczkami chlorowców i rozpuszczalnikiem, prowadzące do powstania kom-

kompleksy międzycząsteczkowe

pleksów międzycząsteczkowych takich, jak np. Cl_2 — benzen czy Cl_2 — aceton:

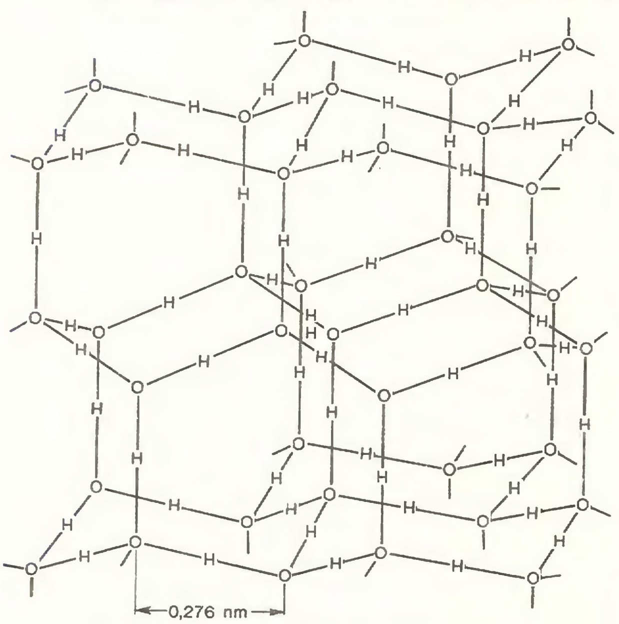


Powstanie tych kompleksów, zwanych kompleksami z przeniesieniem ładunku (*charge-transfer*) powoduje, że w cząsteczkach chlorowców następują zmiany rozkładu ładunku prowadzące do powstania momentu dipolowego. O strukturze i własnościach kompleksów z przeniesieniem ładunku będzie jeszcze mowa obszerniej w rozdziale omawiającym widma elektronowe.

wiązanie wodorowe

Innym typem oddziaływań międzycząsteczkowych badanych metodami spektroskopii oscylacyjnej są oddziaływania zwane wiązaniami wodorowymi. Nie jest przesadą stwierdzenie, że wiązanie wodorowe stanowi główny czynnik wzajemnego powiązania cząsteczek, w tym również cząsteczek układów biologicznych, oraz że decyduje ono w znacznym stopniu o strukturze materii ożywionej. Międzycząsteczkowe wiązanie wodorowe występuje pomiędzy dwiema cząsteczkami, z których jedna zawiera grupę protonodonorową X—H, druga zaś atom lub grupę atomów protonoakceptorowych. Symbolicznym oznaczeniem wiązania wodorowego jest zwykle zapis X—H···Y. Grupami protonodonorowymi mogą być

>O—H, >N—H, >S—H, Cl—H, Br—H, F—H, J—H oraz w mniejszym stopniu C—H. Protonoakceptorami mogą być azot, tlen, chlorowce i ich ujemne jony, siarka a czasami również inne atomy, np. węgiel. Energia wiązań wodorowych waha się

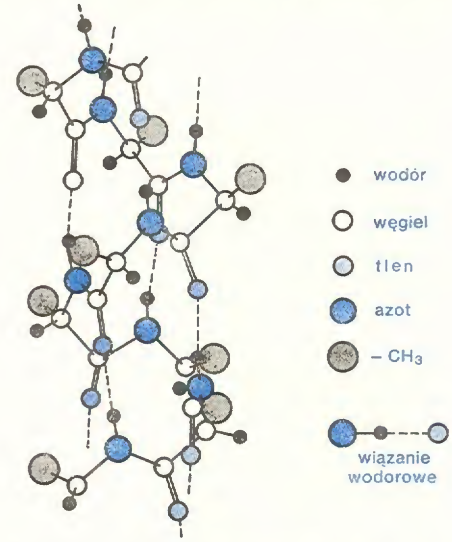


Rys. 24. Struktura lodu (lód heksagonalny)

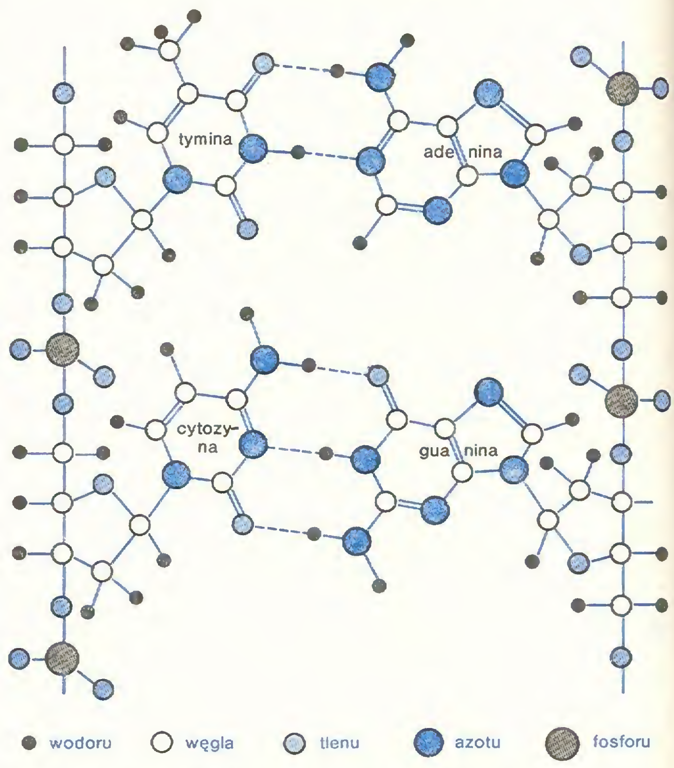
w granicach od kilku do kilkunastu, a nawet w niektórych przypadkach kilkudziesięciu kJ/mol. Własności wielu układów molekularnych określone są przez wiązania wodorowe. Np. w lodzie istnieje regularna sieć wiązań wodorowych (rys. 24). Podczas topnienia część wiązań się rozrywa i to jest przyczyną wzrostu gęstości w temperaturze topnienia.

układy biologiczne

Szczególne znaczenie ma wiązanie wodorowe w układach biologicznych, takich jak białka i kwasy nukleinowe. W białkach wiązania wodorowe typu N—H···O=C utrwalają skręty łańcuchów polipeptydowych (rys. 25). W kwasach nukleinowych, mających również helikalną budowę, łańcuchy tzw. komplementarne są do siebie „dopasowane" dzięki istnieniu wiązań wodorowych między parami zasad adeniną i tyminą oraz guaniną i cytozyną. Fragment kwasu dezoksyrybonukleinowego pokazany jest na rys. 26. Rola wiązań wodorowych łączących pary zasad jest jak widać bardzo duża.



Rys. 25. Fragment heliksu łańcucha polipeptydowego

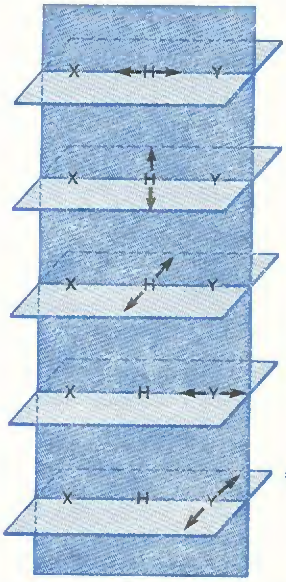


Rys. 26. Fragment komplementarnego łańcucha kwasu dezoksyrybonukleinowego (DNA)

absorpcja w podczerwieni

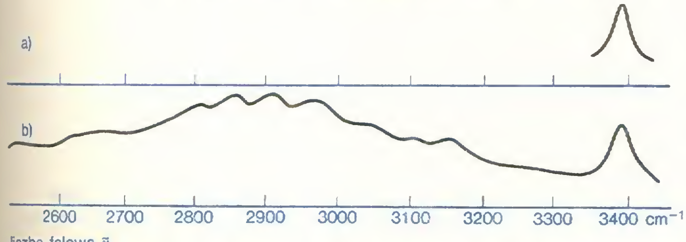
Istnienie wiązania wodorowego przejawia się bardzo wyraźnie w widmie absorpcji w podczerwieni. Widmo to dostarcza następujących ważnych informacji:

1. Wskazuje na obecność wiązania wodorowego — nawet bardzo słabego.
2. Informuje, jaki protonodonor *X*—H bierze udział w wiązaniu i z jakim protonoakceptorem *Y* oddziałuje *X*—H.
3. Mówi o roli środowiska otaczającego cząsteczki związane.
4. Pozwala wyznaczyć energię wiązania wodorowego.



Rys. 27. Możliwe drgania wiązania wodorowego

Drgania układu związanego wiązaniem wodorowym w przybliżeniu trójatomowym *X*—H···*Y* przedstawia rys. 27. Powstaniu wiązania wodorowego towarzyszy pojawienie się nowych częstości w widmie, a mianowicie częstości odpowiadających drganiom międzycząsteczkowego wiązania, przedstawionym na rys. 27 (*4*) oraz (*5*). Są to drgania o niskich częstościach w zakresie tzw. dalekiej podczerwieni (ok. 50–250 cm^{-1}). Oprócz pojawienia się tych nowych częstości, zmianom pod wpływem powstania wiązania wodorowego ulega również widmo donora i akceptora protonu. Największe zmiany zachodzą w obszarze częstości drgania rozciągającego donora protonu *X*—H (drganie *1* na rys. 27). Częstość tego drgania może obniżyć się od kilku do kilkuset cm^{-1} (niekiedy nawet więcej). Pasmo odpowiadające temu drganiu ulega też znacznemu poszerzeniu — nawet do setek cm^{-1}. Zwiększa się również bardzo, często o rząd lub dwa rzędy wielkości, natężenie pasma odpowiadającego temu drganiu. Wiązanie wodorowe powoduje również zauważalne zmiany w widmie akceptora protonu. Np. wiązanie C=O uczestniczące w oddziaływaniu obniża swą częstość. Ilustracją zmian w widmie pod wpływem powstania wiązania wodorowego jest rys. 28. Pokazuje on widmo absorpcji w podczerwieni w zakresie drgań rozciągających wiązanie N—H w cząsteczce znajdującej się w różnych roztworach, a mianowicie w roztworze, w którym nie tworzą się wiązania wodorowe typu N—H···O=C (rys. 28a), oraz w roztworze, w którym powstają te wiązania (rys. 28b). Szerokie pasmo o bogatej strukturze odpowiada cząsteczkom związanym wodorowo, wąskie pasmo odpowiada cząsteczkom nie oddziałującym ze sobą. Mierząc natężenie obu tych pasm w roztworach oraz badając wpływ temperatury na ich natężenia można wyznaczyć energię wiązań wodorowych.



Rys. 28. Widmo absorpcji w podczerwieni 4-keto, 6-metylopirymidyny w roztworze $CDCl_3$; a) małe stężenie badanych cząsteczek, b) większe stężenie badanych cząsteczek

Widma elektronowe cząsteczek

Jerzy Prochorow

O przejściach elektronowych mówimy wtedy, gdy w trakcie pochłaniania (absorpcji) lub wysyłania (emisji) kwantu promieniowania elektromagnetycznego przez cząsteczkę zmienia się energia elektronów cząsteczki. Odpowiednie widma absorpcji bądź emisji, związane z tymi przejściami, nazywamy widmami elektronowymi cząsteczki. Obszar, w którym występować mogą takie widma, rozciąga się od około 120 nm do 1000 nm, a więc obejmuje nadfiolet, obszar widzialny i bliską podczerwień.

Stany elektronowe cząsteczki

Stany energetyczne cząsteczki, a więc i elektronów w cząsteczce, są stanami dyskretnymi. Reguły, które rządzą przejściami między takimi stanami, można zrozumieć i wytłumaczyć tylko na podstawie mechaniki kwantowej. Dlatego w tym miejscu odsyłamy Czytelnika do działu „Chemia kwantowa", ażeby zapoznał się (o ile do tej pory jeszcze tego nie zrobił) z kwantowym opisem cząsteczki i podstawowymi pojęciami tego opisu, z których będziemy w dalszej części korzystali. Na razie spróbujemy tylko krótko uporządkować te pojęcia w taki sposób, abyśmy mogli zrozumieć istotę i prawidłowości rządzące przejściami i widmami elektronowymi — innymi słowy, zapoznamy się na razie z teorią przejść i widm elektronowych.

molekularna funkcja falowa

Stany energetyczne cząsteczki opisują molekularne funkcje falowe ψ_n, które są ścisłymi rozwiązaniami równania Schrödingera. Dla cząsteczek wieloelektronowych takie rozwiązania nie są w ogóle możliwe i do otrzymania i obliczenia energii stanów stacjonarnych cząsteczek wieloelektronowych korzysta się z przybliżonych metod. Jedną z nich jest metoda orbitali molekularnych.

Molekularna funkcja falowa ψ określa orbitę i własności elektronu w cząsteczce. Gęstość rozkładu ładunku elektronowego w każdym punkcie przestrzeni jest proporcjonalna do $|\psi|^2$. Wobec małych rozmiarów elektronu w stosunku do rozmiarów cząsteczki (a nawet w stosunku do rozmiarów atomu), chmura ładunku elektronu skupiona jest w pewnej ograniczonej przestrzeni. Dla każdego więc elektronu

orbital

można wydzielić pewną powierzchnię graniczną, wewnątrz której zawarta jest większość ładunku elektronu, np. 95%. I to jest właśnie wyobrażenie orbitala elektronu. W chemii kwantowej mówi się po prostu o funkcji falowej elektronu jako o orbitalu — wszak jej związek z położeniem elektronu w przestrzeni jest oczywisty!

Złożone wieloatomowe cząsteczki mają dużą liczbę takich orbitali, z których każdy opisany jest odpowiednią funkcją falową ψ. Orbital ψ może być np. zlokalizowany w znacznym stopniu na jednym z jąder cząsteczki i jest to wtedy orbital atomowy, może też obejmować dwa lub kilka jąder jednocześnie i wtedy jest to orbital molekularny. Każdy orbi-

tal w cząsteczce może być, zgodnie z zakazem Pauliego, obsadzony co najwyżej przez dwa elektrony (a w takim przypadku muszą się one różnić funkcją spinową — ich spiny muszą być sparowane, porównaj rys. 41).

Oddziaływanie cząsteczki z promieniowaniem elektromagnetycznym

Tak jak w przypadku widm oscylacyjnych czy rotacyjnych, również i przy przejściach elektronowych cząsteczka znajdująca się w stanie ψ_n może, absorbując kwant promieniowania, przejść do stanu ψ_m, jeżeli spełniony jest warunek Bohra: $E_n - E_m = h\nu$, tzn. jeżeli różnica energii obu stanów elektronowych jest równa energii kwantu promieniowania oddziałującego z cząsteczką. Musi być również spełniony pewien dodatkowy warunek. W pierwszym przybliżeniu oddziaływanie pola elektrycznego promieniowania z elektronami cząsteczki jest oddziaływaniem wektora elektrycznego z elektrycznym momentem dipolowym cząsteczki. I stąd dodatkowy warunek, który mówi, że aby możliwe było przejście elektronowe ze stanu ψ_n do stanu ψ_m, wielkość zwana dipolowym momentem przejścia musi być różna od zera. Wielkość ta obrazująca przesunięcia ładunku, lub inaczej zmiany momentu dipolowego, wyraża się następująco:

dipolowy moment przejścia

$$R_{nm} = \int \psi_n R \psi_m \, d\tau,$$

gdzie $R = e\sum_i r_i$ jest tzw. operatorem elektrycznego momentu dipolowego; r_i jest współrzędną i-tego elektronu, a sumowanie obejmuje wszystkie elektrony; e jest ładunkiem elektronu. Jeżeli warunek ten jest spełniony, to może nastąpić przejście elektronowe, które nazywamy przejściem dipolowym elektrycznym, a prawdopodobieństwo takiego przejścia jest proporcjonalne do $|R_{nm}|^2$.

przejście dipolowe elektryczne

Oddziaływanie promieniowania z układem cząsteczek może prowadzić do trzech zasadniczych typów przejść dipolowych elektrycznych. Cząsteczka może absorbować promieniowanie i jej energia zostaje podwyższona, mówimy wówczas, że cząsteczka została wzbudzona (albo że jest w stanie wzbudzonym). Cząsteczka w stanie wzbudzonym może tracić swoją energię wysyłając kwant promieniowania, który unosi ze sobą traconą przez cząsteczkę energię. Prawdopodobieństwo pierwszego procesu — absorpcji — jest wprost proporcjonalne do gęstości energii $\varrho(\nu)$ promieniowania o częstości ν i do liczby cząsteczek n_i w stanie o niższej energii, z którego następuje absorpcja.

Prawdopodobieństwo drugiego procesu — emisji spontanicznej — jest proporcjonalne do liczby cząsteczek n_j w stanie o wyższej energii, tzn. do liczby cząsteczek wzbudzonych. Trzeci wspomniany proces to emisja wymuszona, której prawdopodobieństwo jest proporcjonalne zarówno do n_j, jak i do $\varrho(\nu)$. Ten ostatni proces jest mało istotny w przypadku rozważanych przez nas widm elektronowych, ma natomiast kapitalne znaczenie dla pracy lasera (→ Lasery — podstawy działania).

współczynniki Einsteina

Nie będziemy próbowali wyprowadzać związków pomiędzy wspomnianymi tutaj prawdopodobieństwami. Podamy je tylko za Einsteinem, który pierwszy je wyprowadził (nazywane są czasem dlatego współczynnikami Einsteina). Jeżeli przez B_{ij}, B_{ji} i A_{ji} oznaczymy odpowiednio prawdopodobieństwo absorpcji, emisji wymuszonej i emisji spontanicznej, przez ν — częstość promieniowania, przez c — prędkość światła, a przez h — stałą Plancka, to wielkości te wiążą się ze sobą następująco:

$$B_{ij} = B_{ji} = (c^3/8\pi h\nu^3)A_{ji}.$$

Prawdopodobieństwo przejścia absorpcyjnego, jak już wspominaliśmy poprzednio, wiąże się z dipolowym momentem przejścia, tak że

$$B_{ij} = (2\pi/3\hbar^2)\,|R_{ij}|^2\ (\hbar = h/2\pi),$$

a z tych dwóch równań wynika, że prawdopodobieństwo emisji spontanicznej

$$A_{ji} = (64\pi^4/3hc^3)\nu^3|R_{ji}|^2.$$

Te dwa równania pokazują, że zarówno absorpcja, jak i emisja zależą od momentu przejścia. Niekiedy, chociaż rzadko, można ocenić moment przejścia bez konieczności wykonywania szczegółowych rachunków. Czasem funkcje falowe stanów uczestniczących w przejściu elektronowym są takie, że moment przejścia jest równy zeru. Wtedy mówimy, że przejście jest przejściem wzbronionym i nie powinno być obserwowane. Jeżeli natomiast moment przejścia jest różny od zera, to przejście jest dozwolone i w zasadzie powinniśmy móc je obserwować. O tym czy jakieś przejście jest wzbronione, czy też nie, mówią reguły wyboru. Określają one warunki, w których dipolowy moment przejścia $R_{ij} = \int \psi_i R \psi_j d\tau = 0$.

reguły wyboru

Można podać szereg reguł wyboru, mniej lub bardziej ogólnych. Dla przykładu przytoczymy dwie z nich: 1) wzbronione są przejścia, podczas których następuje zmiana spinu elektronu; 2) wzbronione są przejścia pomiędzy takimi stanami elektronowymi, których funkcje falowe nie spełniają pewnych określonych warunków symetrii (dokładne rozważenie tych warunków symetrii wymaga znajomości specjalnej teorii, zwanej teorią grup, i dlatego, niestety, nie będziemy mogli tego tutaj zrobić).

Można podać inne reguły wyboru, narzucające warunki np. na nakrywanie się orbitali, czy na zmiany momentu pędu cząsteczki. Należy jednak pamiętać o tym, że wyprowadzenie ścisłych reguł wyboru wymaga znajomości ścisłych funkcji falowych (będących ścisłymi rozwiązaniami równania Schrödingera), a tymi na ogół nie dysponujemy. Dlatego też wspomniane reguły wyboru wyprowadzane są przy pewnych założeniach upraszczających i nie mają wcale sensu absolutnie ścisłych nakazów, tzn. przejścia, które są wzbronione przez te reguły, nie są naprawdę ściśle wzbronione i mogą być w odpowiednich warunkach obserwowane. Po prostu ich prawdopodobieństwo, a tym samym i ich natężenie jest z reguły bardzo małe. I tak np. reguła wyboru ze względu na spin ma charakter reguły ścisłej tylko wtedy, gdy w cząsteczce nie występuje sprzężenie spin-orbita (będziemy jeszcze o tym sprzężeniu mówili szerzej), tymczasem sprzężenie takie jest w praktyce zawsze obecne w cząsteczkach i wobec tego przejścia elektronowe zakazane ze względu na spin są jednak w widmach obserwowane (przede wszystkim jako widma emisyjne). Podobnie ma się rzecz z przejściami zakazanymi ze względu na symetrię. Tę ostatnią regułę wyprowadza się nie uwzględniając faktu, że wraz z przejściami elektronowymi wzbudzane są również przejścia oscylacyjne, a te ostatnie mogą zmieniać symetrię. Mówiąc inaczej, nawet jeśli reguła ta jest spełniana w odniesieniu do prostych stanów elektronowych, to nie musi być ona spełniana dla stanów elektronowo-oscylacyjnych (nazywamy je stanami wibronowymi). Tak więc reguły wyboru mają ograniczony charakter i wskazują raczej na to, czy dla danego przejścia można oczekiwać dużego czy małego natężenia. Na ogół, jeśli spróbujemy porównać natężenie przejść wzbronionych z natężeniem przejść dozwolonych i jeśli przyjmiemy natężenie tego ostatniego jako równe 1, to natężenie przejść wzbronionych ze względu na symetrię jest około 10 do 100 razy słabsze, a natężenie przejść wzbronionych ze względu na spin o około 10 000 do 1 000 000 razy słabsze.

struktura oscylacyjna widm elektronowych

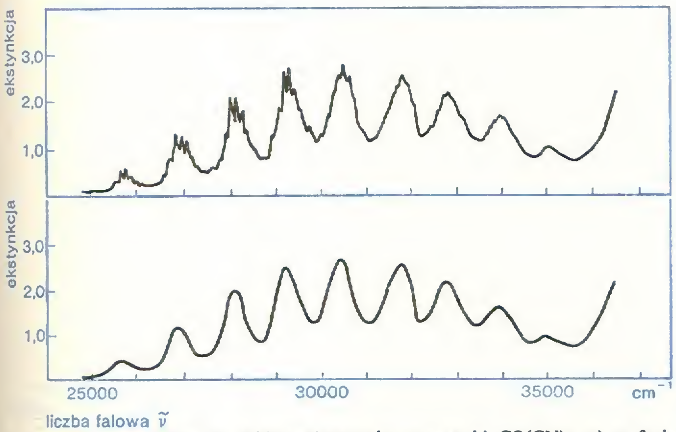
Wspomnieliśmy przed chwilą, że przejściu elektronowemu towarzyszy zmiana energii oscylacyjnej, a więc i przejście oscylacyjne. Jeżeli wrócimy jeszcze raz do rys. 1 ze str. 301, który ilustrował różne po-

ziomy energetyczne cząsteczki, to dojdziemy do wniosku, że w trakcie przejścia elektronowego zmieniać się na pewno może i energia rotacyjna cząsteczki. Stąd wniosek, że widmo elektronowe jest z reguły widmem złożonym: elektronowo-oscylacyjno-rotacyjnym. Bez trudu chyba też uzmysłowimy sobie fakt, jak skomplikowane musi być takie widmo, w którym na przejście elektronowe nałożone są dodatkowo wszystkie sytuacje, które poznaliśmy omawiając widma oscylacyjne i rotacyjne, i jak skomplikowana musi być analiza takich widm. Zdamy sobie jeszcze lepiej z tego sprawę, kiedy spojrzymy na zdjęcie (tzw. spektrofotogram, il. 132, tabl. 33) bardzo niewielkiego, bo obejmującego zaledwie 3 nm wycinka widma elektronowo-oscylacyjnego cząsteczki NCO z uwidocznioną strukturą rotacyjną.

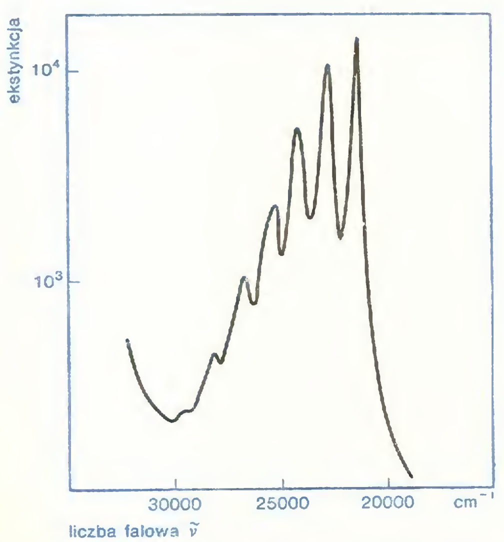
struktura rotacyjna widma elektronowego

W praktyce tak skomplikowane widmo elektronowe z widoczną strukturą rotacyjną jest obserwowane bardzo rzadko i tylko dla małych (głównie dwuatomowych) cząsteczek. Gdy cząsteczka wieloatomowa znajduje się w fazie gazowej i pod bardzo niskim ciśnieniem, to przy bardzo dużych zdolnościach rozdzielczych aparatury, można niekiedy obserwować widmo z wyraźną strukturą rotacyjną. Ponieważ w przeważającej większości wypadków widma dużych cząsteczek badane są w roztworze (lub w fazie stałej)- wobec tego widma absorpcyjne mają charakter szerokich pasm, bez śladów struktury rotacyjnej, z zaznaczoną czasem strukturą oscylacyjną. Rys. 29 ilustruje zmiany w wyglądzie widma cząsteczki obserwowanej wpierw w fazie gazowej a następnie w roztworze. Widać od razu, że przejście od gazu do roztworu spowodowało zamazanie się większości szczegółów struktury. W większości bardzo dużych cząsteczek, nawet tak stosunkowo dobrze rozdzielona, jak na rys. 29, struktura oscylacyjna nie jest obserwowana w roztworze.

widmo cząsteczki w fazie gazowej i w roztworze



Rys. 29. Widma absorpcyjne cząsteczki $CO(CN)_2$; a) w fazie gazowej, b) w roztworze



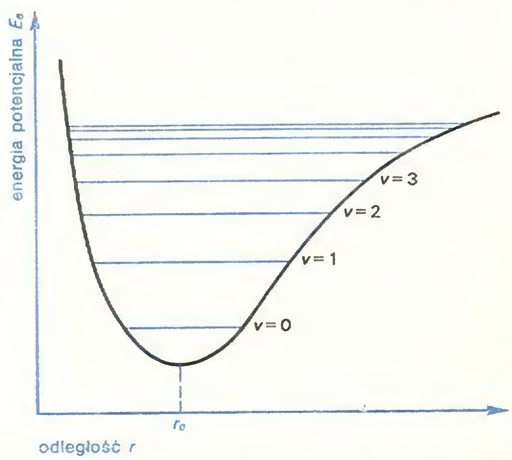
Rys. 30. Długofalowe elektronowe pasmo absorpcyjne cząsteczki naftacenu

Przyjrzyjmy się jeszcze raz rys. 29b. Widać, że w obrębie obserwowanego pasma elektronowego, poszczególne przejścia oscylacyjne mają różne natężenie, jedne są bardzo słabe, inne mają większe natężenie, a największe natężenie mają przejścia oscylacyjne zgrupowane w środkowej części widma. Czy tak zawsze wyglądają widma elektronowe? Otóż nie. Na rys. 30 pokazane jest jedno z pasm elektronowych cząsteczki naftacenu. Widzimy, że w tym wypadku największe natężenie mają pasma oscylacyjne z lewej strony rysunku, są one silniejsze niż pasma znajdujące się w części centralnej pasma elektronowego.

Czy taka informacja o różnym rozkładzie natężeń przejść oscylacyjnych w obrębie pasma elektronowego ma dla nas jakieś praktyczne znaczenie, czy możemy ją wykorzystać do określenia jakichś własności podstawowych cząsteczki, takich jak jej struktura czy geometria? Żeby odpowiedzieć na to pytanie, zastanówmy się, jak w ogóle powstają takie widma, i spróbujmy ustalić związki między obserwowanymi widmami a najprostszymi parametrami cząsteczki (np. długością wiązań). Zrobimy to na przykładzie najprostszym, a mianowicie dwuatomowej cząsteczki AB.

Z rozważań nad oscylacjami cząsteczek wiemy już, że taka cząsteczka ma $3 \cdot 2 - 5 = 1$ drganie normalne. przejawiające się jako zmiana długości wiązania A—B, Wiemy również, jaka może być częstość takiego drgania i jaka jest odległość równowagi r_e w stanie podstawowym. Wiemy, że drgania cząsteczki są skwantowane i że energię potencjalną cząsteczki w podstawowym stanie elektronowym możemy przybliżyć pewną krzywą, zwaną krzywą energii potencjalnej (rys. 31).

krzywa energii potencjalnej



Rys. 31. Schematyczny przebieg rzeczywistej krzywej energii potencjalnej cząsteczki dwuatomowej. Poziomy oscylacyjne zagęszczają się wraz ze wzrostem kwantowej liczby oscylacyjnej v

Krzywa taka ma minimum przy odległości równowagi r_e. Jest to odległość, przy której energia potencjalna drgań jest równa zeru, maksymalna jest natomiast energia kinetyczna. Odwrotnie, w punktach na krzywej (dla danego poziomu oscylacyjnego) energia potencjalna drgań jest maksymalna, a energia kinetyczna jest równa zeru. Te punkty są punktami zwrotnymi oscylacji — po maksymalnym ściśnięciu wiązania A—B (punkt po lewej stronie r_e) następuje jego rozciąganie, aż do wartości maksymalnej (po prawej stronie r_e), a następnie cały cykl oscylacji powtarza się.

Co się dzieje, kiedy cząsteczka zostaje wzbudzona do innego, wyższego stanu elektronowego? Zmienia

wpływ wzbudzenia cząsteczki

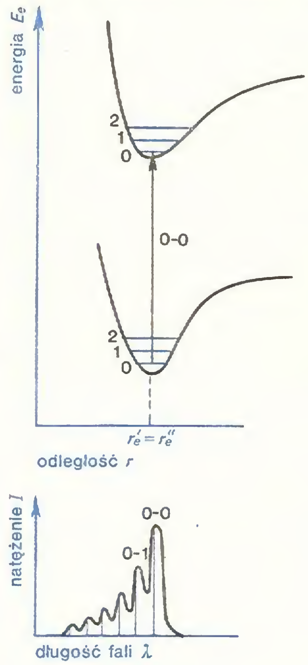
się rozkład ładunku elektronowego w cząsteczce, a to oznacza, że zmieniają się określone siły utrzymujące razem A i B; mogą się one osłabiać lub wzmacniać, a to z kolei oznacza, że długość wiązania, a więc i odległość równowagi we wzbudzonym stanie elektronowym może ulegać zmianie — skróceniu lub wydłużeniu. Mogą również zmieniać się i inne cechy krzywej energii potencjalnej — może mieć ona inny kształt i inną głębokość. To oczywiście prowadzić może do zmiany częstości drgań w stanie wzbudzonym i do innego rozkładu poziomów oscylacyjnych w tym stanie, w stosunku do stanu podstawowego. A zatem krzywe energii potencjalnej w różnych stanach elektronowych są różne i mogą mieć różne położenie minimum (różne odległości równowagi), a bywa i tak, że w ogóle nie mają minimum. Jak się te różnice odbijają na widmach, zobaczymy zaraz na przykładach. Przedtem jednak podamy pewną zasadę o podstawowym znaczeniu dla teorii widm. Zasada ta, zwana zasadą Francka–Condona, bierze swój początek z prostego porównania czasów, w jakich mogą przebiegać pewne określone procesy w cząsteczce. I tak proces przejścia elektronowego, a więc proces przeskoku elektronu w obrębie cząsteczki, jest procesem bardzo szybkim i można oszacować, iż trwa on około 10^{-15} s. Natomiast średni czas trwania oscylacji, tzn. czas w jakim zmieniają swe położenia atomy cząsteczki, jest rzędu 10^{-12} s, a więc jest średnio tysiąc razy dłuższy niż czas przeskoku elektronowego.

zasada Francka–Condona

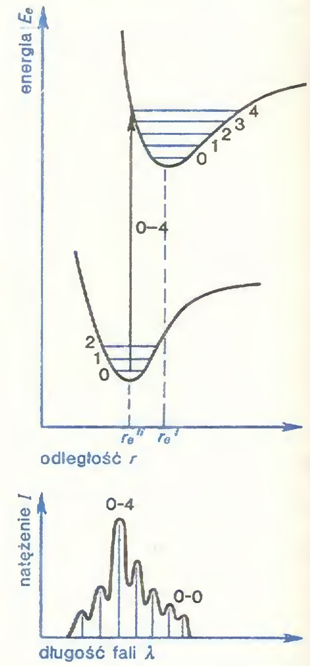
Uwzględniając ten fakt zasada Francka–Condona stwierdza, że w trakcie przejścia elektronowego jądra atomowe w cząsteczce nie zmieniają w praktyce ani swych położeń ani swoich prędkości. Oznacza to, że spośród wszystkich możliwych przejść elektronowo-oscylacyjnych pomiędzy dwoma stanami elektronowymi najbardziej prawdopodobne są te przejścia, w wyniku których nie zmienia się położenie lub, jak czasem mówimy, konfiguracja jąder, ani ich moment pędu.

Wrócimy teraz do zapowiedzianych przykładów. Rozpatrzymy trzy przypadki: a) kiedy odległość równowagi w stanie wzbudzonym r_e' jest taka sama, jak odległość równowagi w stanie podstawowym r_e'', b) kiedy odległość równowagi w stanie wzbudzonym r_e' jest większa niż odległość równowagi w stanie podstawowym r_e'' i c) kiedy r_e' jest mniejsza od r_e''.

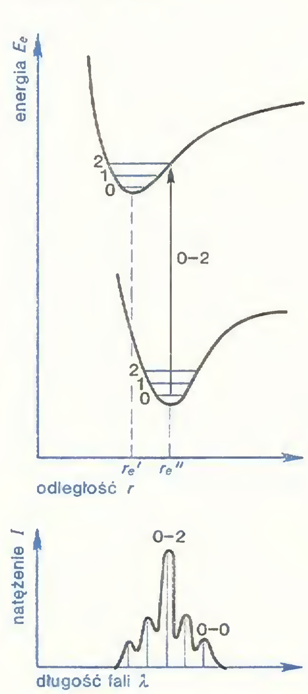
Przypadek a) ilustruje rys. 32. Wiemy, że obsadzenie poszczególnych poziomów oscylacyjnych w stanie podstawowym zależy od temperatury i że w temperaturach niezbyt wysokich w stosunku do temperatury pokojowej rozkład obsadzeń jest taki, iż praktycznie wszystkie cząsteczki są w najniższym stanie oscylacyjnym (o kwantowej liczbie oscylacyjnej $v'' = 0$). W tej sytuacji prawie wszystkie przejścia elektronowo-oscylacyjne ze stanu podstawowego do wzbudzonego stanu elektronowego biorą początek ze stanu oscylacyjnego $v'' = 0$. Ponieważ, w myśl zasady Francka–Condona, przy przejściu elektronowym do wyższego stanu zarówno odległość międzyjądrowa, jak też i prędkość ruchu jąder nie powinna ulegać wyraźnym zmianom, wobec tego możliwe są przejścia, które zaznaczono na rys. 32 linią pionową (lub tylko nieznacznie odchyloną od pionu). Tak więc będzie to przede wszystkim przejście do stanu oscylacyjnego $v' = 0$ górnego stanu elektronowego. Jest to najbardziej prawdopodobne przejście. Inne kolejne przejścia do stanów oscylacyjnych $v' = 1, 2, 3, \ldots$ są coraz mniej prawdopodobne, gdyż wymagają albo coraz większej zmiany konfiguracji jąder w trakcie przejścia (coraz większe odchylenie strzałek, wskazujących przejście, od kierunku pionowego), albo dużej zmiany pędu jąder przy tej samej konfiguracji, a na możliwość poważniejszych zmian którejkolwiek z tych wielkości narzuca ograniczenia zasada Francka–Condona. Nie oznacza to, że przejścia ze stanu podstawowego (z poziomu oscylacyjnego $v'' = 0$) do wyższych poziomów oscylacyjnych ($v' = 1, 2, 3, \ldots$) stanu wzbudzonego są absolutnie niemożliwe. Przejścia takie mogą zachodzić, ale ich prawdopodobieństwo w stosunku do prawdopodobieństwa przejścia z $v'' = 0$ do $v' = 0$ jest bardzo małe i coraz mniejsze w miarę wzrostu kwantowej liczby oscylacyjnej w stanie wzbudzonym v'. W rezultacie widmo, które powinniśmy zaobserwować w rozpatrywanym przypadku, powinno wyglądać tak, jak to pokazano u dołu rys. 32. Największe natężenie ma przejście oscylacyjne z poziomu $v'' = 0$ dolnego stanu do poziomu $v' = 0$ stanu górnego albo, jak po prostu często mówimy, przejście 0–0 (przejście



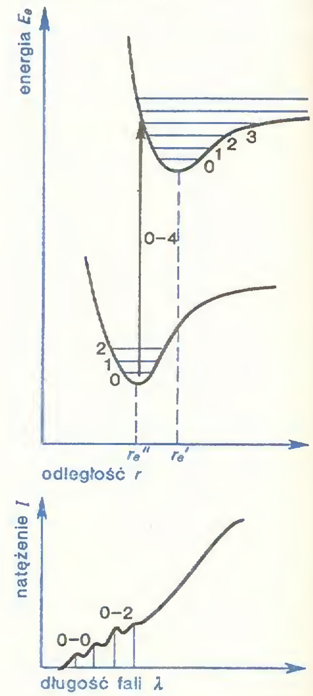
Rys. 32. Ilustracja zasady Francka–Condona: $r_e' = r_e''$



Rys. 33. Ilustracja zasady Francka–Condona: $r_e' > r_e''$



Rys. 34. Ilustracja zasady Francka–Condona: $r_e' < r_e''$



Rys. 35. Ilustracja zasady Francka–Condona; powstawanie widma ciągłego

zero–zero). Natomiast natężenia przejść 0–1, 0–2, 0–3, ... bardzo szybko maleją.

W podobny sposób możemy rozpatrzyć przypadek b), kiedy odległość równowagi w stanie wzbudzonym jest większa niż w stanie podstawowym (rys. 33). Również teraz przejścia biorą swój początek przede wszystkim ze stanu oscylacyjnego $v'' = 0$, stanu podstawowego. Tym razem jednak, zgodnie z zasadą Francka–Condona, największe natężenie ma przejście 0–4, a nie przejście 0–0 i wobec tego widmo wygląda inaczej niż poprzednio, jak to widać u dołu rysunku. Nie będziemy analizować szczegółowo przypadku c), gdy $r'_e < r''_e$, i uzasadniać wyglądu obserwowanego wówczas widma — rys. 34. (Czytelnik sam już doskonale potrafi to w tej chwili zrobić.) Zajmiemy się innym przykładem, który jest często obserwowany w widmach cząsteczek. Ilustruje go rys. 35. Tym razem wzajemne położenie krzywych energii potencjalnych dwóch stanów, pomiędzy którymi zachodzi przejście elektronowe, jest takie, że z dużym prawdopodobieństwem mogą zachodzić przejścia prowadzące do obszaru energii leżącej powyżej granicy dysocjacji w stanie wzbudzonym. Amplituda oscylacji cząsteczki wzbudzonej do takiego obszaru energii jest nieskończenie duża albo, mówiąc inaczej, wiązanie ulega rozerwaniu i cząsteczka rozdziela się na atomy A i B. Energia oscylacji zamienia się na energię kinetyczną obu fragmentów A i B i nie ma już mowy o skwantowanych, dyskretnych poziomach energii takiego układu. Takie przejścia prowadzą w konsekwencji do widma ciągłego, a ich rezultatem jest dysocjacja (fotodysocjacja) cząsteczki. Obraz widma, jakiego wówczas możemy się spodziewać, pokazany jest na rys. 35 u dołu.

dysocjacja

Nasze rozważania nad zasadą Francka–Condona i jej wykorzystaniem do interpretacji widm prowadziliśmy na przykładzie cząsteczki dwuatomowej. Jeśli chodzi o cząsteczki wieloatomowe sytuacja jest bardziej skomplikowana, a to dlatego, że liczba możliwych rodzajów oscylacji (liczba drgań normalnych) wzrasta ze wzrostem liczby atomów w cząsteczce. Do opisu takich złożonych sytuacji nie nadaje się dwuwymiarowa krzywa energii potencjalnej, sporządzana w funkcji długości jednego wiązania. Krzywą taką należałoby zastąpić wielowymiarowymi powierzchniami energii potencjalnej i taki prosty, graficzny sposób analizy, jaki powyżej przeprowadziliśmy, przestaje być w ogóle możliwy. W cząsteczkach wieloatomowych na strukturę oscylacyjną pasma elektronowego składać się mogą przejścia oscylacyjne ze wzbudzeniem różnych rodzajów oscylacji (różnych drgań normalnych, jak mówiliśmy przy omawianiu oscylacji cząsteczek), o bardzo różnych częstościach. Taka właśnie sytuacja występuje w widmie, które pokazano na rys. 29, gdzie w obrębie pasma elektronowego można zidentyfikować dwie częstości oscylacyjne: ok. 1250 cm^{-1}, związaną z drganiami rozciągającymi w grupie karbonylowej C=O (w stanie wzbudzonym), i ok. 125 cm^{-1}, związaną prawdopodobnie ze zmianą kąta, jaki tworzą ze sobą w cząsteczce dwa wiązania —C≡N.

widmo cząsteczki wieloatomowej

Mimo licznych trudności, zasadę Francka–Condona po przyjęciu różnych uproszczeń można zastosować także do analizy widm elektronowych cząsteczek wieloatomowych. Dodajmy jeszcze dla uzupełnienia, że zasada Francka–Condona, którą podaliśmy i omawialiśmy w jej klasycznym sformułowaniu, ma również swoje sformułowanie w ramach mechaniki kwantowej. Z przybliżenia Borna–Oppenheimera (→ Chemia kwantowa) wiemy, że funkcje falowe stanów cząsteczki można przedstawić jako iloczyn odpowiednich funkcji elektronowych φ_e (zależnych od współrzędnych elektronów i jąder) i funkcji oscylacyjnych χ_v (zależnych tylko od współrzędnych jąder), tzn. $\psi = \varphi_e \chi_v$. Jeżeli przypomnimy sobie, że natężenie przejść jest proporcjonalne do kwadratu dipolowego momentu przejścia, to korzystając z tego faktu możemy napisać, iż natężenie I przejścia absorpcyjnego ze stanu oscylacyjnego $\chi_{e''}$ elektronowego stanu $\varphi_{e''}$ do stanu oscylacyjnego $\chi_{v'}$ elektronowego stanu $\varphi_{e'}$ jest proporcjonalne do $|\int \varphi_{e'} R \varphi_{e''} d\tau|^2 \cdot |\int \chi_{v'} \chi_{v''} d\tau|^2$, gdzie R jest znanym już nam operatorem momentu dipolowego. Druga całka zawiera tylko funkcje falowe stanów oscylacyjnych (jest, jak mówimy, całką nakrywania się tych funkcji falowych) i jest tym większa, im większe jest nakrywanie obu funkcji, co ma z reguły miejsce właśnie w punktach zwrotnych oscylacji (gdzie amplituda oscylacyjnych funkcji falowych jest największa, z wyłączeniem stanu $v = 0$; oscylacyjne funkcje falowe pokazane były na rys. 16 w artykule „Chemia kwantowa"). Sformułowanie to potwierdza wniosek, że najbardziej prawdopodobne będą te przejścia, które zaznaczyliśmy na rys. 32–35 pionowo.

zasada Francka–Condona w sformułowaniu kwantowym

Zasada Francka–Condona ma nie tylko duże znaczenie dla zrozumienia postaci obserwowanego widma i powiązania tej postaci ze zmianami podstawowych parametrów cząsteczki, ale jest również ważna dla zrozumienia wielu innych procesów (o niektórych z nich będziemy jeszcze dalej mówili), w których udział biorą wzbudzone stany elektronowo-oscylacyjne (nazywane czasem stanami wibronowymi).

Przedstawiliśmy tylko niezbędne teoretyczne podstawy interpretacji absorpcyjnych widm elektronowych cząsteczek. Teraz zajmiemy się omówieniem niektórych, często spotykanych w praktyce widm.

Widma absorpcyjne

Pierwszym i naturalnym podziałem, z którego zresztą korzystaliśmy już poprzednio, jest podział cząsteczek na małe (przede wszystkim dwuatomowe) i duże. Może to brzmieć nieco paradoksalnie, ale badanie i interpretacja widm prostych, dwuatomowych cząsteczek bywa znacznie bardziej złożonym zagadnieniem niż badanie widm cząsteczek wieloatomowych. Znacznie mniejsza liczba dozwolonych poziomów oscylacyjnych i rotacyjnych, ze stosunkowo dużą odległością między tymi poziomami powoduje, że przy odpowiednich warunkach obserwacji (w fazie gazowej i przy dużej zdolności rozdzielczej) widmo elektronowe małej cząsteczki ujawnia ogromne bogactwo struktury, tak jak to mogliśmy oglądać na il. 132 (tabl. 33). Interpretacja takiego widma, a więc przypisanie poszczególnych przejść oscylacyjnych i rotacyjnych, jest z reguły bardzo żmudnym przedsięwzięciem, chociaż wykonalnym i dostarczającym bardzo istotnych informacji o stałych siłowych i rotacyjnych stanów wzbudzonych i charakterze sprzężeń momentów pędu w cząsteczce. Nie będziemy się jednak zajmować omawianiem i interpretacją takich widm, można to bez trudu znaleźć w każdej książce traktującej o spektroskopii molekularnej.

widma cząsteczek małych

Mówiliśmy już o tym, że dla cząsteczek wieloatomowych konstruowanie krzywych energii potencjalnej w różnych stanach elektronowych jest w praktyce niewykonalne. Energia potencjalna odnosząca się do danego stanu elektronowego cząsteczki wieloatomowej jest zależna od długości wszystkich wiązań i kątów między wiązaniami i do zilustrowania takich przypadków są potrzebne nie krzywe lecz wielowymiarowe powierzchnie energii potencjalnej. Niemożliwa jest zatem analiza widm, taka jaką można przeprowadzić dla cząsteczek dwuatomowych, ze szczegółowym przypisaniem struktury oscylacyjnej i rotacyjnej, tym bardziej, że struktura taka nie jest na ogół obserwowana. Widma elektronowe są szerokimi, prawie pozbawionymi szczegółów pasmami. W tej sytuacji można tylko próbować połączyć obserwowane pasma absorpcyjne z pewnymi zmianami rozkładu ładunku elektrycznego w cząsteczce, a określenie rzeczywistej struktury stanu wzbudzonego byłoby już wielkim osiągnięciem.

widma cząsteczek dużych

Nasze rozważania o widmach elektronowych zaczniemy od widm cząsteczek organicznych. W cząsteczkach tych przejścia elektronowe są przejściami

widma cząsteczek organicznych

elektronów walencyjnych wiązań pojedynczych lub wielokrotnych, bądź też elektronów wolnych par elektronowych, a więc elektronów σ, π lub n (tzn. elektronów znajdujących się na orbitalach wiążących σ i π lub na niewiążącym orbitalu n). W przejściu elektronowym elektrony te mogą być wzbudzone do antywiążących orbitali σ^* i π^*. W cząsteczkach zatem mogą być obserwowane pasma absorpcyjne związane z przejściami typu $\sigma \rightarrow \sigma^*$, $\pi \rightarrow \pi^*$, $n \rightarrow \sigma^*$ i $n \rightarrow \pi^*$ (odczytujemy je jako sigma-sigma z gwiazdką, pi-pi z gwiazdką itd.), w zależności od tego, jakiego typu elektrony posiada cząsteczka: σ, π czy n.

Niekiedy poważnym ułatwieniem przy interpretacji widm elektronowych cząsteczek organicznych może być występowanie w cząsteczce grup chromoforowych. Grupy chromoforowe (są to zespoły atomów i wiązań) mają swoją charakterystyczną absorpcję, która w małym stopniu zależy od pozostałych elementów cząsteczki i przez to ten sam chromofor występujący w różnych cząsteczkach może być stosunkowo łatwo zidentyfikowany. Zajmijmy się teraz nieco bliżej omówieniem charakteru widm pewnych klas związków i grup chromoforowych, próbując powiązać je z odpowiednimi przejściami elektronowymi, ze zmianami struktury elektronowej cząsteczki i z efektami uwarunkowanymi budową różnych cząsteczek.

grupy chromoforowe

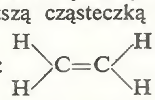
Charakterystyczną cechą pewnych związków organicznych jest to, że wszystkie wiązania w cząsteczce są wiązaniami pojedynczymi. Do tej klasy związków należą nasycone węglowodory — alkany i cykloalkany. Jedynymi możliwymi przejściami elektronowymi w cząsteczkach takich związków są przejścia elektronowe z wiążących orbitali σ na antywiążące orbitale σ^*, tzn. przejścia $\sigma \rightarrow \sigma^*$. Z przejściami $\sigma \rightarrow \sigma^*$ związana jest stosunkowo duża zmiana energii elektronowej cząsteczki i wobec tego widma elektronowe takich cząsteczek leżą daleko w nadfiolecie (w tzw. nadfiolecie próżniowym). Cząsteczki związków nasyconych nie są specjalnie interesujące ze spektroskopowego punktu widzenia, a badanie ich widm jest bardzo trudne, ze względu na obszar, w którym te widma występują. Jeżeli do cząsteczki nasyconej wprowadzony zostanie jakiś atom lub grupa z wolnymi parami elektronowymi (elektrony n), to wówczas możliwe jest przejście elektronu wolnej pary (niewiążącego) na antywiążący orbital σ^*, tzn. przejście $n \rightarrow \sigma^*$. Przejścia $n \rightarrow \sigma^*$ mają niższą energię niż przejścia $\sigma \rightarrow \sigma^*$ i w widmie cząsteczki możemy wówczas zaobserwować bardziej długofalową absorpcję. Węglowodory nasycone, chociaż niezbyt interesujące dla badacza, są bardzo ważne w badaniach spektroskopowych, gdyż używa się ich jako rozpuszczalników, w których badane są widma innych związków.

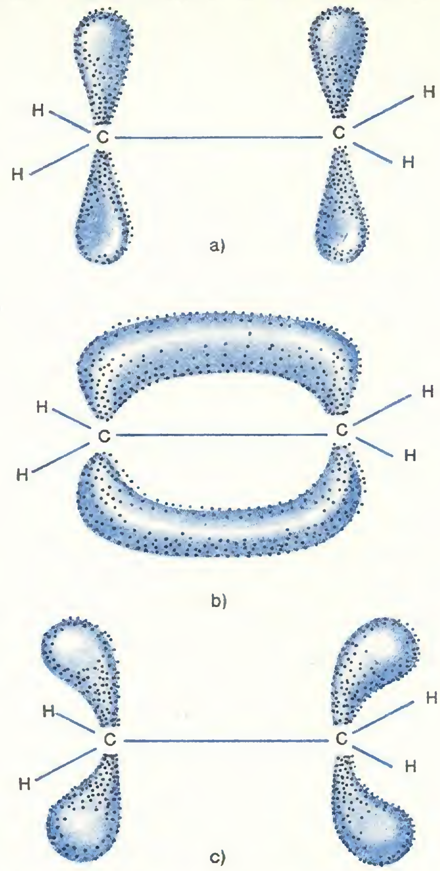
przejście $\sigma \rightarrow \sigma^*$

przejście $n \rightarrow \sigma^*$

Znacznie bardziej interesujące są cząsteczki zawierające nie tylko elektrony σ, ale również elektrony π lub n, bądź też jedne i drugie jednocześnie. Cząsteczki takie, to węglowodory nienasycone (np. alkeny i polieny) i aromatyczne (oraz ich pochodne), a więc cząsteczki, w których występują wiązania podwójne.

cząsteczka etylenu

Najprostszą cząsteczką z wiązaniem podwójnym jest etylen:  . W cząsteczce tej każdy z dwóch atomów węgla tworzy wiązanie z dwoma atomami wodoru i z drugim atomem węgla. Kąty pomiędzy tymi wiązaniami wynoszą około 120° i wszystkie atomy leżą w jednej płaszczyźnie (cząsteczka jest w stanie podstawowym płaska). Oprócz elektronów uczestniczących w wiązaniu σ (a tym samym silnie związanych) na każdym z atomów węgla znajduje się po jednym elektronie zajmującym orbital atomowy p, tak jak to zaznaczono na rys. 36. Orbitale p są prostopadłe do płaszczyzny cząsteczki i oddziałując ze sobą mogą tworzyć orbital molekularny (porównaj opis tworzenia orbitali molekularnych z orbitali atomowych w artykule „Chemia kwantowa" str. 269). Jest to orbital molekularny typu π, a z jego powstawaniem wiąże się zmniejszenie odległości między atomami węgla; odległość ta jest mniejsza niż



Rys. 36. Kształt orbitali w cząsteczce etylenu: a) orbitale atomowe p_z oraz molekularne: b) orbital wiążący π, c) orbital antywiążący π^*

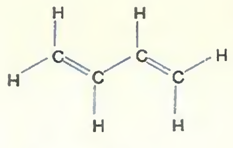
wtedy, gdy atomy węgla związane są ze sobą tylko pojedynczym wiązaniem typu σ. Powstały orbitał molekularny π ma kształt zobrazowany na rys. 36b i jest orbitalem wiążącym, chociaż jego energia wiązania jest mniejsza niż energia wiązania wiązań σ. Orbitalowi wiążącemu towarzyszyć musi jak zawsze orbital antywiążący π^*, a kształt tego orbitala ilustruje rys. 36c. Obydwa elektrony π cząsteczki obsadzają niższy energetycznie orbital wiążący. Najniższe energetycznie przejście elektronowe w cząsteczce etylenu związane jest ze wzbudzeniem elektronu z orbitala wiążącego π do orbitala antywiążącego π^*. Jest to więc przejście typu $\pi \rightarrow \pi^*$.

przejście $\pi \rightarrow \pi^*$

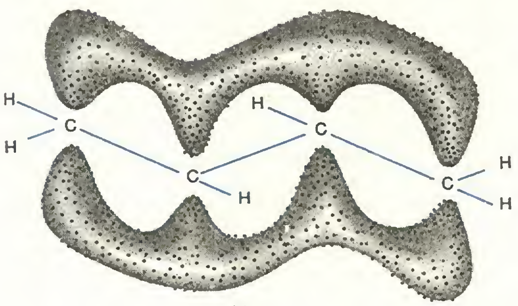
Zwróćmy uwagę na to, że przejście elektronowe $\pi \rightarrow \pi^*$ w etylenie prowadzi do zasadniczej zmiany rozkładu ładunku elektronowego w cząsteczce. W stanie wzbudzonym ładunek zostaje usunięty z obrębu wiązania węgiel-węgiel. Wiązanie ulega zatem osłabieniu i nie ma już charakteru wiązania podwójnego, jakim było w stanie podstawowym. Przejściu elektronowemu $\pi \rightarrow \pi^*$ w cząsteczce etylenu odpowiada pasmo absorpcyjne z maksimum przy ok. 165 nm i jest to absorpcja wywołana ugrupowaniem C=C w cząsteczce.

układ sprzężony

Ciekawie przedstawia się sytuacja w cząsteczkach, w których wiązania podwójne C=C tworzą tzw. układ sprzężony. W układzie takim kolejne wiązania podwójne rozdzielone są wiązaniami pojedynczymi (cząsteczki takie nazywamy polienami). Najprostszym przykładem jest cząsteczka butadienu, pokazana na rys. 37a, w której dwa wiązania podwójne rozdzielone są przez wiązanie pojedyncze. Rozumując w ten sam sposób, jak w przypadku etylenu, dochodzimy



a)



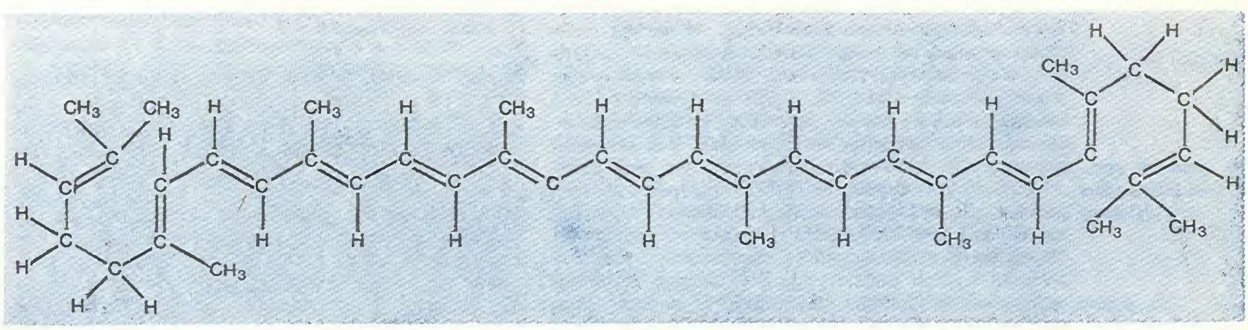
b)

Rys. 37. Cząsteczka butadienu i kształt jednego z jej orbitali wiążących

cząsteczka butadienu

do wniosku, że w cząsteczce butadienu na orbitalach atomowych p (atomów węgla) znajdują się cztery elektrony. Orbitale molekularne typu π, które mogą powstać z kombinacji orbitali atomowych, rozciągają się na całą cząsteczkę, tak jak to np. pokazano na rys. 37b. W rezultacie, elektrony π są zdelokalizowane w całym układzie wiązań sprzężonych. Okazuje się, że konsekwencją delokalizacji elektronów π jest obniżenie energii przejść $\pi \rightarrow \pi^*$, a więc i przesuwanie się widma elektronowego w stronę dłuższych fal. Przesunięcie jest tym większe, im większy jest obszar delokalizacji, tzn. im dłuższy jest układ sprzężony. Na przykład pasmo absorpcyjne $\pi \rightarrow \pi^*$ etylenu ma maksimum przy około 165 nm, a w widmie butadienu maksimum tego pasma leży przy około 217 nm. W miarę jak wydłuża się łańcuch wiązań sprzężonych, widmo przesuwa się coraz dalej w stronę widzialną. W heksatrienie (trzy wiązania podwójne C=C w układzie sprzężonym) maksimum pasma $\pi \rightarrow \pi^*$ leży przy 256 nm, a w dekatetraenie (cztery takie wiązania) przy 310 nm. Wreszcie cząsteczka związku zwanego likopenem, która wygląda tak:

cząsteczka likopenu

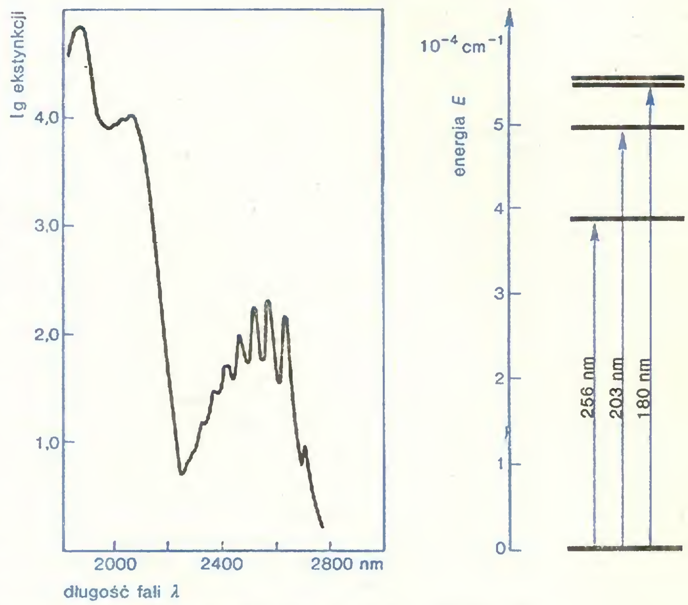


a w której można się doliczyć aż jedenastu wiązań w układzie sprzężonym, ma widmo z maksimum pasma długofalowego przy około 470 nm. Położenie pasma w samym środku obszaru widzialnego sprawia, że to właśnie likopen nadaje czerwone zabarwienie pomidorom i niektórym owocom (il. 18, tabl. 6).

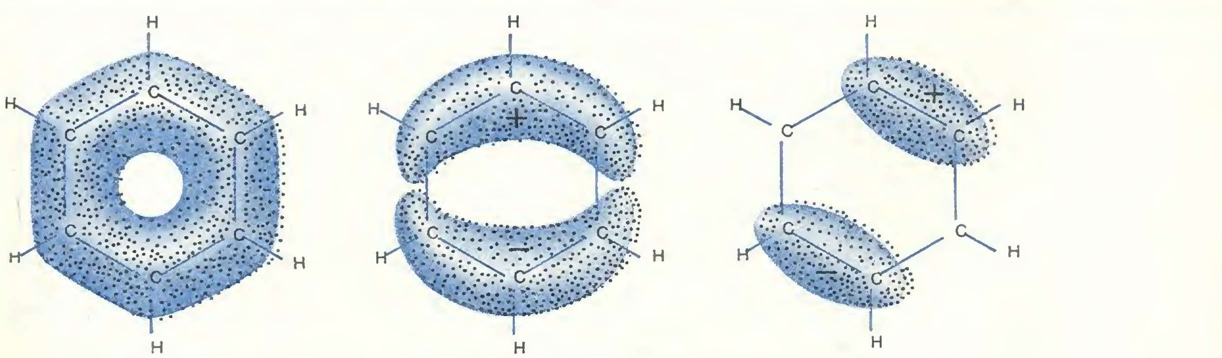
Bardzo ważną w chemii klasą związków, dla których charakterystyczna jest absorpcja wywołana przejściami $\pi \rightarrow \pi^*$, są węglowodory aromatyczne. Widma tych związków poznamy na przykładzie najważniejszego przedstawiciela tej klasy, a mianowicie cząsteczki benzenu.

cząsteczka benzenu

Benzen ma sześć elektronów π i sześć orbitali π — trzy wiążące i trzy antywiążące. Cząsteczkę tę dokładnie omówiono w artykule „Chemia kwantowa" (str. 280). W stanie podstawowym sześć elektronów



Rys. 39. Widmo elektronowe cząsteczki benzenu i schemat jej elektronowych poziomów energetycznych



Rys. 38. Kształt wiążących orbitali molekularnych w cząsteczce benzenu (przekrój w płaszczyźnie cząsteczki)

obsadza trzy wiążące orbitale π, których kształt z grubsza ilustruje rys. 38. Rozważania zmierzające do określenia liczby przejść elektronowych i ich właściwego przypisania są w przypadku cząsteczki benzenu znacznie trudniejsze niż w przypadku etylenu czy butadienu. Nie będziemy tutaj próbowali ich przeprowadzać, powiemy tylko, że rozważenie różnych możliwych konfiguracji elektronowych, z wykorzystaniem elementów symetrii cząsteczki, prowadzi do wniosku, że w benzenie powinny występować trzy przejścia typu $\pi \rightarrow \pi^*$ i widmo absorpcyjne powinno zawierać trzy pasma. Obserwowane widmo benzenu i diagram stanów przewidywany przez teorię pokazuje rys. 39. Widmo benzenu jest bardzo charakterystyczne i cząsteczka ta łatwo może być zidentyfikowana metodami spektroskopowymi. Inne węglowodory aromatyczne mają również charakterystyczne widma, czasem o wysokim stopniu indywidualności.

Zarówno omawiane przez nas wiązanie podwójne C=C w etylenie, jak i pierścień aromatyczny (pierścień taki jak w cząsteczce benzenu) należą do wspomnianych poprzednio grup chromoforowych, tzn. grup posiadających bardzo charakterystyczne widmo absorpcyjne pozwalające na stosunkowo łatwą ich identyfikację nawet wtedy, gdy wchodzą w skład bardziej złożonych cząsteczek. W obu tych grupach charakterystyczna absorpcja związana jest z przejściami $\pi \rightarrow \pi^*$. Jednak nie zawsze charakterystyczna absorpcja chromoforów musi być związana ze wzbudzeniem elektronów π. Na przykład bardzo często spotykaną w różnych związkach grupą chromoforową jest grupa karbonylowa >C=O, której charakterystyczna absorpcja wiąże się z przejściem elektronowym typu $n \rightarrow \pi^*$, tzn. ze wzbudzeniem niewiążących elektronów wolnej pary elektronowej, zlokalizowanych przy atomie tlenu O. Charakterystyczne pasmo absorpcji $n \rightarrow \pi^*$ grupy karbonylowej ma maksimum przy ok. 280 nm.

przejście $n \rightarrow \pi^*$

Charakterystyczna absorpcja, umożliwiająca łatwą identyfikację grup chromoforowych i stwierdzenie ich obecności w cząsteczce, może ulegać zmianom, czasem nawet bardzo znacznym, jeżeli grupa chromoforowa silnie oddziałuje z innymi częściami cząsteczki lub z inną grupą chromoforową obecną w cząsteczce.

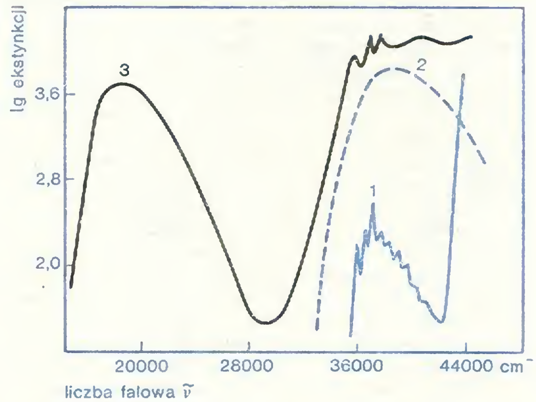
Widma cząsteczek, które dotychczas omówiliśmy, w celu uświadomienia sobie faktu, że widma elektronowe wiążą się z określonymi zmianami w układzie elektronowym cząsteczki, tylko w niewielkim stopniu ukazały niezwykłą złożoność i różnorodność sytuacji, z jakimi można się zetknąć badając elektronowe widma absorpcyjne cząsteczek. Na przykład poważne zmiany w widmie absorpcyjnym grup chromoforowych i cząsteczek może spowodować rozpuszczalnik, w którym znajduje się badana cząsteczka. Są one wynikiem oddziaływania rozpuszczonej cząsteczki z otaczającymi ją cząsteczkami rozpuszczalnika i jeżeli oddziaływania te ulegają zmianom w wyniku wzbudzenia cząsteczki, to widmo ulega przesunięciu. Widmo może przesuwać się zarówno w stronę dłuższych, jak i w stronę krótszych fal. Zależy to od polarności rozpuszczalnika i cząsteczki rozpuszczonej i od tego, jak zmienia się moment dipolowy cząsteczki po wzbudzeniu — rośnie czy też maleje. Są teorie, na podstawie których można przewidywać kierunki przesunięć widm związanych z różnego typu przejściami elektronowymi w cząsteczce. Obserwacja przesunięć widma cząsteczki w różnych rozpuszczalnikach dostarczyć może informacji o charakterze oddziaływań międzycząsteczkowych i o tym, jak zmieniają się w wyniku wzbudzenia różne wielkości charakteryzujące stany elektronowe cząsteczki, np. moment dipolowy.

wpływ rozpuszczalnika na widma

Oddziaływania międzycząsteczkowe wywołują zmiany widma, o czym była już mowa przy okazji rozważań nad widmami w podczerwieni kompleksów z wiązaniem wodorowym. I odwrotnie, badając zmiany widm możemy stwierdzać obecność takich oddziaływań i próbować ustalić ich mechanizm, tzn. ustalić co z czym oddziałuje (np. czy cała cząsteczka, czy też pewien jej fragment) i jakiego rodzaju siły są zaangażowane w oddziaływaniu.

Zajmiemy się teraz pewnym przykładem, który zilustruje nam, jak można dojść do ustalenia charakteru oddziaływania międzycząsteczkowego, badając widma. Od bardzo dawna chemicy znali zjawisko polegające na tym, że jeżeli cząsteczki pewnych różnych związków zmiesza się razem, to mieszanina taka może nabierać zabarwienia niebieskiego, zielonego, czerwonego itp., chociaż roztwory obu substancji przed zmieszaniem były zupełnie bezbarwne. Wiemy jakie są związki pomiędzy barwą roztworu, a jego widmem. Jeśli wyjściowe roztwory były bezbarwne to znaczy, że widma rozpuszczonych cząsteczek leżą w nadfiolecie. Jeżeli po zmieszaniu powstała mieszanina uzyskała pewną barwę to znaczy, że absorbuje ona teraz w obszarze widzialnym. Moglibyśmy przypuszczać, że w wyniku zmieszania widma cząsteczek rozpuszczonych przesunęły się po prostu w stronę dłuższych fal. Jednakże zbadanie widma mieszaniny

widmo a oddziaływania międzycząsteczkowe



Rys. 40. Widma absorpcyjne układu: czterocyjanoetylen (akceptor) i sześciometylobenzen (donor) w roztworze (temperatura −180°C): *1* widmo absorpcyjne sześciometylobenzenu, *2* widmo absorpcyjne czterocyjanoetylenu, *3* widmo absorpcyjne roztworu czterocyjanoetylen + sześciometylobenzen. Pasmo mające maksimum przy ok. 19 000 cm^{-1} jest pasmem absorpcyjnym kompleksu charge-transfer

zaprzecza temu. Widma cząsteczek przed zmieszaniem i po zmieszaniu są prawie niezmienione. Sytuację tę ilustruje rys. 40. Ale w mieszaninie pojawia się nowe pasmo absorpcyjne, które tak, jak się tego spodziewaliśmy, leży w obszarze widzialnym, daleko od widm składników mieszaniny. Pasmo to jest charakterystyczne dla zmieszanego układu, tzn. dla obu rodzajów cząsteczek występujących w mieszaninie, np. zamiana jednej z nich na inną spowoduje pojawienie się pasma w całkiem innym obszarze. Mówimy, że pasmo to jest pasmem absorpcyjnym charakterystycznym dla oddziałującego układu. Dokładne badania pokazały, że pewna liczba cząsteczek obu związków tworzy w mieszaninie układy, w których obie cząsteczki znajdują się bardzo blisko siebie, tworząc układ luźno związany. Wzbudzenie elektronowe w takim układzie wiąże się ze wzbudzeniem elektronu jednej z cząsteczek, ale elektron ten nie przechodzi do wyższego stanu wzbudzonego tej cząsteczki, lecz do wyższego stanu wzbudzonego drugiej cząsteczki (mówiąc inaczej jest przenoszony z orbitala molekularnego pierwszej cząsteczki na nieobsadzony orbital molekularny zlokalizowany na drugiej cząsteczce). Tak więc proces wzbudzenia prowadzi do przeniesienia ładunku elektronu z jednej cząsteczki (zwanej donorem elektronu) do drugiej cząsteczki (zwanej akceptorem elektronu). Takie oddziałujące układy nazywa się w chemii kompleksami z przeniesieniem ładunku (lub czasem kompleksami *charge-transfer*,

kompleksy z przeniesieniem ładunku

pasma absorpcyjne charge-transfer

od angielskiej nazwy procesu przeniesienia ładunku). Pasma absorpcyjne charakterystyczne dla tych kompleksów nazywamy pasmami absorpcyjnymi *charge-transfer*. Proces przenoszenia ładunku jest procesem dosyć powszechnie spotykanym wśród układów molekularnych, a kompleksy z przeniesieniem ładunku odgrywają zapewne bardzo poważną rolę m.in. w procesach biologicznych. Właśnie badanie widm tych układów pozwoliło zrozumieć naturę zjawiska i własności takich kompleksów.

Możliwość poznawania oddziaływań międzycząsteczkowych na podstawie badania widm sprawia, że spektroskopia molekularna jest potężnym instrumentem poznawania otaczającego nas świata, gdzie większość zjawisk, z którymi stale się spotykamy, jest właśnie wynikiem oddziaływań między różnymi elementami, z których zbudowani jesteśmy my sami i to co nas otacza.

Stany wzbudzone cząsteczek

Mówiliśmy do tej pory o widmach absorpcyjnych, a więc widmach wywołanych przejściami elektronowymi ze stanu podstawowego do stanu wzbudzonego. Zastanawialiśmy się, jakie zmiany w strukturze elektronowej pociąga za sobą przejście, jak zmienia się rozkład gęstości ładunku elektronowego w wyniku przejścia i jak zmiany takie można odczytać z obserwacji widma. Nie obchodziły nas dotychczas dalsze losy cząsteczki w stanie wzbudzonym. Będzie to tematem naszych dalszych rozważań.

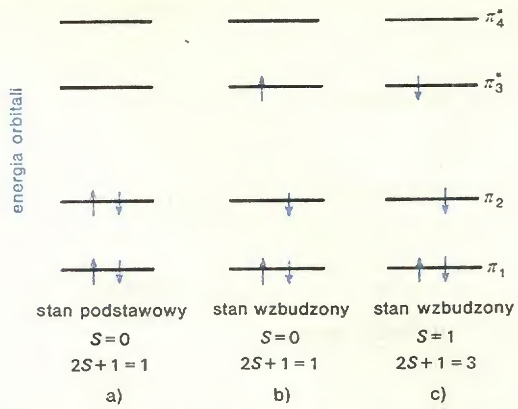
Ze wzbudzeniem elektronowym związana jest stosunkowo duża energia, np. absorpcja promieniowania z zakresu 300–400 nm — wzbogaca cząsteczkę o ok. 400 do 300 kJ/mol lub ok. 4 do 3 eV. Elektronowe stany wzbudzone są więc stanami o znacznym, w stosunku do stanu podstawowego, nadmiarze energii i należy się spodziewać, że cząsteczka nie będzie w takim stanie przebywała zbyt długo. Istotnie doświadczenie pokazuje, że średni czas w jakim cząsteczka przebywa w stanie wzbudzonym — nazywany czasem życia stanu wzbudzonego — jest bardzo krótki, zazwyczaj rzędu 10^{-7}–10^{-9} s.

przejście promieniste — emisja

Cząsteczka wzbudzona może w różny sposób pozbyć się nadmiaru energii i powrócić do stanu podstawowego. Skoncentrujemy się na sposobie powrotu cząsteczki do stanu podstawowego, z którym wiąże się wysłanie nadmiaru energii w postaci promieniowania elektromagnetycznego (uprzednio pochłoniętego). Taki powrót nazywa się przejściem promienistym (lub dezaktywacją promienistą), a jego rezultatem jest emisja promieniowania zwanego najogólniej luminescencją. Końcowym efektem takiego procesu, którego badaniem może zająć się spektroskopia, jest widmo emisyjne — widmo luminescencji. Zanim jednak zajmiemy się nieco szczegółowiej widmami emisyjnymi i tym co może przynieść ich badanie, musimy uzupełnić nasze wiadomości o naturze stanów wzbudzonych i procesach promienistych, a także bezpromienistych.

luminescencja

Do badań stanów wzbudzonych cząsteczki i procesów dezaktywacji (sposobów i dróg jakimi cząsteczka pozbywa się energii wzbudzenia) tych stanów potrzebna jest znajomość schematu stanów energetycznych cząsteczki. Ustalenie takiego schematu jest zresztą jednym z pierwszych zadań i celów spektroskopii. Schemat taki, szczególnie gdy chodzi o wieloatomowe cząsteczki, może być bardzo złożony i na ogół nigdy nie jest w pełni znany, ponieważ nasza wiedza w tym przedmiocie ogranicza się do kilku najniżej położonych (o najniższej energii) stanów wzbudzonych. Już z artykułu „Chemia kwantowa" wiemy, że dodatkową komplikacją schematu energetycznego cząsteczki jest występowanie układów stanów o różnych multipletowościach. Przypomnijmy to sobie raz jeszcze na prostym przykładzie zilustrowanym rys. 41. Na rysunku tym zaznaczono schematycznie cztery orbitale molekularne (w skali energii) hipotetycznej cząsteczki — dwa o niższej energii, obsadzone przez cztery elektrony, i dwa nie zapełnione (tak



Rys. 41. Schemat energetyczny orbitali molekularnych i możliwe konfiguracje elektronowe w stanie podstawowym i wzbudzonym

multipletowość stanów

właśnie mógłby wyglądać schemat orbitali cząsteczki butadienu). Strzałki na rysunku symbolizują orientację spinów elektronów. Zgodnie z zasadą Pauliego w stanie podstawowym oba elektrony muszą mieć przeciwne spiny, co zaznaczono przeciwnymi zwrotami strzałek (rys. 41a) i wobec tego wypadkowy spin $S = 0$, a multipletowość stanu jest $2S+1 = 1$ i stan podstawowy jest stanem singletowym.

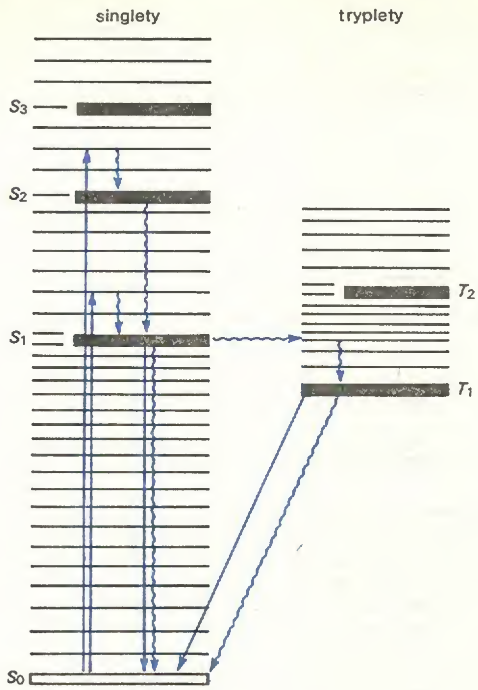
Wzbudzenie elektronowe jest równoważne przejściu jednego z elektronów na wyższy nie obsadzony orbital. Mamy teraz do czynienia z inną konfiguracją elektronową; jest to konfiguracja stanu wzbudzonego. Teraz jednak zakaz Pauliego już nie obowiązuje, gdyż oba elektrony znajdują się na różnych przestrzennie orbitalach i wobec tego możliwe są dwie sytuacje. Albo elektrony będą miały spiny przeciwnie (rys. 41b), albo też jednakowo skierowane (rys. 41c). Przy pierwszej ewentualności wypadkowy spin będzie, tak jak w stanie podstawowym, równy zeru. Natomiast druga ewentualność prowadzi do wypadkowego spinu $S = 1$ i multipletowości stanu $2S+1 = 3$. Tej samej więc konfiguracji elektronowej odpowiadają dwa stany o różnych multipletowościach — jeden jest singletem, drugi natomiast trypletem. Z taką sytuacją mamy do czynienia w cząsteczkach. Z naszego rysunku wynika, że oba stany mają tę samą energię, jednak naprawdę ich energia nie jest jednakowa. To, że elektrony w stanie trypletowym znajdują się na różnych przestrzennie orbitalach sprawia, iż energia tego stanu jest z reguły niższa niż stanu singletowego tej samej konfiguracji elektronowej (dokładnie wyjaśniono ten efekt w artykule „Chemia kwantowa".

schemat Jabłońskiego

Spróbujmy teraz podać ogólny schemat energetyczny cząsteczki i zobrazować na nim drogi powrotu ze stanów wzbudzonych do stanu podstawowego. Taki schemat nosi w spektroskopii nazwę diagramu (lub schematu) Jabłońskiego (od nazwiska wybitnego współczesnego fizyka polskiego A. Jabłońskiego, który pierwszy taki schemat zastosował do wyjaśnienia różnych obserwowanych widm emisyjnych).

Na diagramie (rys. 42) zaznaczono kilka pierwszych stanów z sekwencji stanów singletowych S_1, S_2, S_3 oraz dwa pierwsze stany T_1, T_2 z sekwencji stanów trypletowych. Dla każdego z tych stanów zaznaczono również kilka pierwszych stanów oscylacyjnych. S_0 jest stanem podstawowym cząsteczki. Absorpcja promieniowania o odpowiedniej długości fali prowadzi ze stanu S_0 do stanów wzbudzonych S_1, S_2 lub wyższych. Absorpcja prowadząca ze stanu S_0 do stanów T_1, T_2, ... jest, o czym już wiemy, procesem wzbronionym, gdyż wiąże się ona ze zmianą spinu. Zakaz ten jak wiemy, nie jest zakazem ścisłym i w pe-

wnych warunkach absorpcja taka może być obserwowana. Jednak w normalnych warunkach nie można jej na ogół zaobserwować.



Rys. 42. Diagram Jabłońskiego stanów energetycznych cząsteczki. Linie ciągłe łączące różne stany odpowiadają procesom promienistym, linie faliste – procesom bezpromienistym

absorpcja i emisja na diagramie Jabłońskiego

Absorpcja promieniowania może prowadzić do każdego ze stanów wzbudzonych S_1, S_2, S_3, ..., S_n, ale emisja w praktyce zawsze bierze swój początek z najniższego stanu wzbudzonego S_1, a do tego również z najniższego, zerowego stanu oscylacyjnego ($v = 0$) tego stanu elektronowego. Dlaczego właściwie nie obserwujemy fluorescencji z wyższych stanów wzbudzonych? Nasz diagram energetyczny jest wprawdzie przejrzysty, ale daleko odbiega od rzeczywistej sytuacji w wieloatomowej cząsteczce. Wiemy, że ze wzbudzeniem elektronowym związane jest również wzbudzenie oscylacji. Przypuśćmy, że wzbudziliśmy naszą cząsteczkę, tę, której odpowiada diagram na rys. 42, do drugiego wzbudzonego stanu singletowego S_2 i że to wzbudzenie sięga do poziomu oscylacyjnego np. $v = 3$. Nasz diagram tego nie pokazuje, bo cały rysunek stałby się nieczytelny, ale gdybyśmy chcieli uzupełnić go dalszymi poziomami oscylacyjnymi stanu S_1, to okazałoby się, że w okolicy poziomu oscylacyjnego $v = 3$ w stanie S_2 znalazłoby się wiele poziomów oscylacyjnych stanu S_1 (poziomów o bardzo wysokich liczbach oscylacyjnych). Byłoby ich tym więcej im większa jest cząsteczka (pamiętamy, że liczba drgań normalnych rośnie z liczbą atomów w cząsteczce). Gęstość stanów oscylacyjnych stanu S_1 (ich liczba przypadająca na jednostkę odstępu energetycznego, np. na 1 cm^{-1}) w okolicy poziomu $v = 3$, stanu S_2, jest bardzo duża — czasem tak duża, że odstępy między poszczególnymi poziomami oscylacyjnymi są nierozróżnialne i poziomy te tworzą pewien kwaziciągły obszar energii. Ta duża gęstość stanów oscylacyjnych niższego stanu elektronowego w obszarach energii odpowiadających wyższym stanom elektronowym stanowi właśnie w cząsteczce „kanał", przez który energia wysokiego wzbudzenia elektronowego (w naszym przykładzie wzbudzenia do S_2 i $v = 3$) może zostać odprowadzona z wyższeg owzbudzonego stanu elektronowego do niższego stanu wzbudzonego (S_1) i w ten sposób „rozdzielona" miedzy różne oscylacje tego stanu. W ten sposób energia wzbudzenia elektronowego w wyższym stanie zostaje szybko zamieniona na energię oscylacji w niższym stanie.

W podobny sposób przebiegać może powrót cząsteczki nie tylko ze stanu S_2 do S_1, ale również ze stanów wyższych S_n do S_1 (przy pośrednictwie stanów znajdujących się pomiędzy S_n i S_1). Ogólnie rzecz biorąc, każde wzbudzenie do wyższych stanów wzbudzonych kończy się w efekcie tym, że cząsteczka bardzo szybko wraca do najniższego poziomu oscylacyjnego pierwszego wzbudzonego stanu elektronowego S_1. Doświadczenie pokazuje, że prawdopodobieństwo takiego powrotu z wyższych stanów wzbudzonych do pierwszego stanu wzbudzonego jest bardzo duże. Dodajmy jeszcze do tego, że również bardzo prawdopodobnym procesem jest strata części energii wzbudzenia podczas zderzeń wzbudzonej cząsteczki z innymi cząsteczkami w ośrodku. Nie potrafimy jednak jeszcze określić, jaka część energii wzbudzenia tracona jest na tej drodze.

Trochę inaczej wygląda sprawa przejścia cząsteczki ze stanu S_1 do stanu podstawowego S_0. Oczywiście i w tym wypadku gęstość stanów oscylacyjnych stanu S_0 w obszarze zerowego stanu oscylacyjnego stanu S_1 jest ogromna i moglibyśmy spodziewać się również, że opisany poprzednio mechanizm będzie i teraz wydajny. Rzecz w tym jednak, że obraz, który podaliśmy, jest zbyt uproszczony. Gęstość stanów oscylacyjnych, to tylko jeden z czynników warunkujących omawiany mechanizm. Innym ważnym czynnikiem, o którym nie wspominaliśmy do tej pory, jest odstęp energii między poszczególnymi stanami elektronowymi. Otóż jeśli odstęp ten jest duży, to nawet wielka gęstość stanów oscylacyjnych nie wystarcza, aby zapewnić wydajną zamianę energii wzbudzenia elektronowego (do jednego z wyższych stanów) na energię oscylacji w stanie niższym. W cząsteczkach sytuacja jest na ogół taka, że odstęp energetyczny między stanem podstawowym S_0 a pierwszym wzbudzonym stanem singletowym S_1 jest znacznie większy niż pomiędzy S_1 i S_2 i kolejnymi wyższymi stanami wzbudzonymi. To właśnie sprawia, że omówiony mechanizm jest znacznie mniej wydajny i ważny dla przejścia $S_1 \rightsquigarrow S_0$ niż dla przejść $S_2 \rightsquigarrow S_1$, $S_3 \rightsquigarrow S_2$, itp. Proces powrotu cząsteczki z wyższych stanów singletowych do niższych jest procesem bardzo szybkim i zachodzi w czasie, który średnio zawarty jest w przedziale 10^{-11}–10^{-14} s. W porównaniu z nim proces $S_1 \rightsquigarrow S_0$ jest procesem bardzo wolnym, gdyż charakterystyczny dla niego czas jest rzędu 10^{-5}–10^{-7} s.

przejście bezpromieniste

Przejścia $S_n \rightsquigarrow S_{n-1}$, ..., $S_2 \rightsquigarrow S_1$, $S_1 \rightsquigarrow S_0$ są przejściami, w których energia wzbudzenia tracona przez cząsteczkę nie jest zamieniana na energię kwantów promieniowania elektromagnetycznego. Takie przejścia noszą ogólną nazwę przejść bezpromienistych (oznacza się je czasem i my będziemy się takiego oznaczenia trzymali, liniami falistymi). Przejścia bezpromieniste zachodzące między stanami o tej samej multipletowości, jak w omawianym przez nas wypadku, nazywają się konwersją wewnętrzną.

konwersja wewnętrzna

Jakie mogą być dalsze losy cząsteczki, która po wzbudzeniu znalazła się, w wyniku szybkich przejść bezpromienistych, w stanie wzbudzonym S_1? Wiemy już, że może ona powrócić bezpromieniście do stanu podstawowego S_0, chociaż proces takiego powrotu nie należy do zbyt wydajnych. Drugą możliwością, którą teraz szerzej się zajmiemy, jest powrót promienisty. Jego rezultatem jest emisja promieniowania i takie właśnie procesy prowadzą do obserwacji widm emisyjnych. Emisja ta nazywa się fluorescencją. Staranne badania wykazały, że charakterystyczny czas przejść promienistych, w większości badanych cząsteczek, zawiera się w granicach 10^{-7}–10^{-9} s. A zatem fluorescencja może z powodzeniem konkurować z wewnętrzną konwersją $S_1 \rightsquigarrow S_0$ w cząsteczce. Staje się jednocześnie zrozumiałym fakt, dlaczego w praktyce obserwowana fluorescencja bierze swój początek zawsze ze stanu S_1.

fluorescencja

przejścia międzysystemowe

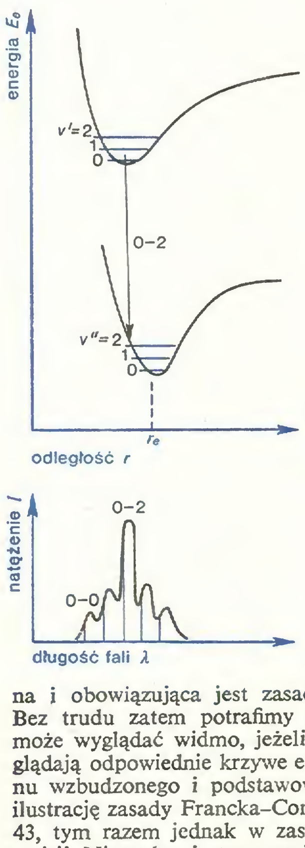
Wewnętrzna konwersja $S_1 \rightsquigarrow S_0$ i emisja $S_1 \to S_0$ nie są jedynymi możliwymi procesami, w wyniku których opróżniany jest wzbudzony stan singletowy S_1. Na rys. 42 zaznaczony jest jeszcze jeden proces polegający na bezpromienistym przejściu ze stanu S_1 do stanu trypletowego T_1. Przejście takie, zwane przejściem międzysystemowym, przeprowadza cząsteczkę do stanu o innej multipletowości niż stan wyjściowy i chociaż wiemy, że przejścia ze zmianą multipletowości są na ogół bardzo mało prawdopodobne, to wbrew naszym przypuszczeniom przejścia takie są w cząsteczkach zjawiskiem dosyć powszechnym. Czynnikiem umożliwiającym występowanie przejść międzysystemowych jest obecność tzw. sprzężenia spin-orbita. Jest ono rezultatem oddziaływania pomiędzy momentami magnetycznymi, orbitalnymi i spinowymi elektronów w cząsteczce.

Zanim zajmiemy się bardziej szczegółowym omówieniem widm fluorescencji cząsteczek, zapoznamy się pokrótce z najważniejszymi wielkościami charakteryzującymi emisję i widma emisyjne cząsteczek. Wielkości te są zazwyczaj wyznaczane w pomiarach luminescencyjnych i stanowią podstawową informację o widmie fluorescencyjnym, a tym samym o stanie wzbudzonym, z którego bierze początek fluorescencja. Pozwalają one również na ocenę prawdopodobieństwa zachodzenia innych (na ogół bezpromienistych) procesów w tych stanach. Najważniejszymi wielkościami charakteryzującymi fluorescencję są: energia, czas zaniku i wydajność kwantowa.

widmo fluorescencji

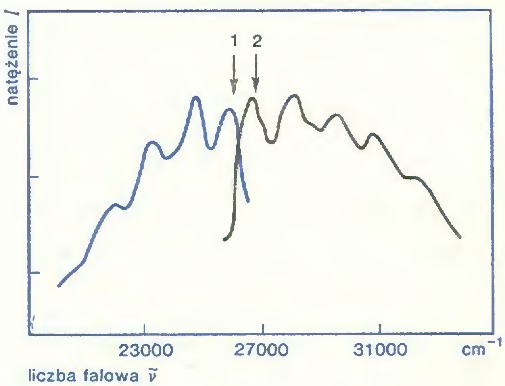
Spośród tych trzech wielkości energia jest tą, którą wyznacza się najłatwiej, bo z bezpośredniej obserwacji widma fluorescencji. Podobnie jak i przy przejściach absorpcyjnych, również i dla przejść emisyjnych słuszna i obowiązująca jest zasada Francka–Condona. Bez trudu zatem potrafimy sobie wyobrazić, jak może wyglądać widmo, jeżeli tylko wiemy jak wyglądają odpowiednie krzywe energii potencjalnej stanu wzbudzonego i podstawowego. Znaną już nam ilustrację zasady Francka–Condona przedstawia rys. 43, tym razem jednak w zastosowaniu do procesu emisji. Nietrudno jest zrozumieć, że sytuacja taka jak na rys. 43 prowadzi do widma fluorescencji, w którym największe natężenie winno mieć pasmo oscylacyjne 0–2, a rozkład natężeń w obrębie całego pasma elektronowego taki jak pokazany u dołu rysunku. Zwróćmy uwagę na fakt, że struktura oscylacyjna widma fluorescencji odpowiada oscylacjom w stanie podstawowym, gdyż przejścia ze stanu wzbudzonego prowadzą do różnych stanów oscylacyjnych stanu podstawowego. Zwróćmy również uwagę na to, że przejście emisyjne 0–0 jest przejściem o najwyższej energii w widmie fluorescencji, a więc przejściem najbardziej krótkofalowym. Z drugiej strony przejście 0–0 w widmie absorpcyjnym jest przejściem najniższej energii (porównaj rys. 32, 33, 34), tzn. najbardziej długofalowym w widmie absorpcyjnym. W obu wypadkach energia przejścia 0–0 jest przynajmniej teoretycznie jednakowa i wobec tego jest to jedyne przejście wspólne dla obu widm. A zatem możemy oczekiwać, że widmo fluorescencji będzie zawsze leżało w obszarze bardziej długofalowym niż widmo absorpcji (to spostrzeżenie nazywa się czasem regułą Stokesa, a przesunięcie między widmem absorpcji i fluorescencji — przesunięciem Stokesa lub stokesowskim). Możemy również spodziewać się, że między widmem absorpcji i fluorescencji występować będzie zwierciadlana symetria, szczególnie wyraźna wtedy, gdy oscylacje cząsteczki w stanie podstawowym i w stanie wzbudzonym nie będą się różniły. Długość fali wspólna dla obu widm (albo inaczej punkt przecięcia widm absorpcji i emisji) wyznacza w takich wypadkach energię przejścia 0–0, a więc i odstęp energetyczny między stanem podstawowym i wzbudzonym.

reguła Stokesa



Rys. 43. Ilustracja zasady Francka–Condona i rozkład natężeń w widmie emisyjnym

Oczywiście nasze oczekiwania, jak zresztą zawsze, dotyczą przypadków idealnych, z którymi w praktyce prawie nigdy się nie stykamy. Wspomniana symetria zwierciadlana jest tylko symetrią przybliżoną, a nigdy dokładną, co widać na rys. 44. Podobnie w widmie cząsteczki znajdującej się w roztworze przejścia 0–0 w absorpcji i emisji mają z reguły inną energię.



Rys. 44. Widmo absorpcji i fluorescencji cząsteczki antracenu. Strzałki pokazują położenie przejścia 0–0 w emisji (*1*) i absorpcji (*2*)

promienisty czas życia

Zajmiemy się teraz czasem życia stanu wzbudzonego i czasem zaniku fluorescencji. Jeżeli cząsteczka znajduje się w stanie wzbudzonym, a jedyną drogą jej powrotu do stanu podstawowego jest przejście promieniste (nieobecne lub nieaktywne są wszelkie procesy bezpromienistej dezaktywacji), to szybkość, z jaką zachodzi proces promienisty, określa znany już nam współczynnik Einsteina emisji spontanicznej A_{ji} (przejście $j \to i$). Mówiąc inaczej współczynnik ten jest stałą szybkości zaniku stanu wzbudzonego na drodze promienistej, gdy nieobecne są wszystkie inne konkurencyjne procesy dezaktywacyjne (bezpromieniste). Odwrotność stałej szybkości jest naturalnym czasem życia τ_0 stanu wzbudzonego ze względu na emisję, albo jak często mówimy — promienistym czasem życia; $\tau_0 = 1/A_{ji}$. Wiemy, że A_{ji} jest proporcjonalne do B_{ji} — współczynnika absorpcji. Wobec tego bez trudu domyślimy się, iż możliwe jest ustalenie związków między naturalnym czasem życia stanu wzbudzonego i wielkościami charakteryzującymi przejście absorpcyjne do tego stanu. Mamy wszelkie podstawy, aby spodziewać się, że własności stanu wzbudzonego warunkujące absorpcję i emisję

nie są wielkościami od siebie niezależnymi, lecz że są współzależne i że znając jedne potrafimy określić czy przewidzieć drugie. Tak jest w istocie. Można podać szereg związków łączących naturalny czas życia z wyznaczonymi doświadczalnie wielkościami charakteryzującymi widma absorpcyjne. Na przykład jednym z takich związków jest poniższa zależność:

$$\tau_0 \approx 3{,}5 \cdot 10^8 (\overline{\nu}^2 \cdot \Delta\nu_{1/2} E_{max})^{-1};$$

tutaj $\overline{\nu}^2$ jest średnią częstością pasma absorpcyjnego, $\Delta\nu_{1/2}$ — szerokością połówkową pasma, E_{max} — współczynnikiem ekstynkcji w maksimum pasma. Korzystając z tej zależności spróbujemy oszacować, jakiego rzędu jest naturalny czas życia wzbudzonego stanu singletowego typowej cząsteczki. W przypadku cząsteczek organicznych (np. węglowodorów aromatycznych) ekstynkcja pasm dozwolonych, leżących w obszarze około 30 000 cm^{-1}, jest rzędu ok. 10^4–10^5, co przy typowych szerokościach połówkowych rzędu kilku tysięcy cm^{-1} prowadzi do wniosku, że naturalny czas życia jest rzędu 10^{-8}–10^{-9} s.

Podana wyżej zależność wykazuje, że im bardziej jest prawdopodobne jakieś absorpcyjne przejście elektronowe (tzn. im większe jest jego natężenie), tym krótszy jest naturalny czas życia stanu wzbudzonego osiąganego w wyniku takiego przejścia. Oznacza to z kolei, że tym większa jest możliwość, że przejście odwrotne ze stanu wzbudzonego do podstawowego będzie przejściem promienistym, a więc odbywającym się z emisją promieniowania.

Mówiliśmy w swoim czasie o przejściach wzbronionych, wskazując na to, iż są one mało prawdopodobne, a więc mają bardzo niskie natężenie w porównaniu z natężeniem przejść dozwolonych. Podaliśmy liczbę 10^6 jako czynnik wskazujący, ile razy średnio niższe natężenie mają przejścia zakazane ze względu na zmianę multipletowości, a więc np. przejścia $S_0 \rightarrow T_1$. Istotnie przejścia absorpcyjne $S_0 \rightarrow T_1$, które w pewnych szczególnych warunkach można czasem obserwować, mają bardzo małe natężenie — współczynniki ekstynkcji są rzędu 10^{-2}–10^{-3}. Nietrudno jest wobec tego ustalić, że promienisty czas życia stanu trypletowego T_1, musi być rzędu 10^{-3} s lub dłuższy. Stany o tak długim promienistym czasie życia nazywane są stanami metatrwałymi. Zrozumiałe jest też, że emisja z tych stanów jest w normalnych warunkach bardzo mało prawdopodobna, a o konsekwencjach takiego stanu rzeczy będziemy jeszcze mówić.

stany metatrwałe

Podkreślmy raz jeszcze, że promieniste czasy życia określają średni czas życia stanu wzbudzonego tylko i wyłącznie ze względu na promienisty powrót cząsteczki z tych stanów. Promienisty czas życia może być traktowany jako średni czas życia cząsteczki w danym stanie wzbudzonym tylko wtedy, gdy niemożliwe są żadne inne drogi powrotu cząsteczki do stanu podstawowego. Obserwowane wtedy natężenie fluorescencji, po przerwaniu wzbudzenia, maleje zgodnie z prawem wykładniczego zaniku: $I = I_0 e^{-t/\tau_0}$, gdzie I_0 jest początkowym natężeniem fluorescencji, I — natężeniem po czasie t, a τ_0 — promienistym (naturalnym) czasem życia stanu. Po czasie $t = \tau_0$ natężenie wynosi $I = I_0/e$, czyli zmniejsza się e razy. Taki czas, po którym natężenie fluorescencji zmniejsza się e razy, nazywa się czasem zaniku fluorescencji. Wtedy gdy emisja jest jedyną drogą dezaktywacji stanu wzbudzonego, czas zaniku fluorescencji jest równy promienistemu czasowi życia stanu.

czas zaniku fluorescencji

Obserwowany czas zaniku fluorescencji jest zawsze inny niż czas promienisty, gdyż całkowita szybkość opróżniania stanu wzbudzonego jest sumą szybkości wszystkich procesów, zarówno promienistych jak i bezpromienistych, i wobec tego czas zaniku fluorescencji $\tau = 1/(k_e + k_i)$. Przez k_e oznaczyliśmy tutaj stałą szybkości przejścia promienistego, a przez k_i sumaryczną stałą szybkości wszystkich procesów bezpromienistych (konwersji wewnętrznej, bezpromienistego przejścia międzysystemowego itp.). Ponieważ naturalny czas życia jest równy $\tau_0 = 1/k_e$, wobec tego czas zaniku fluorescencji jest zawsze krótszy niż naturalny, promienisty czas życia. Procesy bezpromieniste skracają czasy zaniku fluorescencji, a zatem badanie czasów zaniku fluorescencji w funkcji różnych czynników zewnętrznych lub wewnętrznych może być źródłem informacji o procesach dezaktywacyjnych.

Poświęćmy teraz trochę uwagi trzeciej wielkości charakteryzującej fluorescencję. Jest nią wydajność kwantowa fluorescencji, którą definiuje się jako:

wydajność kwantowa fluorescencji

$$\text{wydajność kwantowa fluorescencji} = \frac{\text{liczba kwantów wyemitowanych}}{\text{liczba kwantów zaabsorbowanych}}.$$

Jeżeli jedyną możliwą drogą opróżniania stanu wzbudzonego byłyby przejścia promieniste, to oczywiście każdemu kwantowi zaabsorbowanemu w układzie odpowiadałby kwant wyemitowany i wydajność kwantowa byłaby równa 1. Ponieważ jednak w praktyce z taką sytuacją nigdy nie mamy do czynienia, wobec tego i wydajność kwantowa fluorescencji jest zawsze mniejsza od jedności. Można pokazać, że wydajność kwantowa

$$\Phi = \frac{k_e}{k_e + k_i},$$

gdzie k_e, tak jak poprzednio, jest stałą szybkości przejścia promienistego, a k_i — sumaryczną stałą szybkości wszystkich procesów bezpromienistych.

Możemy teraz ustalić związki pomiędzy wydajnością kwantową, a czasem zaniku fluorescencji τ i promienistym czasem życia. Jeśli wykorzystamy podane poprzednio zależności, to znajdziemy, że

$$\Phi = k_e \tau \quad \text{lub} \quad \Phi = \tau/\tau_0.$$

Zauważyliśmy już na pewno, że pomiar czasu zaniku i wydajności kwantowej fluorescencji pozwala na wyznaczenie stałych szybkości k_e i k_i. Niestety, ani pomiary czasu zaniku, ani pomiary wydajności kwantowych nie są doświadczalnie łatwe i wymagają stosowania specjalnych, czasem bardzo precyzyjnych technik pomiarowych. Dlatego też, chociaż obserwacja i badanie widm fluorescencyjnych jest na ogół zadaniem niezbyt skomplikowanym (ale niekoniecznie łatwym), to uzyskanie pełnego obrazu mechanizmów i procesów zachodzących w stanach wzbudzonych cząsteczek jest zadaniem bardzo trudnym i nie zawsze do końca wykonalnym.

Jak wynika z rys. 42, stan S_1 może dezaktywować się nie tylko na skutek przejścia promienistego (fluorescencji) oraz wewnętrznej konwersji, ale również w wyniku przejścia bezpromienistego do stanu trypletowego T_1. Wspominaliśmy już, że z przejściem tym, zwanym przejściem międzysystemowym, związana jest zmiana multipletowości (spinu). I chociaż przejścia ze zmianą multipletowości są przejściami o bardzo małym prawdopodobieństwie, to jednak w wyniku sprzężenia spin-orbita są one silnie wzmacniane. Okazuje się, że w wieloatomowych cząsteczkach przejścia $S_1 \rightsquigarrow T_1$ mogą nawet konkurować z fluorescencją i przewyższać swoją szybkością proces konwersji wewnętrznej $S_1 \rightsquigarrow S_0$. Typowe stałe szybkości k_{ISC} przejść $S_1 \rightsquigarrow T_1$ są rzędu 10^6–10^7 s^{-1}. A więc istotnie jest to jeszcze jeden ważny sposób dezaktywacji bezpromienistej stanu singletowego S_1. Fakt ten ma bardzo ważne i interesujące konsekwencje i te za chwilę poznamy.

W wyniku przejścia bezpromienistego $S_1 \rightsquigarrow T_1$ cząsteczka przechodzi do wzbudzonego stanu trypletowego. Innym sposobem osiągnięcia stanu T_1 byłaby bezpośrednia absorpcja promieniowania przez cząsteczkę w stanie podstawowym S_0 z przejściem $S_0 \rightarrow T_1$. Wyjaśniliśmy już poprzednio, że przejście takie ma znikomo małe natężenie w porównaniu z przejściami dozwolonymi, a nawet w porównaniu z przejściami wzbronionymi, np. ze względu na symetrię. Oczywiście jeśli istnieje sprzężenie spin-orbita w cząsteczce przejście $S_0 \rightarrow T_1$ będzie wzmacniane i moglibyśmy oczekiwać, że będzie ono wzmacniane w takim samym

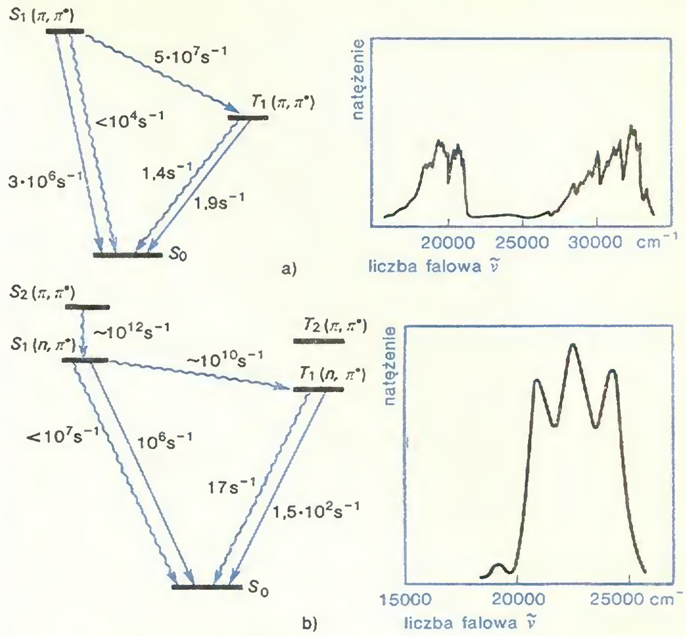
stopniu, jak przejście $S_1 \rightsquigarrow T_1$. Tak jednak nie jest, gdyż i w tym wypadku istotną rolę w wielkości wzmocnienia przejścia odgrywa wspomniany już przez nas czynnik odstępu energetycznego pomiędzy oddziałującymi stanami. Odległość między S_0 i T_1 jest z reguły dużo większa niż odległość między S_1 i T_1 i to właśnie sprawia, że oba przejścia nie są jednakowo prawdopodobne.

Moglibyśmy zapytać się, czy ta różnica ma jakieś istotne następstwa? Otóż tak, różnica ta ma konsekwencje natury praktycznej. Bardzo małe prawdopodobieństwo absorpcji $S_0 \rightarrow T_1$ sprawia, że w ten sposób właściwie nie można osiągnąć stanu trypletowego; oczywiście w zbiorze cząsteczek stan trypletowy jest osiągalny w wyniku absorpcji $S_0 \rightarrow T_1$, ale procent ogólnej liczby cząsteczek wzbudzonych w ten sposób do stanu T_1 jest znikomo mały — nie do wykrycia zwykłymi metodami. Mówimy, że obsadzenie (populacja) stanu T_1 osiągane w ten sposób jest niemierzalnie małe. Skoro tak, to pozostaje nam ewentualnie próba osiągnięcia stanu trypletowego w bardziej skomplikowany sposób, a więc przez wzbudzenie cząsteczek do stanu singletowego S_1 (lub wyższego) w nadziei, że przejście bezpromieniste $S_1 \rightsquigarrow T_1$ będzie procesem na tyle wydajnym, ażeby liczba cząsteczek przechodzących do stanu T_1 stanowiła znaczny procent ogólnej liczby cząsteczek wzbudzonych. Tak też jest w istocie. Zaproponowany przez nas sposób osiągania znacznego obsadzenia stanu trypletowego jest realizowany w cząsteczkach i stanowi właściwie jedyny sposób obsadzania stanu trypletowego.

Cząsteczka, która znalazła się w stanie trypletowym T_1, ma teraz do wyboru zasadniczo dwa sposoby opuszczenia tego stanu. Może tego dokonać albo przez przejście promieniste $T_1 \rightarrow S_0$ albo przejście bezpromieniste $T_1 \rightsquigarrow S_0$. Przejściu promienistemu towarzyszy oczywiście emisja. Jest to jednak emisja różna od znanej nam już fluorescencji i dla odróżnienia od niej nazywana fosforescencją. Po pierwsze jest ona, jak wynika ze schematu energetycznego cząsteczki (rys. 42), bardziej długofalowa niż fluorescencja. Po drugie, mówiąc o czasach życia stanów wzbudzonych doszliśmy do wniosku, że dla stanu trypletowego czas ten jest znacznie dłuższy niż dla stanu singletowego. Oczywiście i czas zaniku fosforescencji będzie wobec tego znacznie dłuższy od czasu zaniku fluorescencji. Czas ten, przez analogię do czasu zaniku fluorescencji, możemy wyrazić jako $\tau_p = 1/(k_p+k_d)$, gdzie tym razem k_p jest stałą szybkości przejścia promienistego $T_1 \rightarrow S_0$, a k_d — stałą szybkości przejścia bezpromienistego $T_1 \rightsquigarrow S_0$. Dosyć trudno jest podać typowe czasy zaniku fosforescencji, a to dlatego, że są one bardzo zależne od warunków zewnętrznych. Na przykład czas zaniku fosforescencji cząsteczki naftalenu może zmieniać się blisko milion razy w zależności od tego, w jakim ośrodku fosforescencja jest obserwowana. W ośrodku ciekłym, w temperaturze pokojowej, może on być rzędu 10^{-5} s, a w ośrodku usztywnionym (w tak zwanym szkliwie niskotemperaturowym, które uzyskuje się przez ochłodzenie mieszaniny pewnych węglowodorów do temperatury ciekłego azotu, tj. do 77 K) może sięgać 2 s. Te ogromne zmiany czasu zaniku wynikają głównie z tego, że promienisty czas życia stanu trypletowego jest długi i wobec tego stan ten może ulegać dezaktywacjom bezpromienistym w wyniku zderzeń z cząsteczkami ośrodka i taka dezaktywacja — bardzo wydajna w ośrodku ciekłym — może być w znacznym stopniu zahamowana w ośrodku usztywnionym, gdzie cząsteczki nie mają swobody ruchu. Konsekwencją takiego stanu rzeczy jest fakt, że fosforescencji — w odróżnieniu od fluorescencji — nie można w praktyce zaobserwować w ośrodkach ciekłych, można ją dopiero zaobserwować w ośrodkach usztywnionych. Na rys. 45 dla ilustracji naszych wywodów pokazano schematy energetyczne dwóch różnych cząsteczek, podano wartości stałych szybkości różnych procesów promienistych i bezpromienistych, a także widma fluorescencji i fosforescencji.

fosforescencja

Patrząc na ten rysunek, można zadać pytanie, w jaki sposób wyznaczono te wszystkie stałe szybkości. Sprawa niestety nie jest prosta i uzyskanie pełnego



Rys. 45. Schemat poziomów i widmo emisji: a) cząsteczki chloronaftalenu; stała szybkość przejścia bezpromienistego $S_1 \rightsquigarrow T_1$ jest tylko nieco większa od stałej szybkości przejścia promienistego $S_1 \rightarrow S_0$ i w widmie emisji można obserwować zarówno fluorescencję jak i fosforescencję — pasmo ok. 20 000 cm^{-1}; b) cząsteczki benzofenonu; stała szybkości przejścia bezpromienistego $S_1 \rightsquigarrow T_1$ jest blisko 10 000 razy większa niż stała szybkości przejścia promienistego $S_1 \rightarrow S_0$ i dla tej cząsteczki w ogóle nie obserwuje się fluorescencji, jej emisja — to wyłącznie fosforescencja

obrazu procesów zachodzących w stanach wzbudzonych jest rzeczą raczej dosyć trudną, a czasem wręcz niemożliwą. Podstawowe informacje uzyskuje się oczywiście z pomiarów czasów zaniku fluorescencji i fosforescencji oraz z pomiarów ich wydajności kwantowych. Dla fosforescencji, tak jak i dla fluorescencji, można zdefiniować wydajność kwantową, chociaż tym razem odpowiednie wyrażenie będzie bardziej złożone i będzie musiało zawierać również stałą szybkości przejścia bezpromienistego $S_1 \rightsquigarrow T_1$, w wyniku którego obsadzone są tryplety. Oczywiście uzyskane z pomiarów czasów zaniku i wydajności informacje są na ogół niewystarczające i trzeba się uciekać do innych pomiarów pomocniczych, bądź też czynić różne założenia upraszczające problem. Dodatkowe komplikacje wynikają z faktu, że nasze schematy obrazują cząsteczkę izolowaną. W rzeczywistości cząsteczka występuje zawsze w pewnym otoczeniu, z którym w mniejszym lub większym stopniu oddziałuje. Te oddziaływania wywołują różne procesy (o niektórych z nich będziemy jeszcze mówili), które znakomicie komplikują podany schemat dezaktywacji. Chcąc uwzględnić te dodatkowe procesy musimy wprowadzić do naszego schematu opisujące je dodatkowe stałe szybkości (oczywiście musimy sobie przedtem zdawać sprawę z tego, jaki jest mechanizm tych procesów — jak one przebiegają). Nakreślony obraz jest zapewne zniechęcający, ale naprawdę nie jest aż tak źle i w wielu wypadkach potrafimy poradzić sobie z rozwikłaniem tych problemów, jeśli nawet niezupełnie ściśle, to na ogół z dokładnością wystarczającą do wyrobienia sobie poglądu na to, co się dzieje i co może się dziać ze wzbudzoną cząsteczką. Niestety, nie będziemy tutaj mogli przedstawić różnych metod eksperymentalnych i sposobów postępowania, za pomocą których rozwiązuje się tego typu problemy; wszystkich, którzy pragną osiągnąć wyższy stopień

wydajność kwantowa fosforescencji

wtajemniczenia odsyłamy do literatury tego przedmiotu, podanej na końcu tego rozdziału.

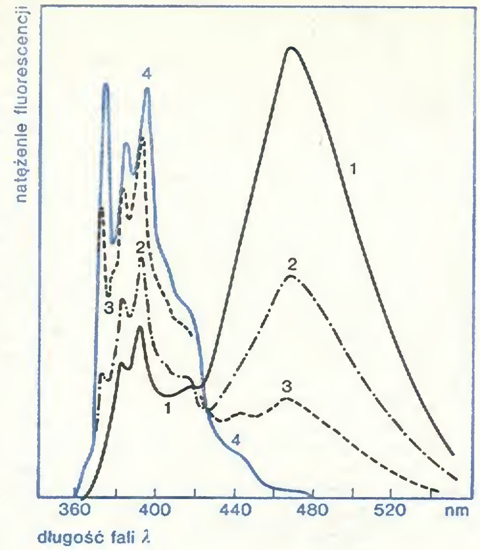
fluorescencja opóźniona

Fluorescencja biorąca swój początek ze stanu singletowego S_1 i fosforescencja biorąca początek ze stanu trypletowego T_1 nie wyczerpują problemu procesów emisyjnych, z jakimi można się spotkać w cząsteczce. Czasem można zaobserwować emisję, której mechanizm jest bardziej złożony niż poznane dotychczas. Opisując różne przejścia promieniste i bezpromieniste pominęliśmy bowiem jeszcze jedną możliwość, polegającą na przejściu odwrotnym $T_1 \rightsquigarrow S_1$. Ponieważ stan T_1 ma niższą energię niż stan S_1, przejście takie wymaga dodatkowej energii, która musi być dostarczona cząsteczce z zewnątrz, np. w wyniku zderzenia z cząsteczkami ośrodka. Przejście takie wymaga, jak mówimy, pewnej energii aktywacji. Jeżeli odstęp pomiędzy stanami S_1 i T_1 nie jest zbyt duży, to energia ruchów cieplnych w ośrodku może okazać się wystarczająca do zaktywizowania przejścia $T_1 \rightsquigarrow S_1$. W ten sposób cząsteczka znajdzie się z powrotem w stanie S_1, a stąd może powrócić do stanu podstawowego, np. w wyniku przejścia promienistego, emitując oczywiście promieniowanie. Ciąg procesów prowadzących do takiej emisji jest następujący: absorpcja $S_0 \rightarrow S_1$, bezpromieniste przejście $S_1 \rightsquigarrow T_1$, przejście aktywowane termicznie $T_1 \rightsquigarrow S_1$ i wreszcie emisja $S_1 \rightarrow S_0$. Widmo takiej emisji pokrywa się z widmem fluorescencji, ale między tymi dwoma rodzajami emisji jest istotna różnica. Cząsteczka, która przebywała w stanie trypletowym i powróciła do stanu singletowego, przebywała w stanie wzbudzonym (a ściślej w dwóch stanach wzbudzonych) znacznie dłużej niż cząsteczka, która powróciła do stanu podstawowego bezpośrednio ze wzbudzonego stanu singletowego. A zatem czasy zaniku obu fluorescencji są inne. Dlatego też emisja, o której mówimy, nazywa się fluorescencją opóźnioną, a jej czas zaniku jest określony przez czas życia stanu trypletowego.

Poznaliśmy różne typy emisji prowadzące do różnych widm emisyjnych w cząsteczkach. Poznaliśmy również różne procesy bezpromieniste konkurujące z emisją, bądź też czasem umożliwiające jej obserwację (jak to się dzieje z obserwacją fosforescencji i przejściem bezpromienistym $S_1 \rightsquigarrow T_1$). Oczywiście, to czy dla danej cząsteczki można obserwować fluorescencję czy fosforescencję, czy też obie jednocześnie, a także czy nie można ich w ogóle obserwować, zależy przede wszystkim od budowy, struktury elektronowej cząsteczki i innych naturalnych czynników, takich jak np. sprzężenie spin-orbita, które możemy określić ogólnym mianem — własności cząsteczki. Zdajemy sobie również sprawę, że bardzo ważną rolę w obserwacji widm emisji cząsteczek odgrywają własności ośrodka, które mogą zarówno wzmacniać, jak i osłabiać emisję. Znając naturalne własności cząsteczki i ośrodka, w którym się znajduje, możemy przewidywać, jak wygląda emisja i jej podstawowe parametry, takie jak czas zaniku i wydajność. I odwrotnie, badając widma emisyjne i ich parametry możemy wnioskować o budowie i własnościach cząsteczki (a także i ośrodka, w którym cząsteczka się znajduje) oraz o ewentualnych oddziaływaniach międzycząsteczkowych. Dlatego też na zakończenie naszych rozważań podamy przykład takich oddziaływań, o których można wnioskować na podstawie badania widm emisji.

fluorescencja pyrenu

Zacznijmy od przykładu, który ilustruje rys. 46. Rysunek ten przedstawia zmiany, jakim ulega widmo fluorescencji pyrenu, wtedy gdy zmieniane jest stężenie pyrenu w roztworze. Gdy stężenie pyrenu jest stosunkowo niskie, wówczas jego widmo fluorescencji jest takie, jak widmo *4* na rys. 46. Kiedy stężenie pyrenu w roztworze wzrasta (kolejne widma *3*, *2* i *1*) natężenie tego pasma fluorescencji maleje, a jednocześnie pojawia się nowe pasmo fluorescencji, bardziej długofalowe i pozbawione struktury oscylacyjnej. Natężenie tego drugiego pasma wzrasta w miarę wzrostu stężenia, a natężenie pierwotnego pasma maleje. Mówimy, że fluorescencja ta jest wygaszona, co możemy dokładnie prześledzić na przykładzie widm *3*, *2*, *1* na rysunku. Dodajmy jednocześnie, że w widmie



Rys. 46. Widmo pyrenu w roztworze w różnym stężeniu: *4* niskie stężenie pyrenu, *3*, *2*, *1* stężenie coraz większe. Pozbawione struktury pasmo o maksimum przy ok. 470 nm jest pasmem fluorescencji ekscimerów

absorpcji nie obserwuje się żadnych zmian przy zmianach stężenia. Co jest przyczyną obserwowanych zmian widma fluorescencji? Może i tutaj, tak jak w przypadku kompleksów z przeniesieniem ładunku, mamy do czynienia z powstawaniem kompleksów. Otóż tak jest w istocie, ale są to innego rodzaju kompleksy. Brak zmian widma absorpcji świadczy o tym, że kompleksy te nie mogą być tworzone w stanie podstawowym, innymi słowy, cząsteczki pyrenu w stanie podstawowym nie oddziałują ze sobą wzajemnie, a ściślej mówiąc siły, z jakimi na siebie oddziałują, są tak słabe, że nie prowadzą do wytworzenia kompleksu w stanie podstawowym. Kompleksy takie powstają jednak w stanie wzbudzonym, o czym właśnie świadczą obserwowane zmiany w widmie fluorescencji. Dokładne badania ujawniły, że kompleksy te, zwane ekscimerami, powstają w rezultacie oddziaływania dwóch cząsteczek pyrenu — jednej wzbudzonej, drugiej niewzbudzonej. Nie będziemy tutaj dociekali charakteru takich oddziaływań i bez ich znajomości potrafimy zrozumieć, dlaczego do utworzenia ekscimeru i zaobserwowania zmian w widmie potrzebne jest duże stężenie cząsteczek w roztworze.

ekscimery

Aby mógł powstać ekscimer z dwóch cząsteczek, z których jedna jest wzbudzona, a druga znajduje się w stanie podstawowym, obie cząsteczki muszą znaleźć się dostatecznie blisko jedna od drugiej. Przy tym spotkanie takie musi nastąpić w czasie krótszym niż czas życia wzbudzonej cząsteczki (a jest on w tym wypadku rzędu $3 \cdot 10^{-7}$ s). Jest rzeczą zrozumiałą, że im więcej jest cząsteczek w roztworze, tym większa jest szansa, że w tak krótkim czasie czasteczka wzbudzona natrafi na cząsteczkę niewzbudzoną, bądź odwrotnie, i powstanie wzbudzony kompleks — ekscimer, którego jedną z własności jest to, że najniższy stan, z którego zachodzi emisja, ma niższą energię niż wzbudzony stan singletowy swobodnej cząsteczki pyrenu. Stąd też fluorescencja ekscimeru jest zawsze bardziej długofalowa niż fluorescencja cząsteczki pyrenu (ta ostatnia, w kontekście omawianych kompleksów, nazywana jest czasem fluorescencją monomerową).

Znamy wiele cząsteczek, które mogą tworzyć kompleksy w stanie wzbudzonym. Okazuje się ponadto, że kompleksy takie mogą być również tworzone po-

hetero-ekscimery (ekscipleksy)

między różnymi cząsteczkami, z których jedne są wzbudzone, a drugie w stanie podstawowym. Takie kompleksy nazywane są heteroekscimerami lub ekscipleksami. I w takich przypadkach obserwujemy wygaszanie widma fluorescencji jednej z cząsteczek, przy wzroście stężenia drugiej cząsteczki (niewzbudzonej) w roztworze, z jednoczesnym pojawieniem się nowego pasma fluorescencji — fluorescencji kompleksu.

międzycząsteczkowe przenoszenie energii wzbudzenia

Na zakończenie wspomnimy jeszcze o pewnej niezwykle ważnej kategorii zjawisk, które są przedmiotem badań spektroskopii molekularnej i które ujawniają się przez zmiany widm fluorescencji bądź fosforescencji. Zjawiska te określamy ogólną nazwą międzycząsteczkowego przenoszenia energii wzbudzenia. Procesy przenoszenia energii polegają na bezpromienistym przenoszeniu energii wzbudzenia elektronowego od wzbudzonej cząsteczki zwanej donorem do niewzbudzonej cząsteczki zwanej akceptorem. Rezultatem przenoszenia energii jest powrót wzbudzonego donora do stanu podstawowego i przejście niewzbudzonego akceptora do stanu wzbudzonego. Schematycznie proces taki możemy zapisać następująco:

$$D+h\nu \rightarrow D^{*}+A \rightarrow A^{*}+D,$$

gdzie D i A oznaczają cząsteczki donora i akceptora w ich stanach podstawowych, a A^* i D^* — ich stany wzbudzone.

Mechanizmy fizyczne przenoszenia energii są różne i zależą między innymi od tego, z jakich stanów donora (singletowego czy trypletowego) do jakich stanów akceptora jest ona przenoszona. Aby proces przenoszenia energii był efektywny, wymagany jest w niektórych wypadkach bezpośredni kontakt cząsteczki donora i akceptora — tak jest wtedy, gdy energia przenoszona jest między stanami trypletowymi donora i akceptora. W innych wypadkach taki bezpośredni kontakt nie jest w ogóle potrzebny i energia wzbudzenia może być przenoszona na bardzo znaczną odległość nawet rzędu kilku nanometrów. Jeżeli proces przenoszenia energii jest efektywny, to wówczas emisja donora — fluorescencja bądź fosforescencja — jest w obecności cząsteczek akceptora silnie wygaszona, pojawia się natomiast odpowiednia emisja — fluorescencja bądź fosforescencja akceptora. Taką emisję akceptora, która nie jest bezpośrednio wzbudzana, lecz jest wynikiem procesu przenoszenia energii nazywa się czasem sensybilizowaną (uczuloną) fluorescencją.

fluorescencja sensybilizowana

N. L. Albert, W. E. Kaiser, H. A. Szymański *Spektroskopia w podczerwieni — teoria i praktyka*, Warszawa 1974; P. W. Atkins *Molekularna mechanika kwantowa*, Warszawa 1974; G. Herzberg *Molecular Spectra and Molecular Structure* v. 1–3, Princetown 1966; Z. Kęcki *Podstawy spektroskopii molekularnej*, Warszawa 1972; *Organic Molecular Photophysics*, ed. J. B. Bergs, v. 1, 2, 1975; C. N. R. Rao *Spektroskopia elektronowa związków organicznych*, Warszawa 1981; J. P. Simons *Fotochemia i spektroskopia*, Warszawa 1976.

Spektroskopia mikrofalowego rezonansu rotacyjnego

Jan Stankowski

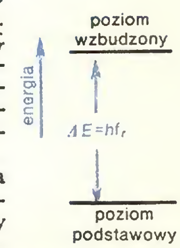
Spektroskopia mikrofalowego rezonansu rotacyjnego (obejmująca badania substancji w stanie gazowym) jest jednym z działów radiospektroskopii, czyli spektroskopii długofalowej części widma promieniowania elektromagnetycznego o długościach fal większych od 0,1 cm. Innymi ważnymi metodami radiospektroskopii są metody rezonansów magnetycznych (→ Spektroskopia rezonansów magnetycznych).

Mikrofalowy rezonans rotacyjny (MRR — ang. *Microwave Rotational Resonance*) obserwuje się w cząsteczkach obdarzonych elektrycznym momentem dipolowym; wartość momentu zmienia się w procesie absorpcji lub emisji fotonu. Typowymi cząsteczkami tego rodzaju są: HCN, NH_3, H_2O.

rozszczepienie inwersyjne NH_3

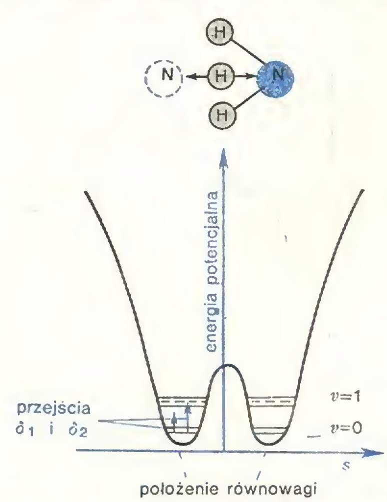
W rozwoju mikrofalowej spektroskopii gazów doniosłą rolę odegrała cząsteczka amoniaku. Cząsteczka amoniaku NH_3 jest spłaszczonym bąkiem symetrycznym. Poziomy oscylacyjne (→ Spektroskopia molekularna, rozdz. Widma oscylacyjne i rotacyjne) cząsteczki NH_3 ulegają rozszczepieniu wskutek inwersyjnego przemieszczania atomu azotu, polegającego na jego przenikaniu przez barierę potencjału oddzielającą dwa położenia równowagi leżące po obu stronach płaszczyzny wyznaczonej przez atomy wodoru (rys. 1). (Zjawisko przenikania cząstki o energii mniejszej od wysokości bariery nazywa się efektem tunelowym, gdyż zachodzi ono jakby dzięki tunelowi w barierze potencjału; proces ten można zrozumieć przyjmując falowy obraz cząstki, bowiem tylko w takim przedstawieniu istnieje skończone, różne od zera, prawdopodobieństwo znajdowania się cząstki poza barierą potencjału). Wartość rozszczepienia inwersyjnego NH_3 wynosi 0,8 cm^{-1} (leży więc w obszarze mikrofal) dla oscylacyjnego poziomu podstawowego ($v = 0$) oraz 36,5 cm^{-1} dla poziomu wzbudzonego ($v = 1$). Rozszczepienie oscylacyjnego poziomu podstawowego (0,8 cm^{-1}) jest wykrywane metodą MRR — odkryli je C. E. Cleeton i N. H. Williams (1933 r.).

Spektroskopia polega na badaniu rezonansowego oddziaływania fotonów z układem molekularnym. Rezonans zachodzi wtedy, gdy energia fotonu hf_r (rys. 2) odpowiada różnicy energii dwóch dyskretnych poziomów energetycznych układu molekularnego. W rezonansie atomy (cząsteczki) silnie absorbują lub emitują promieniowanie co prowadzi do wystąpienia linii rezonansowej (rys. 3).



Rys. 2.

Badanie absorpcji w pasmie mikrofal polega na obserwowaniu zmian dobroci rezonatora mikrofalowego (→ Generacja mikrofal), gdy w sposób liniowy zmienia się częstość źródła fal elektromagnetycznych. Mikrofalowe widmo amoniaku uzyskane dla róż-

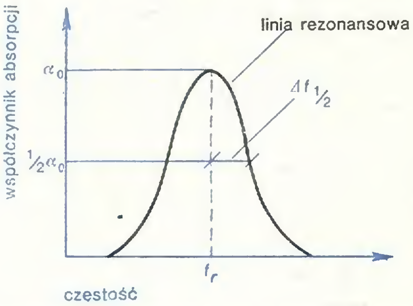


Rys. 1. Zależność energii potencjalnej atomu azotu w cząsteczce NH_3 od wychylenia S od położenia równowagi; δ_1 i δ_2 składowe dubletu deformacyjnego, uwarunkowane inwersyjnymi rozszczepieniami poziomu $v = 1$

nych ciśnień gazu w rezonatorze przedstawia rys. 4. Pod normalnym ciśnieniem obserwuje się silną absorpcję w szerokim zakresie częstości, a obserwowana całkowita szerokość widma amoniaku jest odwrotnie proporcjonalna do czasu między kolejnymi wzajemnymi zderzeniami cząsteczek. W miarę obniżania ciśnienia zderzenia stają się coraz rzadsze i szerokie pasmo absorpcyjne zwęża się; ukazuje się nam wieloskładnikowe widmo inwersyjne NH_3 (rys. 4b).

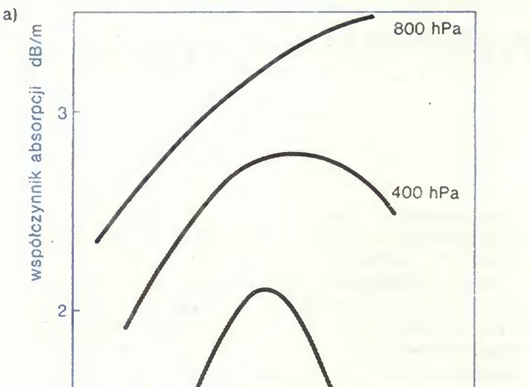
zjawisko Starka

Ze względu na to, że widmo amoniaku związane jest z elektrycznymi przejściami dipolowymi, częstość rezonansowa poszczególnych linii silnie zależy od zewnętrznego pola elektrycznego (zjawisko Starka).



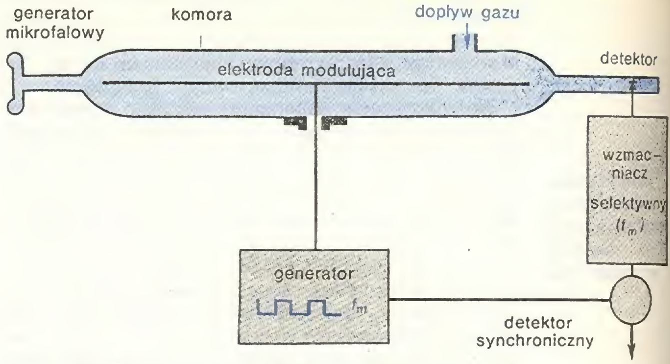
Rys. 3. Linia rezonansowa; szerokością połówkową $\Delta f_{1/2}$ linii rezonansowej jest różnica częstości w połowie wysokości linii

widmo mikrofalowe amoniaku



Rys. 4. Widmo mikrofalowe amoniaku przy różnym ciśnieniu; przy ciśnieniu 1 Pa ujawnia się struktura inwersyjna. Cyfry przy składowych oznaczają liczby kwantowe *J*, *K* charakteryzujące rotacyjny ruch cząsteczki

Położenie linii rezonansowej można zmieniać za pomocą sygnału prostokątnego o częstości f_m uzyskując periodyczną modulację linii rezonansowej. Mo-



Rys. 5. Schemat spektrometru MRR z modulacją starkowską

modulacja starkowska

dulację za pomocą pola elektrycznego, zwaną modulacją starkowską, wykorzystuje się do zwiększenia czułości spektrometrów MRR. Najczęściej spotykane rozwiązanie spektrometru MRR jest pokazane na rys. 5. Główną częścią spektrometru jest falowód zawierający badany gaz. Do elektrody przykłada się prostokątne impulsy wysokiego napięcia wytwarzające pole elektryczne między elektrodą i ścianką falowodu. Modulacja starkowska powoduje periodyczną zmianę położenia linii rezonansowej, co pociąga za sobą periodyczną, z częstością modulacji, zmianę mocy na detektorze. Umieszczony za detektorem wzmacniacz nastrojony na częstość modulacji starkowskiej f_m, rejestruje linię rezonansową podczas liniowej zmiany częstości źródła mikrofal.

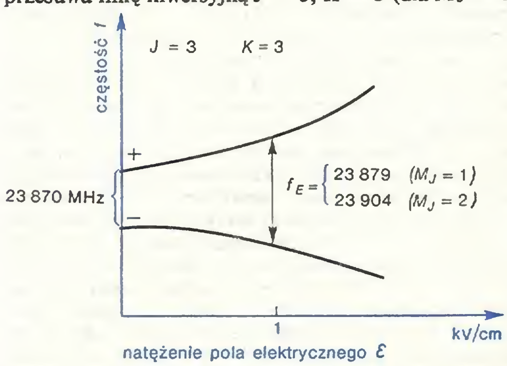
Maser i wzorce częstości

W cząsteczce NH_3 występuje kwadratowy efekt Starka, tzn. przesunięcie linii rezonansowej w polu elektrycznym jest proporcjonalne do kwadratu przyłożonego pola elektrycznego. Dla określonego stanu rotacyjnego cząsteczki przesunięcie częstości rezonansowej $\Delta f = f_E - f_r$ w polu E opisuje wyrażenie:

$$\Delta f = AE^2,$$

kwadratowy efekt Starka

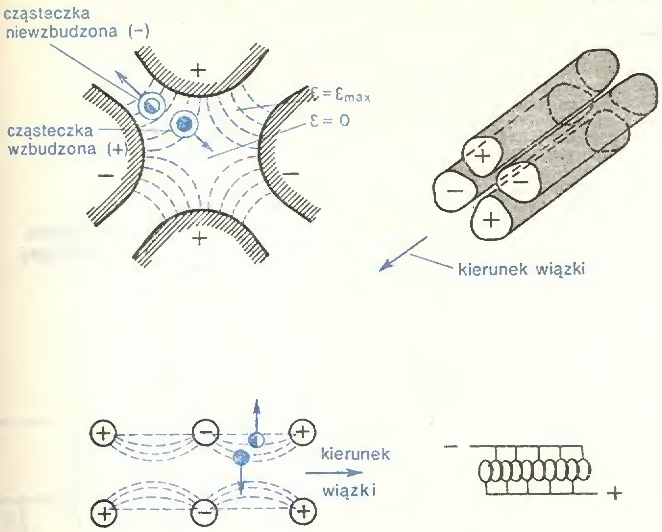
gdzie A jest wielkością zależną od momentu dipolowego badanej cząsteczki oraz od wartości rzutu całkowitego momentu pędu J na kierunek pola E, który oznaczamy przez M_J. Dla stanów inwersyjnych cząsteczek amoniaku kwadratowy efekt Starka przedstawiono na rys. 6. Pole o natężeniu 1 kV/cm przesuwa linię inwersyjną $J = 3$, $K = 3$ (dla $M_J = 1$)



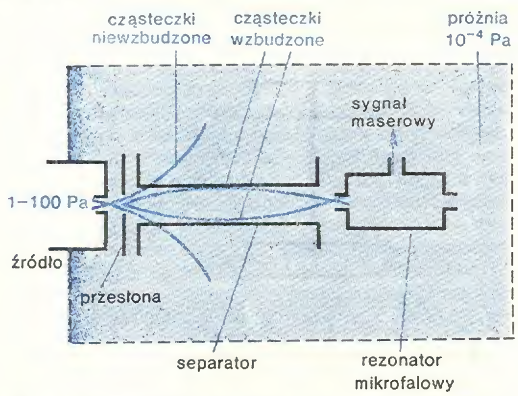
Rys. 6. Kwadratowy efekt Starka w widmie inwersyjnym amoniaku

od częstości $f_r = 23\,870$ MHz do wartości $f_E = 23\,879$ MHz. Widać stąd, że silnie niejednorodne pole elektryczne będzie separować obydwa stany inwersyjne.

Takie oddziaływanie niejednorodnego pola elektrycznego na cząsteczki amoniaku wykorzystano w części separująco-ogniskującej (rys. 7) urządzenia zwanego



Rys. 7. Separator cząsteczek wzbudzonych: a) kwadrupolowy, b) kółkowy



Rys. 8. Schemat masera z wiązką molekularną amoniaku

maserem (z ang. *microwave amplification by stimulated emission of radiation*). Z wiązki molekularnej biegnącej wzdłuż osi separatora cząsteczki niewzbudzone (−) będą do pola wciągane, podczas gdy cząsteczki wzbudzone (+) będą z niego wypychane.

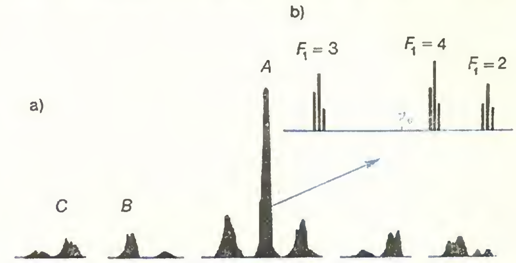
maser amoniakalny

Schemat masera amoniakalnego przedstawiony jest na rys. 8. Ze źródła, w którym amoniak znajduje się pod ciśnieniem ok. 10 Pa cząsteczki NH_3 wpadają do obszaru wielkiej próżni (10^{-4} Pa). Tutaj na całej swojej drodze od źródła do rezonatora cząsteczki nie zderzają się wzajemnie. W obrębie separatora wiązka molekularna pozbawiona jest cząsteczek niewzbudzonych i do rezonatora mikrofalowego, nastrojonego na linię inwersyjną wpadają tylko cząsteczki wzbudzone. Ponieważ w zakresie mikrofal cząsteczki wzbudzone przechodzą do stanu podstawowego tylko w wyniku wymuszonej emisji promieniowania, tylko zderzenie między cząsteczkami lub kwant o częstości rezonansowej może spowodować akt emisji. W wyniku przypadkowego aktu emisji cząsteczki wzbudzają elektromagnetyczne drgania w rezonatorze. Dla małego strumienia cząsteczek drgania te gasną wskutek strat energii pola elektromagnetycznego w ściankach rezonatora. Zwiększając strumień cząsteczek ze źródła możemy spowodować, że energia wnoszona do rezonatora przez wzbudzone cząsteczki stanie się większa od strat rezonatora i pole elektromagnetyczne będzie podtrzymywane przez emisję cząsteczek NH_3. Tak powstające pole w.cz. będzie wymuszało dalsze przejścia w cząsteczkach wpadających do rezontatora i to spowoduje trwałe wzajemne sprzężenie rezonatora mikrofalowego z wiązką molekularną.

Maser jest źródłem drgań pola w.cz. o nadzwyczajnie dużej stabilności, gdyż drgania te są utrzymywane przez promieniowanie cząsteczek o określonej budowie, których częstość mało się zmienia pod działaniem czynników zewnętrznych. Za pomocą masera zostały określone najsubtelniejsze oddziaływania w molekule NH_3 (rys. 9).

maser wodorowy

Promieniowanie o najwyższej stabilności częstości można uzyskać za pomocą masera wodorowego. Maser ten pracuje przy częstości 1420 MHz równej rozszczepieniu poziomów: singletowego $F = 0$ i trypletowego $F = 1$ wodoru. W maserze wodorowym wykorzystuje się magnetyczne przejścia dipolowe i dlatego wiązkę atomów wzbudzonych uzyskuje się w separatorze magnetycznym o silnej niejednorodności pola magnetycznego. Graniczną stabilność tego masera określa się na 10^{-14} co stanowi światowy rekord stabilności częstości. Maser wodorowy jest



Rys. 9. Nadsubtelna struktura linii $J = 3$, $K = 3$ amoniaku $^{14}NH_3$: a) A, B i C struktura kwadrupolowa, b) struktura magnetyczna linii A

urządzeniem kontrolnym dla wzorca cezowego, spełniającego obecnie rolę światowego wzorca częstości.

wzorzec cezowy

Wzorzec cezowy jest spektrometrem absorpcyjnym z wiązką atomów cezu, w którym zastosowano sprzężenie zwrotne tak, że częstość źródła mikrofal ustala się za pomocą wybranej linii rezonansowej. Wybrana umownie linia cezowego wzorca ma częstość $f_{Cs} = 9\,192\,631\,770$ Hz. Widmo cezu jest dzisiaj podstawą atomowej skali czasu przewyższającej swoją precyzją skale astronomiczne związane z periodycznym ruchem ciał niebieskich. Ruch planet i wszelkie zjawiska astronomiczne opisuje się właśnie w atomowej skali czasu, wyznaczonej przez laboratoria czasu posługujące się zegarami cezowymi i maserami wodorowymi.

Masery zapoczątkowały elektronikę kwantową i wskazały na możliwość realizacji laserów, które spowodowały wspaniały rozwój optyki nieliniowej i zadecydowały o dalszych sukcesach elektroniki kwantowej (→ Optyka nieliniowa, Lasery — podstawy działania).

ujemna temperatura bezwzględna

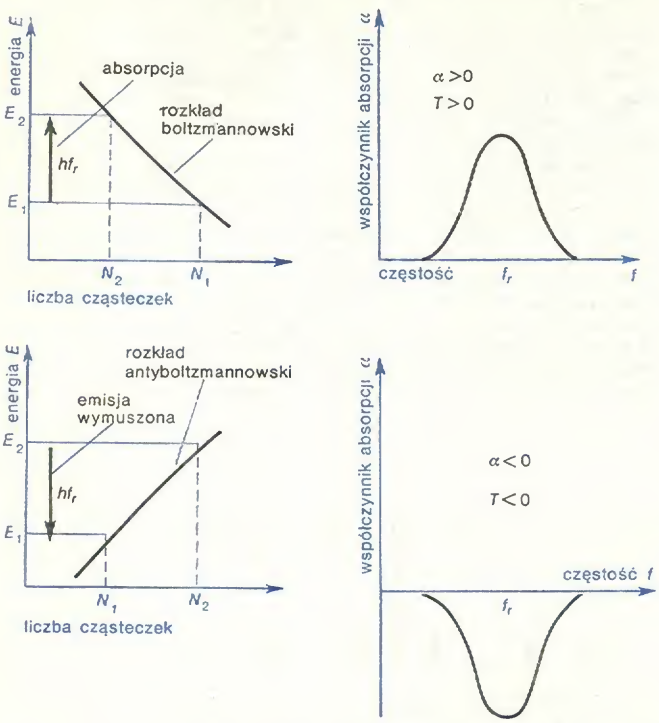
Masery i lasery spowodowały wzrost zainteresowania stanem układów znajdujących się w ujemnej temperaturze bezwzględnej. Termin ten zrazu szokuje, gdyż wiemy, że zero bezwzględne jest nieosiągalne. Nie jest on użyty tutaj w dotychczasowym sensie, gdyż bezwzględną temperaturę ujemną nie określa energia kinetyczna chaotycznego ruchu cząsteczek, lecz energia ich dyskretnych stanów. Temperatury nie mierzy się termometrem lecz za pomocą obsadzenia dyskretnych stanów: ujemnej temperaturze odpowiada antyboltzmannowskie obsadzenie stanów — w wyższym stanie znajduje się więcej cząsteczek niż w niższym. Nie ma również sprzeczności z naszą wiedzą o zerze bezwzględnym, gdyż ze strony ujemnych temperatur bezwzględnych zero bezwzględne również jest nieosiągalne.

W warunkach zwykłej równowagi termodynamicznej, energia poszczególnych atomów i cząsteczek opisana jest za pomocą tzw. rozkładu boltzmannowskiego (rys. 10). Zgodnie z tym rozkładem w zbiorze cząsteczek najliczniejsze są cząsteczki o najmniejszej energii. Rozkład boltzmannowski opisuje również obsadzenie stanów o skwantowanej energii: liczbę cząsteczek N_1 i N_2 znajdujących się w dwu dowolnych stanach *1* i *2* określa funkcja wykładnicza (rys. 10). Gdy $E_1 < E_2$ stosunek

$$N_2/N_1 = e^{-\Delta E/kT}, \quad (1)$$

gdzie $\Delta E = E_2 - E_1$, zaś k jest stałą Boltzmanna. Jeżeli przyj-

miemy to wyrażenie za wzór definiujący temperaturę, za pomocą obsadzenia poziomów energetycznych możemy określić temperaturę dowolnego przejścia rezonansowego. „Termometrem” jest wówczas absorpcja P_{abs} mocy mikrofalowej, która jest wprost proporcjonalna do różnicy obsadzeń dwu rozważanych poziomów energetycznych cząsteczki: $P_{abs} \sim N_1 - N_2$. W gazie,



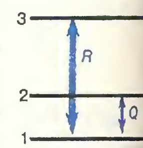
Rys. 10. Obsadzenie poziomów określa temperaturę układu: a) $T > 0$, b) $T < 0$

w którym cząsteczki we wszystkich stanach energetycznych znajdują się w równowadze z otoczeniem, moc absorbowana jest dodatnia: $P_{abs} > 0$, a więc mamy do czynienia z dodatnią temperaturą bezwzględną dla dowolnie wybranego dwupoziomowego układu. Zupełnie inną sytuację mamy wówczas, gdy dla dwóch wyizolowanych poziomów $N_2 > N_1$, tzn., że górny stan jest silniej obsadzony niż dolny. Zgodnie z tym co wyżej powiedziano wystąpi ujemna absorpcja równoważna emisji, a układ rozważany możemy opisać za pomocą ujemnej temperatury bezwzględnej. Widać stąd natychmiast, że układy o temperaturze dodatniej ($T > 0$) absorbują energię, gdyż $P_{abs} > 0$, a układy o ujemnej temperaturze ($T < 0$) ją emitują, bowiem dla nich współczynnik absorpcji jest mniejszy od zera ($\alpha < 0$). Dzięki maserom rozwinięto termodynamikę układów o ujemnej temperaturze bezwzględnej, w której określono warunki jakie musi spełniać taki układ: 1) W układzie dwupoziomowym musi istnieć równowaga w sensie termodynamicznym, co w doświadczeniach mikrofalowych zapewnia szybko ustalające się oddziaływanie spin-spin. 2) Rozważany układ dwupoziomowy jest adiabatycznie odizolowany od układów o $T > 0$ (duża różnica czasów relaksacji spin-sieć i spin-spin). 3) Energia układu musi być ograniczona od góry, co jest spełnione w układzie o skwantowanych poziomach energetycznych, gdyż jest to po prostu energia górnego poziomu.

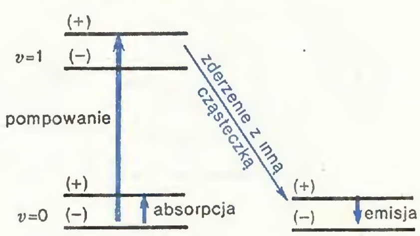
Zasadnicza różnica układów o dodatniej i ujemnej temperaturze bezwzględnej polega na tym, że w układach z temperaturą $T > 0$ rozkład obsadzeń jest wynikiem chaotycznego ruchu cieplnego cząsteczek (ich kinetycznej energii), zaś ujemna temperatura bezwzględna jest związana z energią potencjalną poszczególnych cząsteczek tworzących układ. Osobliwością układów o ujemnej temperaturze bezwzględnej jest to, że układy o $T < 0$ są „gorętsze” od układów o $T > 0$, bowiem zawierają one energię zmagazynowaną we wzbudzonych stanach cząsteczek, którą na drodze wymuszonej emisji promieniowania poszczególne cząsteczki oddają polu elektromagnetycznemu.

rezonans podwójny

Innymi ważnymi metodami mikrofalowej spektroskopii gazów jest rezonans podwójny i molekularny rezonans rotacyjny. Rezonans podwójny stosowany jest do detekcji przejść, gdy współczynnik pochłaniania dla danego przejścia leży poniżej czułości spektrometru MRR. W metodzie tej stosuje się równocześnie dwa sygnały mikrofalowe: sygnał o dużej mocy powodujący zmianę obsadzeń poziomów energetycznych (przejście *R*, rys. 11) i słaby sygnał wykrywający przejście *Q*. W układzie trzech poziomów, między którymi zachodzą dwa przejścia *R* i *Q*, pompowanie silnym sygnałem *R* opróżnia w istotny sposób poziom *1*, a tym samym staje się możliwa obserwacja słabego przejścia *Q*. Metodą podwójnego rezonansu zbadano cząsteczkę HDCO: przejście *R* (35 GHz) naświetlane silnym sygnałem mikrofalowym umożliwiło obserwację przejścia *Q* (13 GHz).



Rys. 11. Rezonans podwójny



Rys. 12. Molekularny rezonans rotacyjny

molekularny rezonans rotacyjny

Molekularny rezonans rotacyjny (rys. 12) polega na przekazywaniu energii rotacyjnej podczas zderzeń cząsteczek, w wyniku czego zmienia się obsadzenie inwersyjnych poziomów energetycznych. Jest to ważna metoda badania energii zderzeń cząsteczek.

G. M. Barrow *Wstęp do spektroskopii molekularnej*, Warszawa 1968: J. Stankowski, A. Graja *Wstęp do elektroniki ciała stałego*, Warszawa 1972.

Spektroskopia rezonansów magnetycznych

Marek Gutowski

rezonans magnetyczny

Wyobraźmy sobie pewną substancję umieszczoną w stałym polu magnetycznym. Jeżeli się okaże, że ta substancja pod wpływem przyłożonego pola magnetycznego zaczyna pochłaniać fale elektromagnetyczne pewnych długości, zjawisko nazwiemy zjawiskiem rezonansu magnetycznego. Badanie zależności pochłaniania od wielkości indukcji pola magnetycznego *B* i od energii padających fal elektromagnetycznych jest właśnie istotą spektroskopii rezonansów magnetycznych. Mówimy „rezonansów”, a nie „rezonansu”, bo znamy kilka rodzajów tych zjawisk: rezonans jądrowy, rezonans elektronowy, rezonans ferromagnetyczny i in. Istnieje jeszcze jeden rodzaj rezonansowego pochłaniania energii w obecności stałego pola magnetycznego, a mianowicie tzw. rezonans cyklotronowy. Nie będziemy go tu omawiać, ponieważ jest to z mikroskopowego punktu widzenia zupełnie inne zjawisko.

Badania rezonansów magnetycznych zostały zapoczątkowane w 1945 r. przez radzieckiego fizyka E. Zawojskiego. Do rozwoju badań eksperymentalnych przyczynili się również E. M. Purcell, H. C. Torrey, R. V. Pound, C. Kittel, F. Bloch, C. Gorter i in. Dzisiaj metody badawcze wykorzystujące rozmaite rezonanse magnetyczne stały się silnym narzędziem eksperymentatorów i badaczy nie tylko w dziedzinie fizyki, ale również chemii, biologii i medycyny. Skonstruowano przyrządy pozwalające określać parametry materiałów używanych w różnych gałęziach przemysłu.

Do sprawy zastosowań rezonansów magnetycznych jeszcze powrócimy. Teraz omówimy samo zjawisko. Najpierw podamy krótki słowniczek objaśniający często używane skróty:

NMR (JRM) — jądrowy rezonans magnetyczny (ang. *nuclear magnetic resonance*),

EPR (ERP, ESR) — elektronowy rezonans paramagnetyczny (ang. *electron paramagnetic resonance*), zwany też rezonansem spinowym,

FMR (RFM) — rezonans ferromagnetyczny (ang. *ferromagnetic resonance*).

Spotyka się niekiedy jeszcze inne skróty, jak np. AFMR, NQR, EQR, ENDOR, które oznaczają kolejno: rezonans antyferromagnetyczny, jądrowy rezonans kwadrupolowy, elektronowy rezonans kwadrupolowy i podwójny rezonans elektronowo-jądrowy. Nie będziemy się tu nimi zajmować.

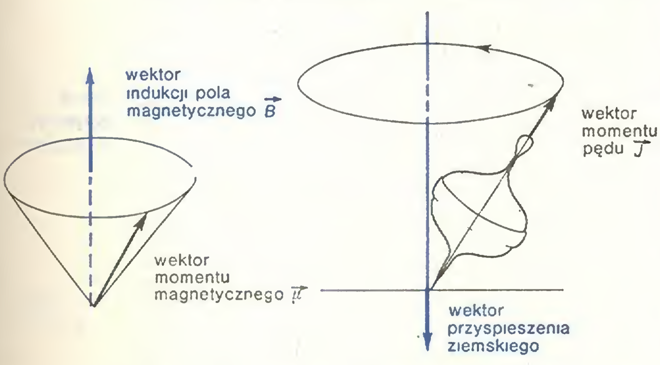
Zjawisko rezonansu magnetycznego

Zjawisko rezonansu magnetycznego można ściśle opisać tylko na gruncie mechaniki kwantowej. Spróbujemy tu mimo to podać pewne analogie klasyczne, przemawiające do wyobraźni i ułatwiające zrozumienie istoty procesu. Za rezonansowe pochłanianie energii odpowiedzialne są cząstki obdarzone momentem magnetycznym. Mogą to być jądra atomowe, jony lub inne mikroobiekty. Skąd się bierze moment magnetyczny jakiejś cząstki? Wyobraźmy sobie naładowaną cząstkę krążącą po zamkniętym torze. Ruch tej cząstki jest równoważny z przepływem prądu elektrycznego w zamkniętej pętli. Jak wiemy, jednym ze skutków przepływu prądu elektrycznego jest wytworzenie pola magnetycznego. Cząstka krążąca po zamkniętej orbicie jest więc czymś w rodzaju elementarnego magnesu. W mechanice kwantowej nie wyobrażamy sobie, że cząstki krążą po ustalonych orbitach lub że wirują wokół jakiejś osi, lecz mówimy, że są obdarzone pewnym momentem pędu. Teraz już możemy odpowiedzieć na postawione pytanie. Moment magnetyczny jest związany z ładunkiem elektrycznym cząstki i z jej momentem pędu. Oznaczmy wektor momentu magnetycznego naszej cząstki przez $\vec{\mu}$.

moment magnetyczny

W zewnętrznym, stałym polu magnetycznym wektor $\vec{\mu}$ wykonuje precesję wokół wektora $\vec{B}$ — indukcji pola magnetycznego. Jest to zjawisko analogiczne do krążenia (precesji) osi bąka-zabawki w polu grawitacyjnym (rys. 1). Istotna różnica pomiędzy tymi zjawiskami polega na tym, że oś bąka może być dowolnie nachylona w stosunku do pionu, a położenie wektora $\vec{\mu}$ nie może być dowolne. Wektor $\vec{\mu}$ musi wykonywać precesję w taki sposób, aby rzut momentu pędu na kierunek wektora $\vec{B}$ przybierał jedną z $2J+1$ wartości:

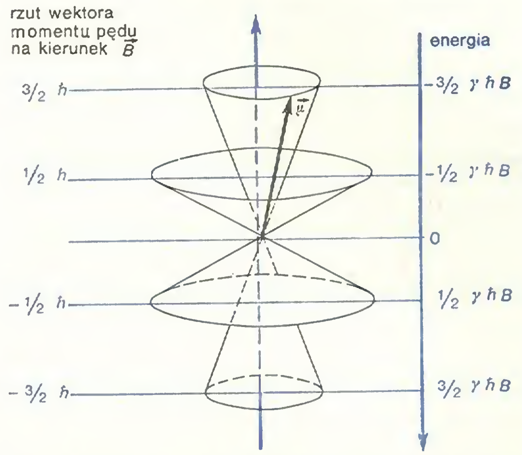
precesja wektora momentu magnetycznego



Rys. 1. Porównanie precesji wektora momentu magnetycznego w jednorodnym polu magnetycznym i precesji bąka-zabawki w ziemskim polu grawitacyjnym

$$-J\hbar,\ (-J+1)\hbar\ \ldots\ (J-1)\hbar,\ J\hbar;$$

J jest tu tzw. liczbą kwantową momentu pędu, a $\hbar$ — stałą Plancka dzieloną przez 2π. Liczba J może być całkowita lub połówkowa (0, $^1/_2$, 1, $^3/_2$, 2...). Na przykład, jeśli cząstka ma moment pędu równy $^3/_2\hbar$, $J = {^3/_2}$ (nie jest to ścisłe powiedzenie,



Rys. 2. Możliwe ustawienia wektora momentu magnetycznego scharakteryzowanego liczbą kwantową momentu pędu $J = {^3/_2}$ w polu magnetycznym i odpowiadające im energie

gdyż całkowity moment pędu cząstki opisanej liczbą kwantową J jest równy $\hbar\sqrt{J(J+1)}$), to możliwe są cztery ustawienia wektora w polu magnetycznym (rys. 2). Każdemu ustawieniu odpowiada inna energia. Jak wiadomo, energia momentu magnetycznego w polu magnetycznym wyraża się wzorem

energia momentu magnetycznego

$$E = -\vec{\mu}\vec{B} = -\mu B\cos\theta;$$

θ oznacza tu kąt pomiędzy wektorami $\vec{\mu}$ i $\vec{B}$, czyli połowę kąta rozwarcia stożka, po którym się porusza wektor momentu magnetycznego $\vec{\mu}$. W mechanice kwantowej oblicza się, że moment magnetyczny opisany liczbą kwantową J ma w polu magnetycznym energię

$$E_J = \gamma J\hbar B. \qquad (1)$$

Współczynnik γ (tzw. stosunek magnetomechaniczny lub giromagnetyczny) ma różne wartości — w zależności od tego, jaką cząstkę rozpatrujemy. Jeśli np. elektron walencyjny, to $\gamma = \gamma_e = g\mu_B$, jeśli zaś jądro atomowe, to $\gamma = \gamma_N = g_N\mu_N$. Magneton Bohra μ_B i magneton jądrowy μ_N są elementarnymi jednostkami momentu magnetycznego odpowiednio dla elektronu i jądra atomowego. Bezwymiarowy współczynnik g nazywa się współczynnikiem rozszczepienia spektroskopowego, czynnikiem Landego lub po prostu czynnikiem g. Jego wartość możemy obliczać teoretycznie lub otrzymywać z eksperymentu. Z relatywistycznej teorii Diraca wynika, że g swobodnego elektronu w próżni równa się 2,0023.

Z równania (1) widać, że różnica energii dwóch położeń wektora $\vec{\mu}$ (czyli dwóch stanów energetycznych), różniących się liczbą kwantową J o jedność, wynosi:

$$\Delta E_{J,J-1} = E_J - E_{J-1} = \gamma\hbar B.$$

Jeśli energia padającego fotonu $E = h\nu = \hbar\omega$ jest równa ΔE, to jest spełniony warunek Bohra i foton może być pochłonięty. Jest to podstawowy warunek rezonansu. Zapiszemy go krótko jeszcze raz:

podstawowy warunek rezonansu

$$\Delta E = \hbar\omega_0 = \gamma\hbar B, \text{ czyli } \omega_0 = \gamma B. \qquad (2)$$

Podkreślmy przy tym, że B oznacza indukcję pola magnetycznego, które działa na nasz moment magnetyczny. Nie zawsze jest to takie pole, jakie przykładamy do próbki z zewnątrz. Może się przecież zdarzyć, że w naszej próbce istnieją poza polem zewnętrznym jakieś wewnętrzne pola magnetyczne. Musimy pamiętać, że B w naszym podstawowym wzorze (2) oznacza wypadkową wszystkich tych pól.

drugi warunek rezonansu

Jest jeszcze drugi warunek rezonansu, łatwo zrozumiały z klasycznego punktu widzenia: składowa magnetyczna $\vec{b}$ padającej fali elektromagnetycznej powinna być prostopadła do wektora $\vec{B}$. Jeśli kierunek wirowania wektora $\vec{b}$ będzie taki sam jak kierunek precesji wektora $\vec{\mu}$ (rys. 3), to można będzie przyspieszyć precesję wektora $\vec{\mu}$, układ pochłonie energię. Jest więc rzeczą konieczną, aby padająca fala elektromagnetyczna miała składową magnetyczną prostopadłą do pola stałego.

równania ruchu momentu magnetycznego

O tym, jak się zachowuje wektor momentu magnetycznego w zewnętrznym polu magnetycznym, możemy się dowiedzieć rozwiązując jego równania ruchu. Nie będziemy tu tego robić, napiszemy tylko te równania:

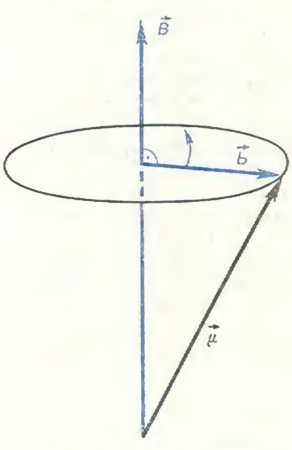
$$\frac{d}{dt}\vec{\mu} = \gamma\vec{\mu}\times\vec{B}. \qquad (3)$$

Są to właściwie trzy równania, ponieważ niewiadomymi są tu trzy składowe wektora $\vec{\mu}$ jako funkcje czasu. Zapisaliśmy je dla skrócenia w formie wektorowej. Jest to dobrze znane równanie mechaniki klasycznej (pochodna momentu pędu układu jest równa momentowi sił zewnętrznych działających na układ), z tą jedynie różnicą, że obie strony równania pomnożono przez γ i otrzymano z lewej strony moment magnetyczny zamiast momentu pędu. Można się przekonać, że równanie (3) opisuje precesję wektora $\vec{\mu}$ wokół wektora $\vec{B}$. Pamiętamy jednak, że każdy rzeczywisty układ fizyczny dąży do takiego stanu, w którym jego energia byłaby możliwie mała. W danym wypadku oznaczałoby to, że wektor $\vec{\mu}$ powinien się w końcu ustawić równolegle do wektora $\vec{B}$. Zachowanie jego winno być mniej więcej takie jak na rys. 4, tzn. powinien on krążyć po okręgu o coraz mniejszym promieniu. Przebieg czasowy zjawiska jest więc następujący: 1) moment magnetyczny jest ustawiony równolegle do pola magnetycznego, 2) po pochłonięciu porcji energii występuje precesja, 3) energia jest wytracana (mówimy, że następuje relaksacja) w pewnym charakterystycznym czasie, zwanym czasem relaksacji. W obrazie kwantowym możemy to sobie wyobrazić jako przeskoki wektora $\vec{\mu}$ z jednego stożka precesji na drugi. Klasyczne równania ruchu, takie jak równania (3), ale zawierające człony opisujące relaksację, nazywają się równaniami ruchu Blocha. Podobne równania ruchu podali niezależnie F. Gilbert i L.D. Landau. Jeśli na podstawie równań ruchu Blocha obliczymy szybkość pochłaniania energii przez układ, otrzymamy następujące wyrażenie zależne od częstości padającej fali elektromagnetycznej:

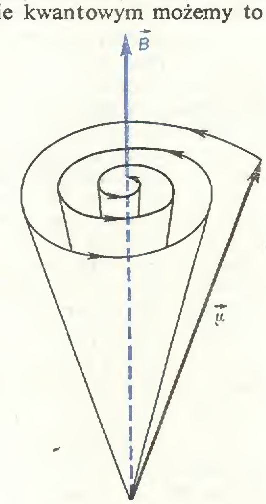
krzywa Lorentza

$$P(\omega) = \frac{\omega\gamma\mu\tau b^2}{1+(\omega-\omega_0)^2\tau^2}, \qquad (4)$$

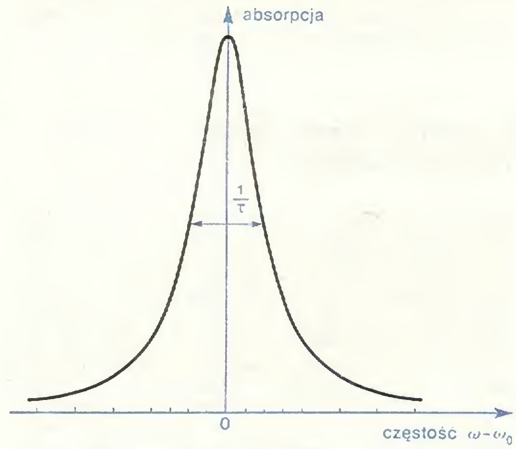
gdzie b jest wielkością składowej magnetycznej padającej fali elektromagnetycznej, która wywołuje rezonans. Jeśli wykreślimy tę krzywą, otrzymamy obraz taki jak na rys. 5. Jest to dobrze znana w fizyce krzywa Lorentza. Jej szerokość na połowie wysokości równa się $1/\tau$; ω_0 jest określone przez warunek rezonansu (2).



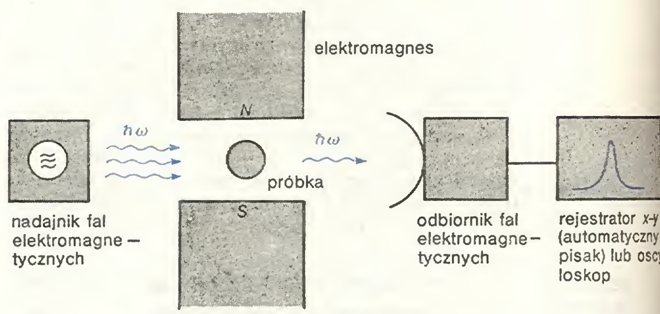
Rys. 3. Wzajemne ustawienie wektora pola magnetycznego, składowej magnetycznej padającej fali elektromagnetycznej wywołującej rezonans i wektora momentu magnet. cząstki



Rys. 4. Klasyczny obraz precesji wektora momentu magnetycznego cząstki, która stopniowo traci energię



Rys. 5. Krzywa absorpcyjna rezonansu magnetycznego



Rys. 6. Schemat aparatury do obserwacji rezonansu magnetycznego

układ eksperymentalny

Schemat typowego układu do eksperymentalnego badania rezonansów magnetycznych przedstawiony jest na rys. 6. Próbka znajduje się w polu magnetycznym wytwarzanym przez elektromagnes lub magnes nadprzewodnikowy. Fale elektromagnetyczne wysyłane z nadajnika docierają do próbki z lewej strony, następnie są pochłaniane, a ta część energii, która nie została zaabsorbowana przez próbkę, dociera do odbiornika. Z odbiornikiem połączony jest oscyloskop lub rejestrator, który pozwala na zapis otrzymywanych krzywych rezonansowych. Zwykle w eksperymencie łatwiej jest zmieniać stałe pole magnetyczne niż częstość padających fal elektromagnetycznych. Otrzymujemy wtedy na rejestratorze zapis krzywej pochłaniania w funkcji pola magnetycznego przy ustalonej częstości fal elektromagnetycznych. Często się zdarza, że próbka zawiera niewiele momentów magnetycznych mogących absorbować energię — stosujemy wówczas skomplikowane metody nadawania

i odbioru fal elektromagnetycznych, dzięki którym możemy rejestrować nawet bardzo słabe sygnały od próbki. Powszechnie stosowane są takie metody spektroskopii rezonansów magnetycznych, które pozwalają obserwować sygnały 100 tysięcy razy słabsze od szumów. Nasza aparatura, zwana spektrometrem, daje wtedy na wyjściu zapis nie krzywej rezonansowej, lecz jej pochodnej. Idea pomiaru pozostaje jednak taka sama.

W pierwszych pracach nad rezonansami magnetycznymi używano jako odbiornika kalorymetru (C. Gorter 1948). Pomiar trwał długo i był obarczony dużymi błędami. Trzeba było przy rozmaitych częstościach fal elektromagnetycznych (lub wartościach pól) mierzyć temperaturę i w ten sposób punkt po punkcie określać kształt krzywej absorpcyjnej. Współczesne spektrometry pozwalają na prawie automatyczne zapisywanie całej krzywej absorpcyjnej, a sam pomiar trwa średnio kilka minut. Obecnie wobec postępu w metodach fizyki niskich temperatur powraca się do obserwacji rezonansów magnetycznych przez pomiar temperatury próbki. Jest to jednak możliwe tylko w bardzo niskich temperaturach, poniżej 4 K.

Rezonans jądrowy

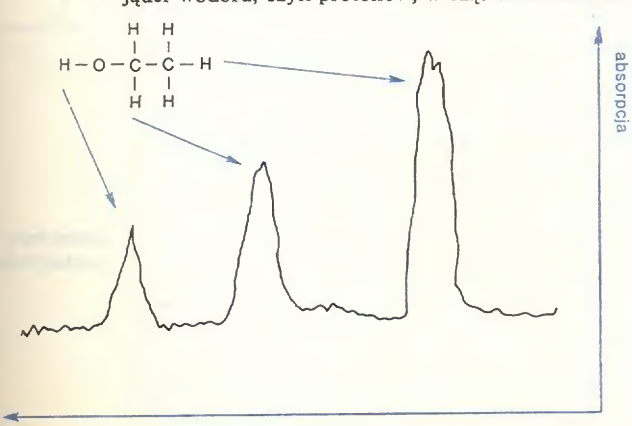
Wszystkie nasze dotychczasowe rozważania pozostaną w mocy, jeżeli zamiast liczby kwantowej J wpiszemy do wzoru (1) liczbę I — spin obserwowanego jądra atomowego. Częstość fal elektromagnetycznych potrzebnych do wywołania rezonansu zależy oczywiście od przyłożonego pola magnetycznego (wzór 2). Częstości rezonansowe niektórych jąder w polu magnetycznym 1 T podane są w tabeli.

Częstości rezonansu jądrowego w polu magnetycznym 1T i wartość spinu jądrowego wybranych jąder

Jądro	Częstość rezonansowa w MHz	Spin
^{1}H (proton)	42,576	1/2
^{2}H (deuter)	6,536	1
^{7}Li	16,55	3/2
^{13}C	10,71	1/2
^{19}F	40,06	1/2
^{23}Na	11,26	3/2
^{127}I	8,52	5/2

Podane częstości pokrywają zakres UKF i fal krótkich stosowanych w radiofonii. Dotyczy to oczywiście jąder swobodnych. Jeśli badane jądra znajdują się w cząsteczkach, efekt może być na przykład taki jak na rys. 7. Jest to widmo rezonansu jądrowego (NMR) jąder wodoru, czyli protonów, w cząsteczce alkoholu etylowego. Zmianę częstości rezonansowej wywołaną ekranującym działaniem chmury elektronowej nazywamy przesunięciem chemicznym. Na rys. 7 widać wyraźnie, że atomy wodoru zawarte w cząsteczce etanolu można podzielić na trzy grupy. Porównanie wzoru strukturalnego etanolu i widma pozwala w tym wypadku łatwo określić, które jądra wodoru w cząsteczce alkoholu są odpowiedzialne za kolejne maksima pochłaniania. Różne częstości rezonansowe tłumaczą się oczywiście tym, że lokalne pole magnetyczne działające na protony jest inne w każdej grupie i nieco inne niż przyłożone pole zewnętrzne. Takimi metodami rozszyfrowuje się strukturę wielu związków organicznych.

przesunięcie chemiczne



Rys. 7. Widmo NMR protonów w cząsteczce alkoholu etylowego

Zastosowanie zjawiska NMR

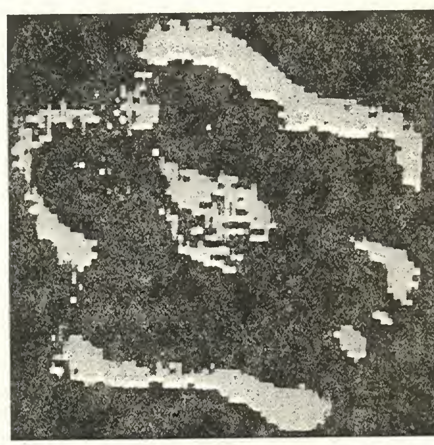
Badanie rezonansu magnetycznego protonów jest szczególnie ważne w chemii organicznej, ponieważ wiele związków organicznych zawiera w swym składzie wodór. W ostatnich latach rozwinęła się również spektroskopia rezonansu jądrowego izotopu węgla ^{13}C, pierwiastka zawartego w związkach organicznych. Badania tego rezonansu są znacznie trudniejsze niż spektroskopia protonów, gdyż izotop ^{13}C stanowi tylko ok. 1,1% wszystkich atomów węgla w przyrodzie. Z kolei spektroskopia rezonansu jądrowego izotopu fluoru ^{19}F, który stanowi prawie 100% fluoru w przyrodzie, jest utrudniona z powodu dosyć wysokiej częstości rezonansowej i toksyczności jego związków. W fizyce ciała stałego bada się NMR wielu innych izotopów, np. glinu ^{27}Al, żelaza ^{57}Fe i kobaltu ^{59}Co.

NMR protonów w wodzie

Zjawisko NMR protonów zawartych w wodzie zostało wykorzystane do precyzyjnego pomiaru pola magnetycznego. Czynnik γ protonów w wodzie jest znany z bardzo wysoką dokładnością. Pozwala to określić wartość indukcji pola magnetycznego z częstości rezonansowej NMR. Pomiar pola tą metodą jest łatwy w zakresie pól 0,1–1,5 T. Do pomiaru większych pól używa się jąder litu ^{7}Li.

miernik wilgotności

Zjawisko NMR protonów w wodzie wykorzystano także w mierniku wilgotności materiałów sypkich. Miernik taki pozwala szybko i nie niszcząc materiału określić z dokładnością ok. 2% zawartość wody w trocinach, piasku, ziarnie i innych materiałach. Pomiar może być prowadzony w sposób ciągły, np. na przesuwającej się taśmie w magazynie, zakładzie produkcyjnym, w porcie.



Rys. 8. Rozkład gęstości protonów w przekroju strąka rośliny afrykańskiej okra otrzymany metodą NMR. Średnica strąka ok. 15 mm

Metodą NMR protonów w wodzie można przeprowadzać ciekawe badania w obiektach biologicznych. Jeżeli obszar pola magnetycznego o takim rozkładzie przestrzennym indukcji, przy którym warunki rezo-



Rys. 9. Rozkład protonów w płaszczyźnie przechodzącej przez żołądek szczura otrzymany metodą NMR

badania obiektów biologicznych

nansowe będą spełnione tylko w małym fragmencie próbki, będziemy przesuwać, to w rezultacie można będzie otrzymać obrazy podobne do rys. 8 i 9. Zarejestrowano oczywiście wielkość absorpcji, podczas gdy pole i częstość rezonansowa były ustalone. Tą metodą uzyskano już obrazy tkanek różnych roślin, pierwsze obrazy żywych małży w muszlach i obrazy dłoni ludzkiej.

Metodą NMR badano też zanieczyszczoną wodę morską. Wynikiem tych badań jest wiedza na temat procesów prowadzących do samooczyszczenia wody morskiej z rozlanych olejów i ropy naftowej oraz wpływu na te procesy takich czynników jak nasłonecznienie i falowanie morza.

NMR w ferromagnetykach

W pewnych wypadkach można obserwować NMR bez przykładania zewnętrznego pola magnetycznego. Dzieje się tak w ferromagnetykach, a to dzięki temu, że wewnątrz nich istnieje już pole magnetyczne. Szczególnie interesujące są badania NMR jąder znajdujących się w ściankach domenowych, tj. obszarach rozdzielających leżące blisko siebie obszary ferromagnetyka, tzw. domeny magnetyczne, o różnych kierunkach spontanicznego namagnesowania (→ Struktura domenowa i procesy magnesowania). Są to badania bardzo trudne do interpretacji, ponieważ w ściance domenowej wewnętrzne pole magnetyczne zmienia swój kierunek od jądra do jądra.

Rezonans elektronowy

Rezonans elektronowy (EPR) obserwowano w gazach, cieczach, szkłach, półprzewodnikach i metalach. Ważną klasą obiektów badanych metodą rezonansu elektronowego są wolne rodniki.

Istota zjawiska jest taka sama jak rezonansu jądrowego. Podstawową różnicę stanowi pochodzenie wektora momentu magnetycznego. W EPR obserwujemy rezonansowe pochłanianie energii przez układ elektronów, a nie jąder atomowych. Poza tym całkowity moment pędu elektronu składa się z momentu orbitalnego, wynikającego z ruchu elektronu wokół jądra, oraz z momentu spinowego. Do opisu rezonansu elektronowego używamy innej liczby kwantowej, tzw. spinu efektywnego S, co oczywiście powoduje pewne komplikacje. Ponieważ magneton Bohra jest ok. 2000 razy większy niż magneton jądrowy, to i stosowane częstości fal elektromagnetycznych muszą być o trzy rzędy większe niż przy rezonansie jądrowym.

częstość promieniowania w EPR

Zwykle w eksperymentach EPR częstość promieniowania jest ustalona i wynosi ok. 10 GHz (10^{10} Hz), a typowy zakres pól magnetycznych wynosi 0–1 T. Stosowane częstości mikrofalowe zmuszają do używania podzespołów elektronicznych typowych dla tego zakresu fal: falowodów, cyrkulatorów, rezonatorów, tłumików i specjalnych mikrofalowych diod detekcyjnych oraz generatorów mikrofal. Do generacji mikrofal używa się lamp zwanych klistronami lub (w najnowszych konstrukcjach) generatorów półprzewodnikowych — diod Gunna (→ Generacja mikrofal).

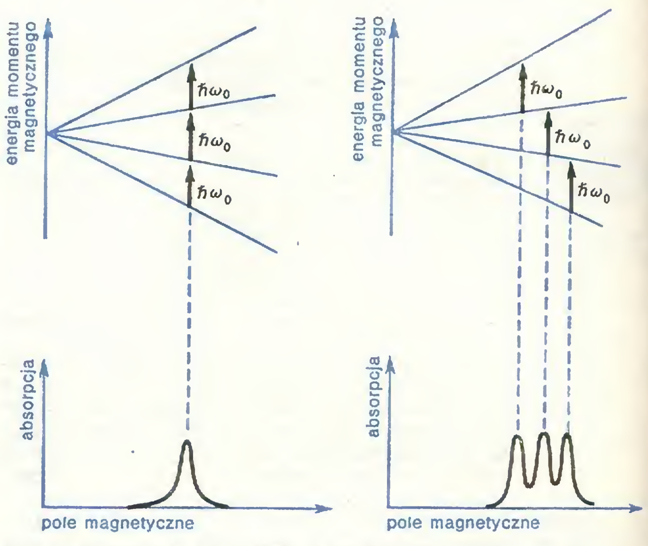
Czułość obecnie używanych spektrometrów EPR jest bardzo wysoka, pozwala wykryć w temperaturze pokojowej $5 \cdot 10^{10}$ cząstek paramagnetycznych w próbce.

Struktura subtelna i nadsubtelna widm EPR

Jeśli liczba kwantowa $S > {}^1/_2$, to poziomy energetyczne elektronu nie są równoodległe. Widmo rezonansu paramagnetycznego nie składa się wówczas z jednej linii, lecz z kilku. Zjawisko polegające na występowaniu kilku linii rezonansowych zamiast jednej nazywamy strukturą subtelną widma EPR. Powinno być $2S$ obserwowanych linii, ponieważ przejścia pomiędzy poziomami, które nie są kolejne, są bardzo mało prawdopodobne. Przykład widma ze strukturą subtelną pokazany jest na rys. 10.

struktura subtelna

W monokryształach wygląd widma zależy ponadto od orientacji pola magnetycznego względem osi krystalograficznych. To zjawisko nosi nazwę anizotropii widma.

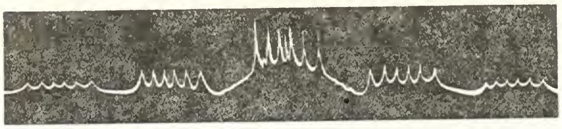


Rys. 10. Powstawanie widma EPR: a) bez struktury subtelnej, b) z rozdzielonymi pikami struktury subtelnej

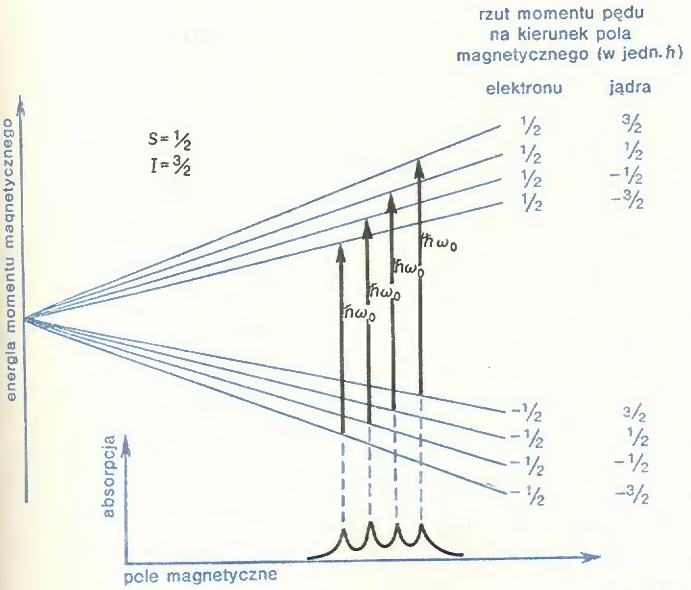
Często się zdarza, że jądro atomowe, wokół którego krąży badany elektron, ma różny od zera spin jądrowy. Energia elektronu zależy wtedy od tego, w jakim stanie energetycznym znajduje się aktualnie moment magnetyczny jądra atomowego. Jest to zupełnie zrozumiałe, bo momenty magnetyczne jądra i elektronu mogą oddziaływać ze sobą zupełnie tak samo jak małe magnesy. W takim wypadku liczba dozwolonych poziomów energetycznych elektronu wynosi $(2S+1) \cdot (2I+1)$, co oczywiście wpływa na wygląd widma; obserwujemy większą liczbę linii rezonansowych. Taką strukturę widma nazywamy nadsubtelną. Przykład widma z dobrze widoczną strukturą nadsubtelną (i jednocześnie subtelną) pokazany jest na rys. 11.

struktura nadsubtelna

Rysunek 12 wyjaśnia sposób powstawania takiego widma. Z wzajemnej odległości linii, które są składnikami struktury nadsubtelnej, można wnioskować



Rys. 11. Widmo EPR jonów Mn^{2+} w apatycie. Widoczne są linie należące do struktury nadsubtelnej. Wąska linia w środku rysunku pochodzi od substancji wzorcowej (znacznika pola magnetycznego)



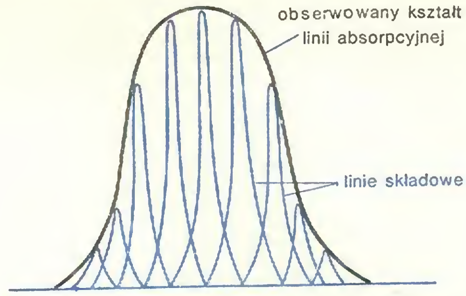
Rys. 12. Sposób powstawania widma ze strukturą nadsubtelną

o sile oddziaływań pomiędzy momentem magnetycznym jądra i elektronu. Jeśli oddziaływanie będzie słabe, to ruch elektronu będzie tylko nieznacznie zaburzony i składowe struktury nadsubtelnej widma będą leżeć blisko siebie. Może się zatem zdarzyć, że obserwacja pojedynczych linii rezonansowych będzie niemożliwa, a o istnieniu oddziaływań nadsubtelnych będzie świadczyć tylko zniekształcona forma linii.

Kształt i szerokość linii NMR i EPR

Szerokość linii absorpcyjnej dla pojedynczego momentu magnetycznego zależy przede wszystkim od procesów relaksacji. Obserwowany czas relaksacji w różnych materiałach waha się od kilku mikrosekund do kilku minut — w zależności od temperatury i rodzaju próbki. Pamiętajmy jednak, że w rzeczywistej próbce znajduje się bardzo wiele momentów magnetycznych, których rezonans obserwujemy równocześnie. Może się zdarzyć, że w próbce będzie kilka rodzajów momentów magnetycznych, rozmaicie zorientowanych, oddziałujących ze sobą i z otoczeniem.

Szczególnie skomplikowana sytuacja występuje w kryształach. Najczęściej współczynnik g zależy od orientacji pola magnetycznego względem osi krystalograficznych; mówimy, że w kryształach g jest wielkością tensorową. Centra magnetyczne w krysztale mogą mieć rozmaite otoczenia, czyli mogą się znajdować w nieco różnych warunkach, np. wskutek istnienia defektów sieci krystalicznej, lokalnych naprężeń. Z tego powodu każdy moment magnetyczny ma nieco inną częstość rezonansową. Wtedy zamiast „czystych" linii rezonansowych obserwujemy linie poszerzone, które są sumą wielu linii rezonansowych leżących blisko siebie. Ilustruje to rys. 13. Ten typ poszerzenia krzywych absorpcyjnych nazywa się po-



Rys. 13. Linia rezonansowa poszerzona niejednorodnie

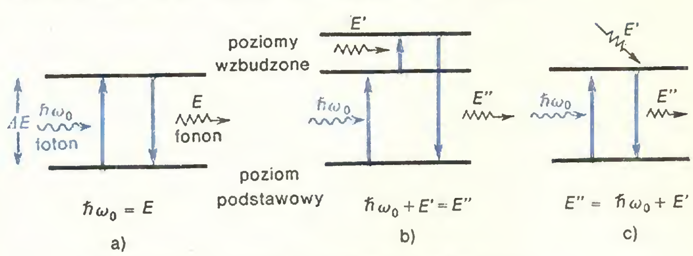
poszerzenie linii absorpcyjnej

szerzeniem niejednorodnym (niehomogenicznym). Identyczny efekt może się pojawić przy niewłaściwym działaniu spektrometru. Zarejestrujemy linie poszerzone, jeśli pole magnetyczne w obszarze próbki będzie niejednorodne lub gdy nadajnik będzie wytwarzał jednocześnie fale o różnych częstościach. Kształt linii poszerzonej niejednorodnie nie jest lorentzowski, lecz np. gaussowski lub zupełnie nieregularny, jak w wypadku proszków otrzymanych z monokryształów.

Szerokość naturalna linii absorpcyjnych rezonansu jądrowego jest na ogół bardzo mała, rzędu ułamka Hz. Zwykle jednak obserwowana szerokość linii, zwłaszcza w ciałach stałych, jest dużo większa. Odpowiedzialne są za to przede wszystkim długozasięgowe oddziaływania dipol–dipol. Inne rodzaje oddziaływań, np. elektrostatyczne lub z drganiami sieci krystalicznej, są silnie wytłumione (ekranowane) przez elektrony danego jonu.

Z inną sytuacją spotykamy się w badaniach EPR. Tu szerokość linii, a więc i czas relaksacji są określone głównie przez oddziaływania z otoczeniem. W kryształach relaksacja zachodzi przy udziale drgań sieci — fononów. Znamy trzy podstawowe mechanizmy relaksacji spin-sieć. Są to: relaksacja Orbacha, relaksacja Ramana i proces prosty. Schematyczny przebieg tych trzech procesów relaksacyjnych ukazuje rys. 14. Jeśli w procesie relaksacji prostej (rys. 14a) zamiast fononu wyemitowany zostanie foton, który z kolei może być pochłonięty przez inny jon, wówczas powiemy o relaksacji spin-spin. W eksperymencie można

relaksacja spin-sieć



Rys. 14. Schematyczny przegląd podstawowych mechanizmów relaksacji spin-sieć: a) relaksacja prosta, b) relaksacja Orbacha, c) relaksacja Ramana

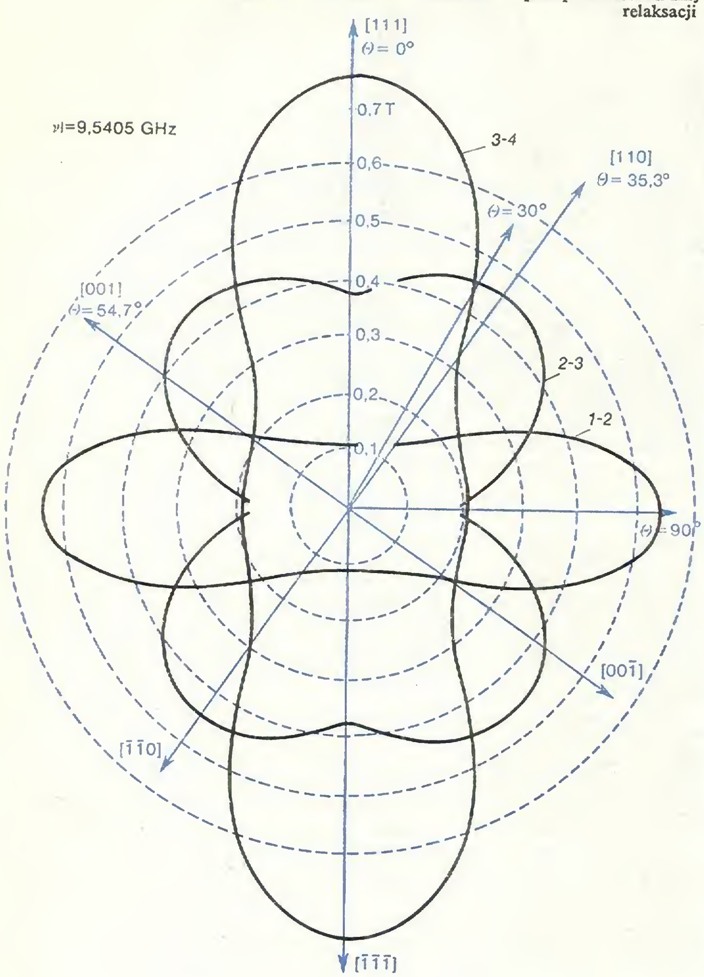
dość łatwo odróżnić wpływ różnych procesów relaksacyjnych na szerokości linii. Czas relaksacji spin-spin praktycznie nie zależy od temperatury, natomiast prawdopodobieństwo relaksacji prostej jest proporcjonalne do temperatury próbki. Prawdopodobieństwo relaksacji typu orbachowskiego zależy od temperatury jak $e^{-\Delta E/kT}$, a relaksacji typu ramanowskiego — jak T^7 lub T^9. Szerokość linii jest proporcjonalna do prawdopodobieństwa relaksacji; wynika to z relacji nieokreśloności: $\Delta E \cdot \tau \approx \hbar$, czyli $\hbar \cdot \Delta\omega \cdot \tau \approx \hbar$. W tym wzorze τ oznacza średni czas relaksacji, uwzględniający wszystkie możliwe mechanizmy relaksacji. Obliczamy go ze wzoru:

czas relaksacji

$$1/\tau = 1/\tau_{\text{relaksacja prosta}} + 1/\tau_{\text{relaksacja Orbacha}} + 1/\tau_{\text{relaksacja Ramana}} + 1/\tau_{\text{relaksacja spin-spin}} + 1/\tau_{\text{inne}},$$

czyli

$$\Delta\omega = \Delta\omega_{\text{procesy proste}} + \Delta\omega_{\text{relaksacja Orbacha}} + \Delta\omega_{\text{relaksacja Ramana}} + \Delta\omega_{\text{relaksacja spin-spin}} + \Delta\omega_{\text{inne mechanizmy relaksacji}}.$$



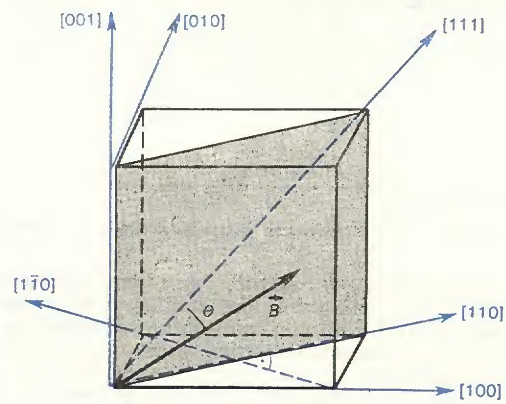
Wróćmy jeszcze na chwilę do wzoru (4) określającego kształt linii pochłaniania. Jest on słuszny, gdy moc padającej fali elektromagnetycznej nie jest zbyt duża. Przy dużej mocy padającego promieniowania obserwuje się tzw. nasycenie rezonansu. Jest to jeszcze jeden z przejawów kwantowego charakteru zjawiska, który krótko opiszemy.

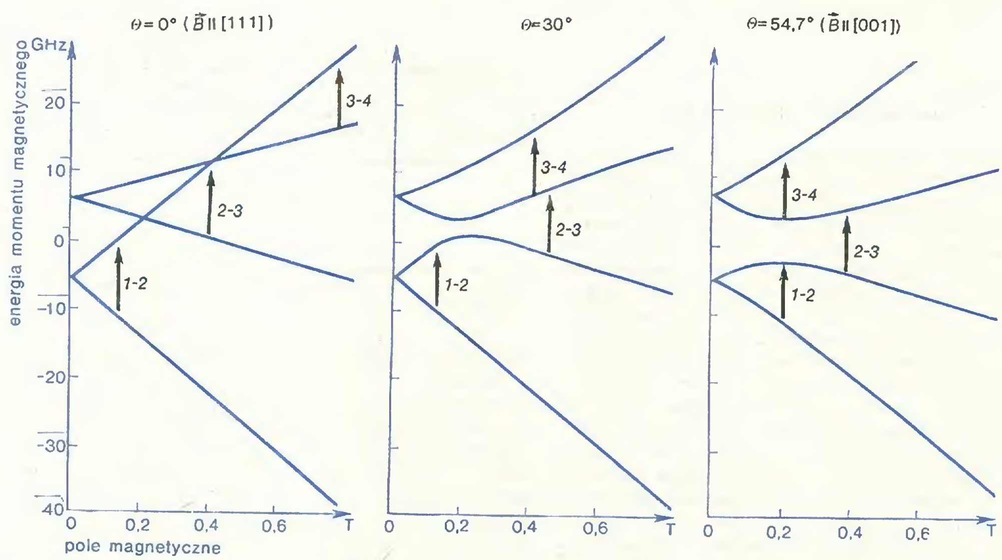
nasycenie absorpcji rezonansowej

Początkowo wiele fotonów zostaje zaabsorbowanych i prawie wszystkie momenty magnetyczne próbki są przeniesione na wyższy poziom energetyczny, z którego nie zdążają powrócić do stanu podstawowego. Podstawowy poziom energetyczny jest więc prawie pusty i nie ma momentów magnetycznych, które by mogły absorbować następne padające fotony. Również elektrony, jony, kompleksy jonów i inne mikroobiekty, które są na wyższym poziomie energetycznym nie mogą pochłaniać energii, gdyż następne dozwolone poziomy energii leżą zbyt wysoko. Zjawisko nasycania absorpcji rezonansowej utrudnia w pewnym stopniu badania, gdyż nie pozwala na użycie fal elektromagnetycznych o zbyt dużej intensywności; stąd konieczność używania spektrometrów o bardzo wysokiej czułości.

zakorkowanie fononowe

Wspomnimy jeszcze o dającym podobne efekty zjawisku „zakorkowania fononowego" (ang. *phonon bottle neck*) lub „wąskiego gardła fononowego", obserwowanym w EPR. W bardzo niskiej temperaturze (poniżej 4 K) relaksacja może zachodzić jedynie przez emisję fononu lub fotonu. Wszystkie emitowane fonony mają tę samą energię równą $\hbar\omega_0$. Oznacza to,





Rys. 15. Pole rezonansowe jonu chromu Cr^{3+} w sieci krystalicznej rubinu. Z prawej strony rysunku — płaszczyzna krystalograficzna, w której przemieszczano stałe pole magnetyczne. Na dole — zależność dozwolonych wartości energii od orientacji zewnętrznego pola magnetycznego dla trzech różnych orientacji. Liczby *1–2* itd. oznaczają numery poziomów energetycznych, między którymi zachodzi przejście rezonansowe

że wzbudzany jest tylko jeden rodzaj drgań sieci krystalicznej. Zamiana takiego uporządkowanego drgania sieci na drgania chaotyczne (cieplne) trwa dość długo, niekiedy czas relaksacji jest rzędu kilku minut. W niskiej temperaturze kryształ jest „spokojny", jego sieć krystaliczna praktycznie nie drga. Istnieje tylko jedna możliwość odprowadzenia nagromadzonego ciepła: przez powierzchnię kryształu (rozpad drgań uporządkowanych na drgania termiczne może też zachodzić na defektach sieci). Ilość drgań określonego rodzaju jest jednak ograniczona, dlatego obserwujemy taki sam efekt jak przy nasycaniu rezonansu zbyt dużą mocą padającego promieniowania.

Zastosowania EPR

Wiele zastosowań EPR wiąże się z fizyką kryształów. Przypuśćmy, że dysponujemy niemagnetycznym kryształem, który jest domieszkowany jonami magnetycznymi. Mogą to być jony tytanu, wanadu, chromu, manganu, żelaza, kobaltu, niklu, miedzi lub ziem rzadkich. Jony takie zachowują się jak sondy, które pozwalają określać parametry elektrostatycznego pola krystalicznego w tym miejscu sieci krystalicznej, w którym się znajduje jon, oraz badać możliwe drgania sieci. Wyjaśnimy to na przykładzie jonów Cr^{3+} w sieci krystalicznej rubinu.

Energia momentu magnetycznego jonu chromu jest określona nie tylko przez zewnętrzne pole magnetyczne, ale i przez pole krystaliczne. W konsekwencji dozwolone wartości energii jonu Cr^{3+} zależą od wielkości pola magnetycznego i od orientacji tego pola względem osi krystalograficznych. Wyjaśnia to bliżej rys. 15; przedstawiono na nim widmo jonu Cr^{3+} w Al_2O_3 (rubinie) we współrzędnych biegunowych. Aby otrzymać taki rysunek, należy wielokrotnie powtórzyć eksperyment badając widmo EPR w różnych kierunkach w krysztale. Rysunek ten ukazuje, w jakiej płaszczyźnie krystalograficznej przemieszczane było pole magnetyczne podczas eksperymentu. Z rys. 15 widać ponadto, jakie elementy symetrii ma w sieci krystalicznej to miejsce, które zajmuje jon chromu. Analizując tego typu wyniki eksperymentalne, można określać położenia wbudowanych w sieć krystaliczną jonów domieszkowych. Jeśli jony domieszkowe wbudowują się w różne położenia w krysztale, to na podstawie kilku nałożonych widm EPR można próbować określić symetrię kryształu jako całości. Jest to więc metoda uzupełniająca w stosunku do rentgenowskich badań strukturalnych.

określenie położeń jonów domieszkowych

określanie symetrii kryształu

Spektroskopia EPR znalazła jeszcze wiele innych zastosowań. Podjęto np. udane próby określania zawartości miedzi w rudach. Dokładność pomiaru wynosi ok. 5%. Jonami, których rezonans badano, były jony Cu^{++}.

Badając widmo EPR w różnych warunkach, można wyciągać wnioski o defektach struktury krystalicznej (dyslokacje, luki, atomy w pozycjach międzywęzłowych, struktura mozaikowa, uszkodzenia radiacyjne i in.), których rozmiary są zbyt małe, aby mogły być wykryte defektoskopem ultradźwiękowym. Na wygląd widma mają również wpływ naprężenia wewnątrz kryształu. Można więc śmiało powiedzieć, że spektroskopia EPR bywa pożytecznym narzędziem do badania jakości monokryształów.

badanie defektów struktury

Wygląd widma EPR jonów jakiegoś pierwiastka zależy od wartościowości tego jonu. Fakt ten jest często wykorzystywany przez technologów materiałów monokrystalicznych do weryfikacji hipotez o rozmieszczeniu jonów domieszki.

W ostatnich latach coraz większą karierę robi spektroskopia EPR w badaniach biologicznych i biochemicznych. Chodzi tu głównie o badania krwi (hemoglobina zawiera w swym składzie jony żelaza) oraz rozmaitych reakcji biochemicznych. Te ostatnie badania stawiają szczególnie wysokie wymagania co do stosowanej aparatury. Przy badaniach reakcji biochemicznych, zwłaszcza z udziałem wolnych rodników, zachodzi potrzeba otrzymywania do tysiąca widm w ciągu sekundy. Tylko w ten sposób można prześledzić przebieg reakcji i jej kinetykę. Wydaje się, że badania tego typu mogą być pomocne w zwalczaniu raka, a także niektórych chorób psychicznych.

badanie krwi

badanie reakcji biochemicznych

Praktyczną dziedziną zastosowania badań EPR jest burzliwie rozwijająca się elektronika kwantowa. Wyniki badań rezonansu paramagnetycznego posłużyły konstruktorom laserów i maserów oraz wzmacniaczy o niskim poziomie szumów, używanych w radioastronomii i telekomunikacji satelitarnej.

Rezonans ferromagnetyczny

Jest to pochłanianie energii pola elektromagnetycznego przez próbkę ferromagnetyczną. Istotną różnicą w stosunku do poprzednio opisanych rezonansów jest to, że w rezonansie ferromagnetycznym (FMR) precesję w polu zewnętrznym wykonuje moment magnetyczny całej próbki. Momenty magnetyczne poszczególnych cząstek, z których się składa próbka, oddziałują ze sobą wzajemnie, a skutkiem tego oddziaływania jest równoległe ustawienie wszystkich momentów magnetycznych w próbce. Oddziaływania te nazywamy wymiennymi. Ich natura jest czysto kwantowa, nie mają one odpowiednika klasycznego. Oddziaływania wymienne sprawiają, że próbka jest namagnesowana nawet w nieobecności zewnętrznego pola magnetycznego. Nasz fundamentalny wzór (2) pozostaje słuszny, ale należy w nim jakoś uwzględnić pole magnetyczne istniejące w próbce. Okazuje się, że warunki rezonansu zależą nie tylko od właściwości materiału próbki, ale i od jej kształtu. Dokładne rozwiązanie można uzyskać tylko w odniesieniu do próbek o kształcie elipsoidy, w szczególności kuli. Szczegółowe obliczenia dotyczące tego problemu były wykonane po raz pierwszy przez C. Kittela.

Badanie tego rezonansu ma duże znaczenie w technice: pozwala wyznaczać parametry materiałów używanych do budowy transformatorów i rdzeni do cewek, materiałów służących do produkcji pamięci maszyn cyfrowych oraz materiałów przeznaczonych na trwałe magnesy. Metodami FMR bada się również materiały przeznaczone do pracy w podzespołach elektroniki mikrofalowej i telewizji oraz w projektowanych układach pamięci holograficznej.

zastosowanie FMR

L. A. Kazicyna, N. B. Kupletska *Metody spektroskopowe wyznaczania struktury związków organicznych*, Warszawa 1974; C. Kittel *Wstęp do fizyki ciała stałego*, Warszawa 1976; M. Suffczyński *Prześwietlanie NMR*, Post. Fiz. 29, 245 (1978); R. Wadas *Biomagnetyzm*, Warszawa 1978; *Encyklopedia fizyki*, t. 3, Warszawa 1974.

Zjawisko Mössbauera

Andrzej Hrynkiewicz

Odkrycie w 1957 r. przez R. Mössbauera zjawiska bezodrzutowej emisji i absorpcji promieniowania γ dostarczyło nauce nowej, niezwykle owocnej metody badawczej, zwanej spektroskopią mössbauerowską lub rezonansową spektroskopią promieniowania γ. Dzięki bardzo wysokiej energetycznej zdolności rozdzielczej

spektroskopia mössbauerowska

rezonansowa spektroskopia promieniowania γ pozwoliła wykonać wiele precyzyjnych pomiarów fizycznych o podstawowym znaczeniu poznawczym oraz stała się powszechnie stosowaną metodą w badaniach z dziedziny fizyki ciała stałego, chemii, fizyki jądrowej i biologii. Coraz szersze jest również zastosowanie techniczne zjawiska Mössbauera, czego przykładem może być jego wykorzystanie do kontroli względnej prędkości i regulacji procesu zbliżania się dwóch pojazdów kosmicznych.

wpływ wiązań krystalochemicznych na procesy jądrowe

Pogląd, że wiązanie krystalochemiczne nie ma wpływu na przebieg procesów jądrowych, już dawno okazał się niesłuszny. Co prawda w większości wypadków, badając procesy jądrowe, można atom wchodzący w skład związku chemicznego traktować jako swobodny, gdyż energia wiązania chemicznego jest na ogół zaniedbywalnie mała w porównaniu z energią nawet niskoenergetycznych przejść jądrowych. Jednak wpływ wiązań chemicznych i struktury krystalicznej na niektóre procesy jądrowe dobrze znano przed odkryciem zjawiska Mössbauera. Przykładami mogą być: zaburzenia korelacji kierunkowych promieniowania γ przez wewnątrzkrystaliczne pola magnetyczne lub elektryczne, wpływ struktury ciał stałych na anihilację pozytonów lub wpływ wiązania chemicznego na prawdopodobieństwo wychwytu elektronu przez jądro atomowe. Dopiero jednak odkrycie zjawiska Mössbauera pozwoliło na badanie bezpośredniego wpływu wiązań chemicznych i struktury krystalicznej na energię promieniowania emitowanego w przejściach jądrowych. Typowa zdolność rozdzielcza osiągana w spektroskopii mössbauerowskiej jest rzędu 10^{-12}, co znaczy, że w przypadku fotonów promieniowania γ o energii 100 keV mogą być obserwowane zmiany energetyczne rzędu 10^{-7} eV. Tę wielkość należy porównać z charakterystycznymi energiami drgań sieci krystalicznej wynoszącymi 0,01–0,1 eV.

Aby wyjaśnić, na czym polega zjawisko Mössbauera, należy zacząć od omówienia zjawiska emisji i rezonansowej absorpcji fotonów.

Rezonansowa absorpcja i rozpraszanie promieniowania

Zjawisko rezonansu występuje wówczas, gdy pewnemu układowi dostarczymy dokładnie takiej ilości energii, jaka jest potrzebna, aby ten układ przeszedł z niższego na wyższy poziom energetyczny. Najczęściej mamy do czynienia z przejściem układu z poziomu podstawowego na któryś z jego poziomów wzbudzonych. Gdy dany układ będziemy naświetlać wiązką promieniowania o takiej długości fali, że energia kwantów promieniowania będzie odpowiadać różnicy energii między stanem wzbudzonym i podstawowym, to zaobserwujemy absorpcję promieniowania, której towarzyszy przechodzenie układu do stanu wzbudzonego. Zjawisko to nosi nazwę absorpcji rezonansowej. Wzbudzony układ, wracając do stanu podstawowego emituje promieniowanie o tej samej długości fali, lecz na ogół w innym kierunku niż kierunek padającej wiązki promieniowania, czyli następuje zjawisko rozpraszania, które w tym wypadku nazywamy rozpraszaniem rezonansowym.

Przedstawiony opis zjawisk jest ścisły tylko wówczas, gdy masa rozpatrywanego układu jest nieskończenie wielka w porównaniu z masą odpowiadającą energii absorbowanych i rozproszonych kwantów.

absorpcja i rozproszenie rezonansowe w optyce atomowej

Jeśli chodzi o atomy, to warunek ten jest spełniony w wystarczająco dobrym przybliżeniu, toteż w optyce atomowej absorpcja i rozproszenie rezonansowe mogą być zaobserwowane bez trudności. Łatwo jest wykonać następujące doświadczenie. Z lampy sodowej, w której świeci rozgrzana para sodu, rzucamy wiązkę charakterystycznego dla sodu żółtego światła na szklane naczynie wypełnione również parą sodu w niższej temperaturze. Ponieważ kwanty światła emitowane z lampy przez pobudzane w niej do świecenia atomy posiadają energię właśnie taką, jaka jest potrzebna do wzbudzenia, znajdujących się w stanie podstawowym, atomów sodu w naczyniu, następuje absorpcja rezonansowa podającego światła. Wiązka światła zostaje w zasadzie całkowicie pochłonięta w cienkiej warstwie pary. Pobudzane atomy tej warstwy, wracając do stanu podstawowego, rozpraszają rezonansowo światło, tak że warstwa pochłaniająca pierwotną wiązkę sama staje się źródłem intensywnego żółtego promieniowania.

absorpcja i rozproszenie rezonansowe w fizyce jądrowej

Analogicznego zjawiska można by się spodziewać w fizyce jądrowej. Można by przypuszczać, że np. promieniowanie γ o energii 411 keV emitowane przez rtęć ^{198}Hg będzie silnie absorbowane rezonansowo w cienkiej warstwie rtęci umieszczonej na drodze wiązki promieni γ. Tak jednak zaprojektowane doświadczenie nie daje oczekiwanego wyniku. Aby wytłumaczyć, dlaczego absorpcję rezonansową można z łatwością obserwować w optyce, a analogiczne doświadczenie w fizyce jądrowej nie daje rezultatu, trzeba rozważyć dokładniej procesy emisji i absorpcji promieniowania a w szczególności rozpatrzeć konsekwencje faktu, że energia kwantu promieniowania γ nie może być pominięta w porównaniu z masą emitującego go jądra atomowego.

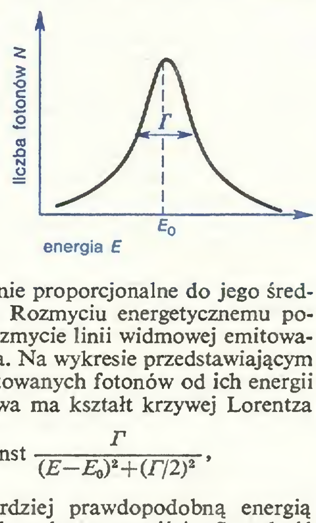
Naturalna szerokość linii widmowej

Fotony promieniowania emitowanego przez zbiór tego samego rodzaju atomów lub jąder nie mają jednakowej, ściśle określonej energii. Jest to konsekwencją zasady nieoznaczoności Heisenberga, która wiąże nieoznaczoność energii ΔE i czasu Δt wzorem:

$$\Delta E \cdot \Delta t = \hbar,$$

rozmycie energetyczne poziomu

gdzie $\hbar = 6{,}6 \cdot 10^{-16}$ eV·s jest stałą Plancka podzieloną przez 2π. Z zasady nieoznaczoności wynika, że rozmycie energetyczne atomowego lub jądrowego



Rys. 1. Naturalny kształt linii widmowej

poziomu jest odwrotnie proporcjonalne do jego średniego czasu życia τ. Rozmyciu energetycznemu poziomu odpowiada rozmycie linii widmowej emitowanego promieniowania. Na wykresie przedstawiającym zależność liczby emitowanych fotonów od ich energii (rys. 1) linia widmowa ma kształt krzywej Lorentza

$$N(E) = \text{const}\,\frac{\Gamma}{(E-E_0)^2+(\Gamma/2)^2},$$

gdzie E_0 jest najbardziej prawdopodobną energią fotonów emitowanych w danym przejściu. Szerokość krzywej Γ w połowie wysokości, charakteryzująca rozmycie linii widmowej, jest zgodnie z zasadą nieoznaczoności związana ze średnim czasem życia poziomu τ za pomocą wzoru:

$$\Gamma = \hbar/\tau \tag{1}$$

i nosi nazwę naturalnej szerokości linii widmowej.

W przypadku atomów sodu emitujących linię żółtego światła o długości fali 589 nm średni czas życia $\tau = 1{,}5 \cdot 10^{-8}$ s, a więc naturalna szerokość linii wynosi $\Gamma = 4{,}4 \cdot 10^{-8}$ eV. W przypadku promieniowania γ rtęci ^{189}Hg średni czas życia jądra w stanie wzbudzonym 411 keV $\tau = 2 \cdot 10^{-11}$ s i obliczona z wzoru (1) naturalna szerokość linii $\Gamma = 3{,}3 \cdot 10^{-5}$ eV. In-

teresujące jest porównanie w obu wypadkach względnych szerokości linii widmowych, to jest stosunków szerokości linii do odpowiednich energii przejść. Dla atomowego przejścia w sodzie energia fotonów promieniowania jest równa 2,1 eV czyli $\Gamma/E = 2\cdot10^{-8}$, a dla jądrowego przejścia w rtęci $\Gamma/E = 8\cdot10^{-11}$. Względna szerokość linii jest miarą dokładności, z jaką można by wyznaczyć energię przejść, gdyby doświadczenie pozwoliło na obserwację naturalnej szerokości linii. Na ogół linia widmowa ulega dodatkowemu, znacznemu poszerzeniu wskutek wielu zjawisk zachodzących w badanych substancjach. Najbardziej istotne dla dalszych rozważań jest tak zwane poszerzenie dopplerowskie, wywołane ruchem cieplnym atomów emitujących promieniowanie.

Dopplerowskie poszerzenie linii widmowych

Jeżeli promieniujący atom porusza się z prędkością v w kierunku obserwatora, to występuje efekt Dopplera polegający na tym, że obserwowana energia fotonu ulega zmianie o

$$\Delta E = Ev/c,$$

gdzie E jest energią fotonu w układzie związanym z poruszającym się atomem, a c prędkością światła $= 3\cdot10^8$ m/s.

W ruchu cieplnym atomów reprezentowane są różne prędkości i różne kierunki ruchu, co prowadzi do poszerzenia (rozmycia) linii widmowej. Ze wzrostem temperatury średnia prędkość ruchu cieplnego rośnie i wywołane przez efekt Dopplera poszerzenie linii staje się coraz większe. Poszerzenie to, zwane temperaturowym, dla źródła gazowego w temperaturze bezwzględnej T może być w przybliżeniu obliczone z wzoru

$$D = \frac{E}{c}\sqrt{\frac{3kT}{M}},$$

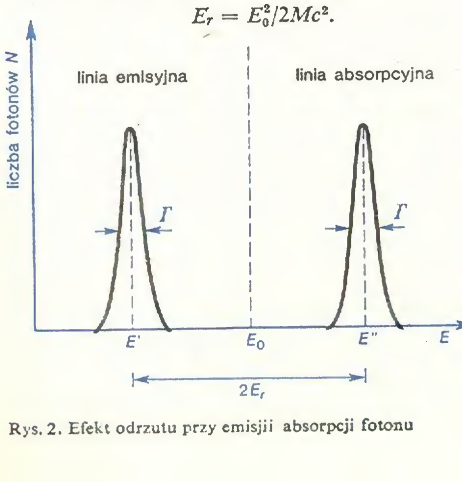
poszerzenie temperaturowe

gdzie $k = 8{,}62\cdot10^{-5}$ eV/K jest stałą Boltzmanna, a M — masą emitującego atomu. W temperaturze pokojowej $T \approx 300$ K dla przejść optycznych w sodzie $D = 3{,}7\cdot10^{-6}$ eV, a dla przejść γ w rtęci $D = 0{,}27$ eV. W obu wypadkach poszerzenie temperaturowe wielokrotnie przewyższa naturalną szerokość linii.

Odrzut spowodowany emisją lub absorpcją promieniowania

Atom lub jądro emitujące foton promieniowania doznaje odrzutu, którego energia może być obliczona z zasady zachowania pędu. Pęd odrzutu p musi być równy pędowi E_0/c emitowanego fotonu. Ponieważ energia odrzutu $E_r = p^2/2M$, więc

$$E_r = E_0^2/2Mc^2. \tag{2}$$



Rys. 2. Efekt odrzutu przy emisjii absorpcji fotonu

Zgodnie z zasadą zachowania energii emitowany foton ma energię E' niższą od energii wzbudzenia E_0, gdyż część energii przejścia zostaje zużyta na odrzut $E' = E_0 - E_r$. Podobnie, w procesie absorpcji, atom lub jądro pochłaniające foton również doznaje odrzutu, a więc energia fotonu, który mógłby być rezonansowo zaabsorbowany, musi mieć energię E'' wyższą o E_r od energii wzbudzenia E_0. Jak stąd widać, występuje niedopasowanie energetyczne fotonów emitowanych i absorbowanych o wartość $2E_r$. Tę sytuację przedstawiono na rys. 2.

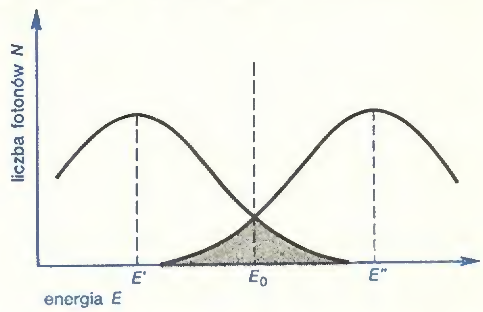
niedopasowanie energetyczne fotonów

Warunki obserwacji absorpcji rezonansowej

Jeżeli przesunięcie energetyczne $2E_r$ linii emisji i absorpcji jest mniejsze od szerokości naturalnej linii, to absorpcja rezonansowa może być łatwo zaobserwowana. Zachodzi to w przypadku przejść optycznych. Dla sodu $\Gamma = 4{,}4\cdot10^{-8}$ eV a $2E_r = 2\cdot10^{-10}$ eV czyli $\Gamma \gg 2E_r$, co tłumaczy pozytywny rezultat opisanego poprzednio prostego doświadczenia z absorpcją i rozpraszaniem rezonansowym żółtej linii w parze sodu.

Dla przejść γ odrzut jądra jest tak duży, że niedopasowanie energetyczne $2E_r$ znacznie przekracza szerokość naturalną linii. Wystarczy dla przejścia γ w rtęci porównać $2E_r = 0{,}92$ eV i $\Gamma = 3{,}3\cdot10^{-5}$ eV. Przesunięcie energetyczne wskutek odrzutu jest 30 000 razy większe od szerokości naturalnej linii. Nic więc dziwnego, że doświadczenia z promieniowaniem γ, przeprowadzone w sposób analogiczny do doświadczeń z dziedziny optyki, nie mogą się udać.

Od 1950 r. wykonano wiele pomysłowych eksperymentów w celu zaobserwowania absorpcji i rozpraszania rezonansowego promieniowania γ. We wszystkich tych doświadczeniach starano się różnymi metodami skompensować przesunięcia energetyczne $2E_r$.



Rys. 3. Poszerzenie temperaturowe linii widmowych umożliwia obserwację absorpcji rezonansowej

sposoby kompensacji przesunięcia energetycznego

Jedna z metod polegała na wykorzystaniu dopplerowskiego, temperaturowego poszerzenia linii. W wysokich temperaturach poszerzenie temperaturowe staje się tak duże, że skrzydła linii emisji i absorpcji pokrywają się w pewnym stopniu i obserwacja absorpcji rezonansowej staje się możliwa. Prawdopodobieństwo absorpcji rezonansowej jest dane przez stosunek powierzchni zakreskowanej na rys. 3 do całkowitej powierzchni linii widmowej.

Inna metoda skompensowania przesunięcia energetycznego polega na nadaniu źródłu w aparaturze pomiarowej dużej prędkości w kierunku absorbenta. Przy dostatecznej prędkości tego ruchu, przesunięcie linii emisji powstające wskutek efektu Dopplera, powoduje jej pokrycie się z linią absorpcji i zjawisko rezonansu może być zaobserwowane. Całkowite skompensowanie niedopasowania energetycznego wymaga stosowania bardzo dużych prędkości. W przypadku ^{198}Hg jest to prędkość ok. 700 m/s. Prędkości tego rzędu realizuje się za pomocą rotora o bardzo wysokiej częstości obrotów, na którego ramionach znajdują się źródła, a skolimowana przez układ przesłon ołowianych wiązka promieni γ wylatuje stycznie do obwodu koła, po którym źródła się poruszają.

Można również do skompensowania rozsunięcia linii emisji i absorpcji wykorzystać ruch jąder wywołany przez odrzut w jakimś procesie jądrowym bezpośrednio poprzedzającym emisję promieniowania γ. Takimi procesami mogą być rozpad β, wychwyt elektronu lub jakaś reakcja jądrowa.

Odkrycie zjawiska bezodrzutowej emisji i absorpcji promieniowania γ przez jądra w kryształach stworzyło zupełnie nowe możliwości badania absorpcji rezonansowej promieniowania γ, wywołało lawinę prac naukowych o podstawowym znaczeniu w różnych dziedzinach fizyki, a R. Mössbauerowi przyniosło nagrodę Nobla w 1961 r.

Zjawisko bezodrzutowej emisji i absorpcji promieniowania

W 1957 r. R. Mössbauer rozpoczął w Heidelbergu doświadczenie nad rezonansową absorpcją promieniowania γ. Badał rozpraszanie rezonansowe linii 129 keV ^{191}Ir. Ponieważ przesunięcie energetyczne $2E_r$ jest w tym wypadku w przybliżeniu równe poszerzeniu temperaturowemu linii γ w temperaturze pokojowej, linie emisji i absorpcji pokrywały się w znacznym stopniu i efekt rezonansowy był obserwowany.

doświadczenie Mössbauera

Mössbauer obniżył temperaturę źródła i absorbenta, spodziewając się zmniejszenia efektu. Tymczasem rozproszenie rezonansowe wzrosło! Mössbauer wytłumaczył ten zaskakujący efekt na podstawie znanej od 20 lat teorii Lamba wpływu wiązania atomów w sieci krystalicznej na przekrój czynny chwytania powolnych neutronów i wykazał w pięknym eksperymencie, że dość znaczny ułamek promieniowania γ irydu nie wykazuje przesunięcia energetycznego, wywołanego przez odrzut jąder, ani też dopplerowskiego poszerzenia temperaturowego. Obserwowana linia γ ma szerokość naturalną. Brak przesunięcia energetycznego tłumaczy się tym, że w pewnej części przypadków pęd odrzutu atomu związanego w sieci krystalicznej, towarzyszący emisji lub absorpcji fotonu γ, zostaje przejęty przez cały kryształ. Wówczas we wzorze (2) na energię odrzutu zamiast masy atomu M wystąpi masa całego kryształu. Nawet w bardzo drobnym proszku krystalicznym poszczególne ziarna zawierają tak wielką liczbę atomów, że energia odrzutu staje się zaniedbywalnie mała, o wiele mniejsza od naturalnej szerokości linii widmowej. Zjawisko emisji i absorpcji można wtedy traktować jako procesy bezodrzutowe. Dopasowanie energii fotonów emitowanych i absorbowanych jest idealne, a więc spełnione są warunki obserwacji absorpcji rezonansowej.

Zniknięcie poszerzenia dopplerowskiego wywołanego ruchami termicznymi atomów jest spowodowane dużą częstością drgań atomów w sieci krystalicznej ($\omega \approx 10^{13}$ s^{-1}). W ciągu czasu życia wzbudzonego stanu jądrowego średnia wartość składowej prędkości atomu w kierunku emisji fotonu jest równa zeru i efekt Dopplera pierwszego rzędu nie występuje.

Wielkość zjawiska Mössbauera zależy od prawdopodobieństwa, że pęd odrzutu zostanie przejęty przez cały kryształ. Prawdopodobieństwo to może być obliczone na podstawie określonego modelu sieci krystalicznej. Ponieważ na energię kryształu składa się energia ruchów drgających atomów wokół położeń równowagi, możemy kryształ traktować jako zbiór oscylatorów, których częstości drgań ω zawarte są w pewnym przedziale. Zmiana energii kryształu polega na tworzeniu lub niszczeniu kwantów $\hbar\omega$ zwanych fononami, czemu towarzyszy odpowiednia zmiana liczb kwantowych oscylatorów. Bezodrzutowa emisja lub absorpcja fotonu γ zachodzi wówczas, gdy odrzut atomu nie prowadzi do wzbudzenia sieci krystalicznej czyli do kreacji fononów, a pęd odrzutu jest przejęty przez cały kryształ.

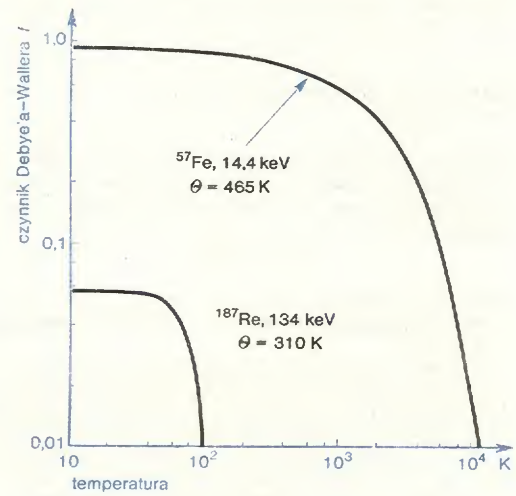
prawdopodobieństwo zjawiska Mössbauera

Prawdopodobieństwo zjawiska Mössbauera można określić przez podanie stosunku liczby fotonów γ emitowanych bezodrzutowo do całkowitej liczby fotonów γ. Ułamek ten zwany czynnikiem Debye'a-Wallera, obliczony na gruncie debye'owskiego modelu sieci krystalicznej, wyraża się wzorem

czynnik Debye'a–Wallera

$$f = \exp\left\{-\frac{3E_r}{2k\theta}\left[1+4\left(\frac{T}{\theta}\right)^2 \int_0^{\theta/T} \frac{x\,dx}{e^x-1}\right]\right\}, \quad (3)$$

gdzie θ jest temperaturą Debye'a danego kryształu związaną z maksymalną częstością drgań atomów ω_{max} zależnością $\theta = \omega_{max}/k$. Czynnik f, jak widać ze wzoru (3), jest tym większy, im mniejsza jest energia odrzutu jądra, im wyższa temperatura Debye'a

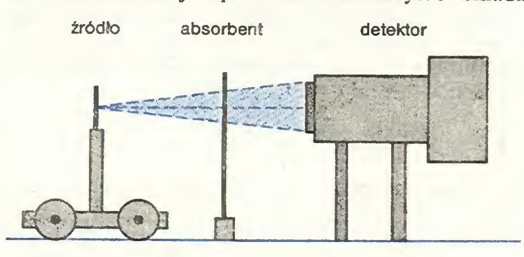


Rys. 4. Zależność czynnika Debye'a–Wallera od temperatury dla przejść 14,4 keV w jądrach żelaza ^{57}Fe i 134 keV w jądrach renu ^{187}Re

i w im niższej temperaturze T przeprowadzany jest eksperyment. Na rys. 4 pokazana jest zależność czynnika Debye'a–Wallera od temperatury dla przejść 14,4 keV w ^{57}Fe i 131 keV w ^{187}Re. Gdy wartości f są małe, doświadczenie należy przeprowadzać w temperaturze ciekłego azotu lub nawet ciekłego helu.

Technika pomiarów w rezonansowej spektroskopii promieniowania γ

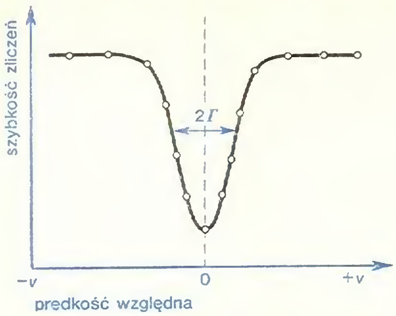
W celu zbadania kształtu linii promieniowania γ korzystamy z efektu Dopplera. Typowa aparatura mössbauerowska jest przedstawiona na rys. 5. Składa



Rys. 5. Schemat typowej aparatury mössbauerowskiej

się ona ze źródła, absorbenta i detektora promieni γ, którym najczęściej jest licznik scyntylacyjny. Źródło umieszczone w ruchomym uchwycie można wprawiać w ruch względem absorbenta za pomocą odpowiedniego mechanizmu. Ruch źródła musi być jednostajny, przy czym powinna być zapewniona możliwość nadawania różnych prędkości. Gdy źródło i absorbent mają identyczną strukturę chemiczną i pozostają względem sobie w spoczynku, energie linii emisji i absorpcji są do siebie dopasowane, absorpcja rezo-

nansowa jest maksymalna i detektor rejestruje najmniejszą liczbę fotonów. Gdy nadamy źródłu pewną prędkość v, efekt Dopplera psuje dopasowanie energii



Rys. 6. Zależność liczby transmitowanych fotonów od względnej prędkości ruchu źródła i absorbenta

o $\Delta E = Ev/c$. Absorpcja rezonansowa maleje i detektor rejestruje więcej fotonów przechodzących przez absorbent. W ten sposób wyznaczając szybkość zliczeń w zależności od prędkości źródła $N(v)$ otrzymujemy kształt linii promieniowania γ (rys. 6), z tym że ma ona podwójną szerokość naturalną 2Γ w wyniku nakładania się linii emisji i absorpcji — każdej o szerokości naturalnej Γ.

Prędkości z jakimi mamy do czynienia w spektroskopii mössbauerowskiej są na ogół małe. Przy energii przejścia γ 50 keV przesunięcie energii o naturalną szerokość linii Γ dla stanu jądrowego o średnim czasie życia 10^{-10} s odpowiada prędkości względnego ruchu źródła i absorbenta 4 cm/s; w przypadku średniego czasu życia 10^{-6} s będzie to prędkość 4 µm/s.

Należy podkreślić, że oprócz jądrowej absorpcji rezonansowej, promieniowanie γ jest pochłaniane i rozpraszane w absorbencie wskutek oddziaływania z elektronami. Efekt Mössbauera występuje więc tym silniej, im większa jest koncentracja nuklidu mössbauerowskiego. W celu zwiększenia efektu stosuje się często absorbenty wzbogacone w badany izotop.

Nuklidy mössbauerowskie

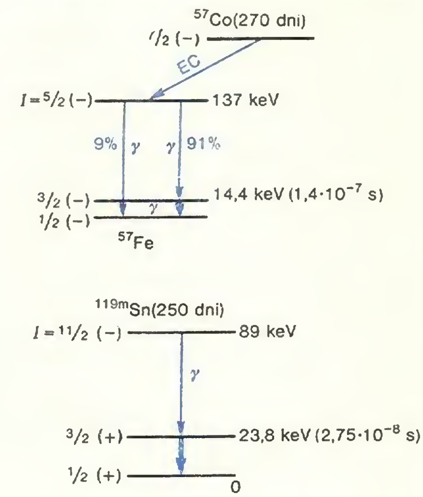
Zjawisko Mössbauera zostało dotychczas zaobserwowane dla 100 przejść γ w 80-ciu izotopach 44 pierwiastków chemicznych. Liczba ta jest ograniczona z następujących względów:

— Energia przejścia γ nie może przekraczać ok. 150 keV, gdyż dla wyższych energii emitowanych fotonów prawdopodobieństwo procesu bezodrzutowego staje się zbyt małe, by efekt Mössbauera mógł być obserwowany.

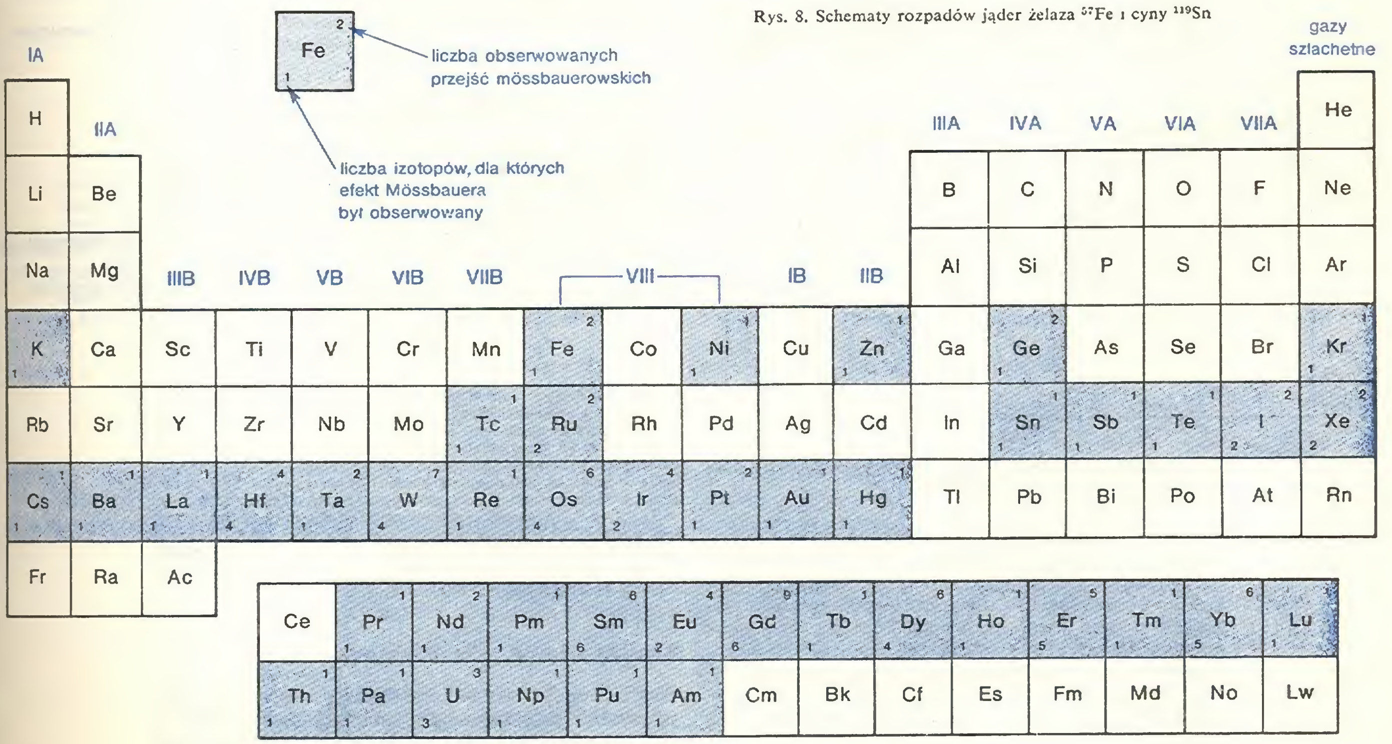
— Przejście γ musi prowadzić do stanu podstawowego stabilnego nuklidu lub nuklidu o tak długim czasie życia, aby móc rozporządzać dostateczną ilością materiału dla zrobienia absorbenta.

— Mössbauerowski stan jądrowy musi być osiągany w rozpadzie promieniotwórczym macierzystego nuklidu o dostatecznie długim czasie życia w porównaniu z czasem trwania eksperymentu, żeby źródło promieniowania γ nie wyczerpało się zbyt szybko.

— Czas życia stanu mössbauerowskiego nie może być zbyt długi ponieważ związana z tym bardzo mała



Rys. 8. Schematy rozpadów jąder żelaza ^{57}Fe i cyny ^{119}Sn



Rys. 7. Nuklidy, dla których obserwowano przejścia mössbauerowskie, na tle układu okresowego

szerokość linii mössbauerowskiej wymaga stosowania specjalnych środków ostrożności, aby występujące w laboratorium drgania nie zniszczyły warunków rezonansu. To ogranicza zbiór potencjalnych kandydatów dla spektroskopii mössbauerowskiej do jąder, w których średni czas życia stanu wzbudzonego jest krótszy od 10^{-5} s. Najdłużej żyjącym stanem, dla którego został dotychczas przeprowadzony eksperyment mössbauerowski, jest stan w ^{67}Zn o średnim czasie życia 9,3 μs i o energii wzbudzenia dziewięćdziesiąt trzy keV.

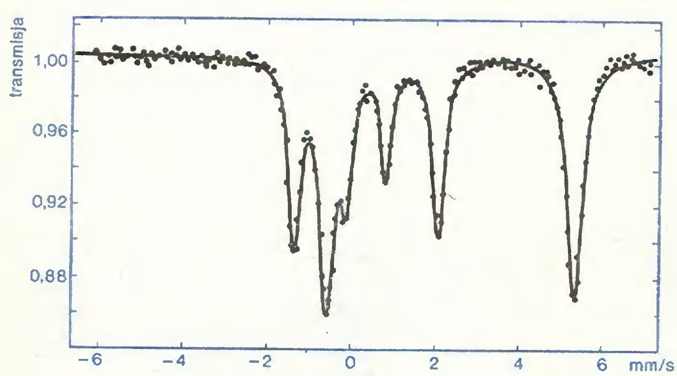
Rysunek 7 przedstawia układ periodyczny pierwiastków z oznaczeniem przypadków obserwacji zjawiska Mössbauera. Na rys. 8 podane są schematy poziomów ^{57}Fe i ^{119}Sn, których związki chemiczne są najczęstszymi obiektami badań spektroskopii mössbauerowskiej. Oba te nuklidy bardzo dobrze spełniają przytoczone wyżej cztery warunki.

Przesunięcie i rozszczepienia linii w widmach mössbauerowskich

Spektroskopia mössbauerowska jest jedyną metodą badawczą pozwalającą na bezpośrednią obserwację przesunięć izomerycznych i nadsubtelnej struktury linii promieniowania γ. Pozwala na to wyjątkowa energetyczna zdolność rozdzielcza. Miarą jej jest stosunek obserwowanej szerokości linii do energii przejścia. W przypadku ^{57}Fe stosunek ten jest równy $3{,}2 \cdot 10^{-13}$. Uwzględniając fakt, że precyzyjna aparatura mössbauerowska pozwala stwierdzić przesunięcie linii z górą sto razy mniejsze od jej szerokości naturalnej, otrzymujemy dokładność pomiaru rzędu 10^{-15}, z czym nie może rywalizować żaden inny pomiar w fizyce. Dobrze to zilustruje następujący przykład. Gdyby udało się zmierzyć odległość od Ziemi do Słońca (150 mln km) z dokładnością do 1 mm, to ta dokładność byłaby jeszcze kilkakrotnie mniejsza niż zdolność rozdzielcza osiągana w spektroskopii mössbauerowskiej związków Fe. Zastosowanie nuklidu mössbauerowskiego o krótszym czasie życia stanu wzbudzonego (np. ^{67}Zn) pozwoliłoby na zwiększenie zdolności rozdzielczej jeszcze o kilka rzędów wielkości.

dokładność metody mössbauerowskiej

Oddziaływanie jąder z otoczeniem w sieci krystalicznej prowadzi na ogół do przesunięć, poszerzeń lub rozszczepień linii promieniowania γ. Przykładowe, skomplikowane widmo mössbauerowskie pokazane jest na rys. 9. Równocześnie występowanie tych efektów w źródle i absorbencie bardzo komplikuje analizę widm. Dla uproszczenia interpretacji złożonego widma stosujemy standardowe źródła dające pojedynczą, możliwie wąską linię emisji. Obserwowane wówczas poszerzenia i rozszczepienia linii możemy przypisać wyłącznie absorbentowi i wnioskować z nich o sytuacji, w jakiej się tam znajdują badane jądra.



Rys. 9. Przykład złożonego widma mössbauerowskiego dla polikrystalicznego $FeCO_3$ w temperaturze 4,2 K

Przesunięcie izomeryczne (chemiczne) linii promieniowania γ

Maksimum absorpcji rezonansowej można obserwować przy prędkości $v = 0$ względnego ruchu źródła i absorbenta tylko wówczas, gdy źródło i absorbent mają identyczną strukturę chemiczną i krystaliczną i gdy znajdują się w tej samej temperaturze.

Różne otoczenia chemiczne badanych jąder powodują wzajemne przesunięcie linii emisji i absorpcji, zw. przesunięciem izomerycznym. W tym wypadku maksimum absorpcji rezonansowej można obserwować przy prędkości różnej od zera. Przesunięcie izomeryczne wywołane jest różnicą energii oddziaływania elektrostatycznego elektronów z ładunkiem jądra w atomach źródła i absorbenta. Składają się na nie dwa czynniki: różnica gęstości ładunku elektrycznego elektronów w obszarze jądra w źródle i w absorbencie oraz różnica rozmiarów jądra w stanie wzbudzonym i podstawowym.

Rolę obu tych czynników widać we wzorze na wartość przesunięcia izomerycznego:

$$\delta = \frac{2\pi}{5} Ze^2(R_w^2 - R_p^2)\,[|\psi_a(0)|^2 - |\psi_{\acute{z}}(0)|^2],$$

gdzie R_w i R_p są średnimi promieniami jądra w stanach wzbudzonym i podstawowym, a kwadraty funkcji falowych elektronów $|\psi_a(0)|^2$ i $|\psi_{\acute{z}}(0)|^2$ opisują gęstości elektronów w miejscu jądra (czyli w „zerze" układu współrzędnych) odpowiednio dla absorbenta i źródła. Z jest liczbą atomową danego pierwiastka a e — ładunkiem elementarnym.

Nadsubtelna struktura linii promieniowania γ

Nadsubtelna struktura, czyli rozszczepienie na składowe linii promieniowania γ jest spowodowane oddziaływaniem dipolowego momentu magnetycznego jądra z polem magnetycznym lub oddziaływaniem elektrycznego momentu kwadrupolowego jądra z gradientem pola elektrycznego. Pomiary rozszczepienia linii wywołanych tymi oddziaływaniami pozwalają wyciągnąć wnioski o polach elektromagnetycznych działających na badane jądra w atomach danej próbki krystalicznej. Jądro spełnia w tym wypadku rolę miniaturowej sondy umieszczonej w centrum atomu.

oddziaływanie magnetyczne

W wyniku oddziaływania momentu magnetycznego jądra μ z polem magnetycznym H poziomy energetyczne jądra ulegają rozszczepieniu na składowe o energiach danych przez wzór

$$E_m = E_0 - \frac{\mu H m_I}{I},$$

gdzie E_0 jest energią nierozszczepionego poziomu, a $m_I = I, I-1 \ldots -I$ są magnetycznymi liczbami kwantowymi jądra o spinie I.

W ogólnym wypadku oba poziomy, między którymi zachodzi przejście γ, są rozszczepione, co z kolei powoduje rozszczepienie linii promieniowania γ na szereg składowych. Liczba składowych zależy od spinów jądra w obu stanach i jest ograniczona przez reguły wyboru dla danego typu promieniowania.

oddziaływanie elektryczne

Oddziaływanie elektrycznego momentu kwadrupolowego jądra eQ z gradientem pola elektrycznego eq również prowadzi do rozszczepienia poziomów jądrowych. Energie poszczególnych składowych, w przypadku gradientu pola elektrycznego o symetrii osiowej, wyrażają się wzorem

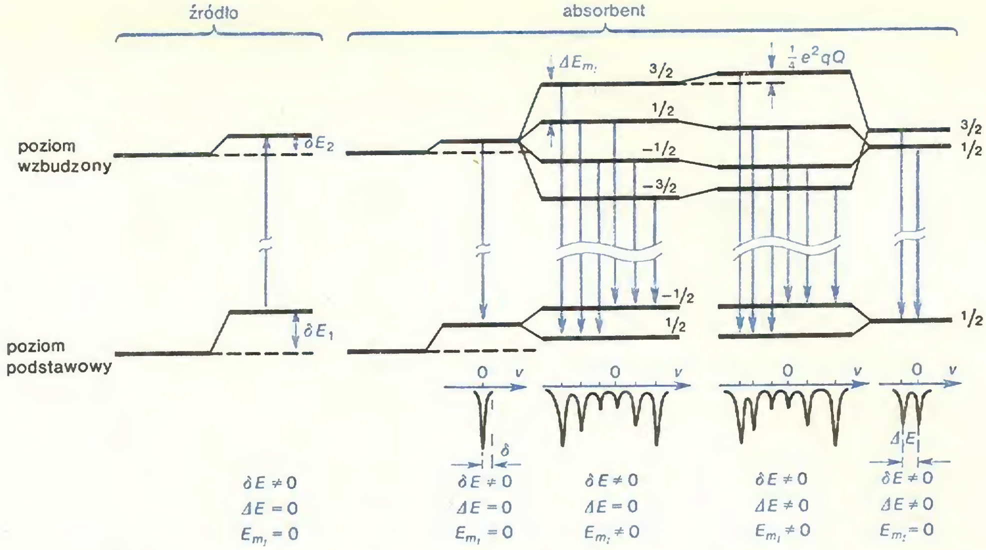
$$E_a = E_0 + \frac{e^2 qQ\,[3m_I^2 - I\,(I+1)]}{4I\,(2I-1)}.$$

Liczba składowych jest w tym wypadku mniejsza niż przy oddziaływaniu magnetycznym, gdyż podstany o magnetycznych liczbach kwantowych $-m_I$ i $+m_I$ mają tę samą energię.

Rysunek 10 przedstawia schematycznie przesunięcie i rozszczepienie poziomów jądrowych ^{57}Fe wywołane oddziaływaniem jądra z powłoką elektronową atomu. W przypadku ^{57}Fe spiny stanu podstawowego i wzbudzonego wynoszą odpowiednio $I = {}^1/_2$ i $I = {}^3/_2$, moment magnetyczny w stanie podstawowym jest dodatni, a w stanie wzbudzonym ujemny. Liczba przejść jest ograniczona regułą wyboru $\Delta m_I = 0, \pm 1$. W dolnej części rysunku pokazane są widma mössbauerowskie odpowiadające różnym przypadkom oddziaływania.

Analiza widma mössbauerowskiego

Analizę widma mössbauerowskiego prowadzi się przeważnie przy użyciu elektronowej maszyny cyfrowej, stosując program, za pomocą którego można



Rys. 10. Przesunięcia i rozszczepienia poziomów jądrowych ^{57}Fe wywołane oddziaływaniem jądra z powłoką elektronową i otoczeniem w sieci krystalicznej

określić położenia poszczególnych linii, ich szerokości oraz powierzchnie opisane przez ich kontury. Dostarczone przez EMC dane są wykorzystane do obliczania wielkości przesunięć izomerycznych, rozszczepień linii na składowe w wyniku ewentualnych oddziaływań nadsubtelnych, a także do wyznaczania stosunków natężeń poszczególnych składowych widma. Wszystkie te dane pozwalają następnie wyciągać wnioski o oddziaływaniu badanych jąder z powłoką elektronową atomów i z ich otoczeniem w sieci krystalicznej. Dla uzyskania dodatkowych informacji o tych oddziaływaniach, a także dla ułatwiania interpretacji widma bardzo często pomiary wykonuje się w różnych temperaturach, co pozwala prześledzić temperaturową zależność parametrów wyznaczonych z widm mössbauerowskich.

Zastosowanie spektroskopii mössbauerowskiej

Wkrótce po odkryciu zjawiska Mössbauera wykonano szereg pomiarów o fundamentalnym znaczeniu w fizyce. Przeprowadzenie ich stało się możliwe dzięki wyjątkowej zdolności rozdzielczej spektroskopii mössbauerowskiej. Do pomiarów o podstawowym znaczeniu można zaliczyć pomiar poprzecznego efektu Dopplera oraz prędkości hipotetycznego „wiatru eteru". Najbardziej znanym z nich jest przeprowadzony przez Pounda i Rebkę pomiar zmiany energii fotonu w polu grawitacyjnym Ziemi. Kwant promieniowania, przelatując między punktami o różnych wartościach potencjału grawitacyjnego, zmienia swą energię. Jest to tzw. przesunięcie ku czerwieni obserwowane w astronomii. Gdy foton spada w polu grawitacyjnym Ziemi, względny przyrost energii jest dany przez wzór:

pomiar zmiany energii fotonu w polu grawitacyjnym Ziemi

$$\frac{\Delta E}{E} = \frac{gH}{c^2},$$

gdzie H jest różnicą wysokości przelatywaną przez foton, a g — wartością przyspieszenia ziemskiego. Wzór ten jest konsekwencją przypisania fotonowi masy $\frac{E}{c^2}$. Przy powierzchni Ziemi na wysokości jednego metra $\frac{\Delta E}{E} = 1{,}1 \cdot 10^{-16}$. Doświadczenia Pounda i Rebki, a następnie Pounda i Snidera, zostały wykonane w Uniwersytecie Harvard w Stanach Zjednoczonych. Wykorzystali oni wieżę, w której między źródłem i absorbentem można było uzyskać różnicę poziomów dwadzieścia dwa metry. Mierzone było przesunięcie energetyczne linii promieniowania γ ^{57}Fe dla fotonów lecących w górę i w dół. Pomiary wymagały bardzo wielkiej precyzji, gdyż oczekiwane przesunięcie stanowiło zaledwie ok. 1/500 obserwowanej szerokości linii. Wynik serii pomiarów, trwających łącznie kilka miesięcy,pozwolił stwierdzić występowanie efektu, którego wielkość, w granicach błędu około jednego procentu, zgadza się z przewidywaniami teorii.

fizyka jądrowa

W fizyce jądrowej efekt Mössbauera pozwolił w kilku przypadkach wyznaczyć różnicę rozmiarów jądra w stanie wzbudzonym i podstawowym, a także zmierzyć dipolowe momenty magnetyczne, elektryczne momenty kwadrupolowe oraz średnie czasy życia wielu stanów jądrowych.

fizyka ciała stałego i chemia

Spektroskopia mössbauerowska stała się rutynową metodą badawczą i pomiarową w fizyce ciała stałego i chemii. Trudno jest podać pełną listę problemów, w których ma ona zastosowanie. Dla przykładu można wymienić następujące problemy:

— struktura wiązań chemicznych i charakter wiązań chemicznych,
— przebieg reakcji chemicznych i rola katalizatorów,
— materiały amorficzne i zestalone roztwory,
— zjawiska relaksacyjne,
— przejścia fazowe i zjawiska krytyczne,
— stopy, zanieczyszczenia i defekty sieci krystalicznej, uszkodzenia radiacyjne,
— rozkłady ładunków i momentów magnetycznych w ciałach stałych,
— struktura magnetyczna i przekazywane oddziaływania nadsubtelne,
— struktura i dynamika cząsteczek biologicznie czynnych,
— analiza minerałów.

O tym, jak szeroki zakres ma spektroskopia mössbauerowska, najlepiej świadczy fakt, że w naukowej literaturze światowej pojawia się rocznie około tysiąca publikacji poświęconych zastosowaniom tej metody.

D. C. Champeney i in. Phys.Rev.Lett. 7, 241 (1963); W. I. Goldanski *Efekt Mössbauera i jego zastosowania w chemii*, Warszawa 1966; U. Gonser *Mössbauer Spectroscopy*, Springer-Verlag 1975; H. J. Hay i in. Phys. Rev.Lett. 4, 165 (1960); A. Hrynkiewicz *Efekt Mössbauera i jego zastosowania w fizyce ciała stałego* w: *Cząstki elementarne, jądra atomowe, promieniotwórczość*, Warszawa 1967; A. Hrynkiewicz, D. Kulgawczuk *Spektroskopia mössbauerowska*, Metody badań minerałów i skał, Warszawa 1979; R. L. Mössbauer Z. Physik 151, 124(1958); R. Y. Pound, G.A. Rebka Phys.Rev.Lett. 4, 274, 337 (1960); A. Wértes i in. *Mössbauer Spectroscopy*, Budapest 1979.

KIERUNKI ROZWOJU OPTYKI

Optyka współczesna

Adam Kujawski

Optyka jest bardzo obszerną dziedziną nauk fizycznych, a niektóre jej zagadnienia zalicza się — ze względu na ich odrębność — również do techniki lub medycyny. W ostatnim ćwierćwieczu zakres badań i zastosowań optyki uległ istotnym przemianom. W rezultacie do optyki współczesnej weszły zupełnie nowe zjawiska i zastosowania, a opis większości znanych zjawisk optycznych uzyskał znacznie głębszą interpretację fizyczną. Wydarzeniem przełomowym — zarówno ze względu na zastosowanie, jak i poszerzenie naszej wiedzy o naturze światła — stało się zbudowanie lasera, źródła światła o bardzo wysokim stopniu spójności. Ten fakt zapoczątkował też rozwój kierunku badań, który przyjęto nazywać elektroniką kwantową.

lasery — przełom w optyce

Dwa artykuły w tym dziale są poświęcone podstawowym zjawiskom fizycznym zachodzącym w laserach i zasadzie działania laserów oraz właściwościom światła spójnego. Osobny artykuł omawia wybrane spośród bardzo licznych zastosowań laserów.

Moc promieniowania laserowego może o wiele rzędów wielkości przewyższać moc źródeł konwencjonalnych. Fakt ten umożliwił odkrycie nowych zjawisk — nieliniowych — oraz znalazł liczne zastosowania. Jednak w tym dziale przede wszystkim podkreśla się inną podstawową cechę światła laserowego (zarówno jego części widzialnej, jak również bliskiego nadfioletu i bliskiej podczerwieni) — jego spójność — i omówieniem tego zagadnienia rozpoczyna się cały dział optyki. Niewątpliwie możliwość uzyskania niemal doskonałej spójności należy do fundamentalnych osiągnięć zarówno przy poszukiwaniu odpowiedzi, jaka jest natura światła, jak i ze względu na zastosowania. Chociaż pojęcie spójności jest tak stare jak zjawisko interferencji, jego istotnie nowe znaczenie zarówno w klasycznym (falowym) jak i kwantowym (korpuskularnym) opisie światła pojawiło się dopiero po zbudowaniu laserów. To, że można otrzymywać światło spójne oraz to, że można wykorzystywać jego różne właściwości statystyczne odgrywa zasadniczą rolę we wszystkich zjawiskach optycznych od fizyki atomowej począwszy, a na optycznym przetwarzaniu danych skończywszy. Należy podkreślić, że obecnie zagadnienie spójności obejmuje nie tylko zjawiska interferencyjne, lecz przede wszystkim problemy właściwości statystycznych światła. Przykładem jest zagadnienie opisu wiązek światła, z których każda może dać ostre prążki w doświadczeniach interferencyjnych, a które — na podstawie doświadczeń ze zliczaniem fotonów — mogą istotnie różnić się od siebie, rozkład statystyczny fotonów może być w nich zupełnie różny.

spójność światła

Wybranym kierunkom badań, z pewnością mającym duże znaczenie dla dalszego rozwoju nauki i techniki, poświęcono cztery artykuły. Dwa z nich „Optyka nieliniowa" oraz „Ultrakrótkie impulsy światła", informują o nowych zjawiskach optycznych i nowych technikach badawczych. Hasła „Holografia" oraz „Optyka fourierowska" są poświęcone dziedzinom, które w oryginalny sposób wykorzystują spójność światła laserowego.

Łatwiej zrozumiemy dlaczego takie kierunki rozwoju obrała optyka współczesna, jeśli sobie uświadomimy, że całe widmo fal elektromagnetycznych, od fal najkrótszych począwszy (promieniowanie γ, częstości rzędu 10^{23} Hz) aż do fal najdłuższych (fale radiowe, częstości do kilku Hz), opisuje teoria Maxwella, która mówi o właściwościach falowych, oraz teoria kwantowa, relacjonująca właściwości korpuskularne. W wypadku fal radiowych energia fotonu $h\nu$ jest bardzo mała i teoria Maxwella odgrywa bardzo dużą rolę, w wypadku fal najkrótszych większa jest rola teorii kwantowej. Jeśli chodzi o zakres optyczny (częstości rzędu 10^{13}–10^{18} Hz; część widzialna stanowi wąskie pasmo wokół wartości 10^{15} Hz), to można powiedzieć, że znajduje się on w środku widma elektromagnetycznego i obydwa opisy teoretyczne są równie przydatne. Dla współczesnej optyki charakterystyczne jest to, że z jednej strony znaczna część technik dotyczących fal długich, takich jakimi się posługuje radiotechnika czy telekomunikacja, stosuje się obecnie w części optycznej widma, z drugiej strony, kwantowa natura promieniowania optycznego uwidacznia się w zupełnie nowy sposób w takich np. zjawiskach, jak fluorescencja rezonansowa i procesy Ramana i Brillouina wymuszonego rozpraszania. W rezultacie, dzięki odkryciu nowych zjawisk i znalezieniu nowych metod eksperymentalnych, zarówno właściwości falowe jak i korpuskularne, odgrywające równorzędną rolę w optycznej części widma elektromagnetycznego, są jeszcze wyraźniej zaakcentowane i udokumentowane.

Charakterystyczne cechy rozwoju współczesnej optyki można podkreślić w trochę inny sposób dwoma następującymi spostrzeżeniami: Po pierwsze — daje się zauważyć tendencja do tego, by większość osiągnięć elektroniki współczesnej, wykorzystującej mikrofale i fale radiowe, przenieść na zakres optyczny. Nazwy nowo powstałych dziedzin, takich jak optoelektronika czy optyka zintegrowana, najlepiej o tym świadczą. Postęp i rozwój obydwu wymienionych dziedzin jest zresztą ściśle związany z opanowaniem nowych technologii wytwarzania materiałów, szczególnie materiałów półprzewodnikowych. Na specjalną uwagę zasługuje tutaj dalsze wykorzystanie wyjątkowych własności fizycznych złącza *p-n*. Oprócz dotychczas znanych przyrządów elektronicznych, wykorzystujących złącze *p-n* — jak np. diody, fotoogniwa, tranzystory — pojawił się laser półprzewodnikowy, który dzięki opanowaniu technologii otrzymywania heterozłącz stał się jednym z najważniejszych elementów optycznych współczesnej elektroniki.

nawiązanie do elektroniki

Drugie spostrzeżenie dotyczy jeszcze bliższego powiązania optyki z fizyką atomową. Ilustracją tego faktu jest pojawienie się zupełnie nowych metod spektroskopii optycznej oraz powstanie dziedziny zwanej optyką kwantową. Ta ostatnia obejmuje m.in. problemy statystyki fotonów i roli właściwości statystycznych światła w procesach nieliniowych. Do nowych kwantowych zjawisk optycznych należą zjawiska koherentnego oddziaływania światła z układami atomowymi, jak np. echo fotonowe i wymuszona przezroczystość ośrodka. Na zakończenie zwróćmy uwagę, że o tym, jak dalece metody klasycznej elektroniki i kwantowego opisu promieniowania elektromagnetycznego połączyły się i przeniknęły wzajemnie w zakresie optycznym widma elektromagnetycznego, świadczy również nazwa elektronika kwantowa. Jest ona często używana dla określania zjawisk i technik laserowych, które w większości można także zaliczyć do optyki współczesnej.

optyka kwantowa

elektronika kwantowa

Wyborem tematów w tym rozdziale można było objąć tylko niektóre z obecnych kierunków rozwoju. Tematy te ilustrują fakt, że optyka współczesna stała się dziedziną wzajemnie powiązaną zarówno z innymi działami fizyki (szczególnie z fizyką atomową i fizyką ciała stałego), jak i z różnymi kierunkami techniki (zwłaszcza z elektroniką i telekomunikacją).

Spójność światła

Adam Kujawski

Rozwój nauk fizycznych pierwszego ćwierćwiecza XX w. przyczynił się między innymi istotnie do lepszego rozumienia, jaka jest natura światła i jaki jest mechanizm świecenia atomów, z których zbudowane są źródła promieniowania świetlnego. Odnosi się to, rzecz jasna, do źródeł konwencjonalnych, a więc takich, jak płomień, żarówka, gaz pobudzony do świecenia itp. Źródła te obecnie przyjęto nazywać, i tak je dalej nazywać będziemy, źródłami termicznymi, w odróżnieniu od źródeł laserowych, lub krótko — laserów, które od 1961 r. znalazły się w centrum zainteresowań naukowych i technicznych. Jedną z najważniejszych cech charakterystycznych dla światła laserowego, istotnie różniących je od światła termicznego, jest wysoki stopień spójności. Fakt ten, obok innych specyficznych właściwości wiązki laserowej, jest zawsze podkreślany przy omawianiu zasady działania lasera i właściwości światła przezeń wytwarzanego. Światło spójne lasera znalazło też liczne zastosowania w praktyce, np. w holografii.

Problem czy dwie wiązki światła są spójne (koherentne), czy nie, jest problemem tak starym, jak badanie zjawisk interferencji światła. W ciągu ostatnich dwudziestu lat uzyskał on zupełnie nowe znaczenie i stał się ważnym rozdziałem badań w optyce. W okresie przed zbudowaniem lasera, to znaczy w latach pięćdziesiątych, a następnie w okresie badań nowych zjawisk optycznych charakterystycznych dla światła laserowego, ukształtowała się teoria spójności światła. Teoria ta, posługując się odpowiednim aparatem matematycznym, ściśle opisuje interferencję, zjawiska w których gra rolę spójność wyższych rzędów, oraz statystyki zliczeń fotonów dla światła termicznego i laserowego. Omówimy tutaj te teoretyczne i doświadczalne aspekty statystycznych właściwości światła, które zalicza się do współcześnie rozumianego pojęcia spójności światła.

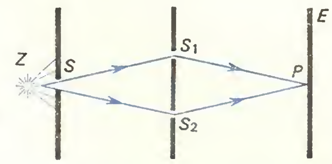
zjawisko interferencji

Interferencja jest zjawiskiem charakterystycznym dla wszystkich ruchów falowych. Właśnie przy obserwacji tego zjawiska wprowadza się pojęcie spójności (koherencji) ruchów falowych biorących udział w nakładaniu się lub, mówiąc inaczej, interferowaniu ze sobą. Najprostszym przykładem, łatwym do zaobserwowania jest nakładanie się dwu fal rozchodzących się na powierzchni wody, źródłami fal mogą być na przykład końce drgających prętów. Gdy drgają one ze stałą częstością, obserwuje się, że w wyniku nałożenia powstaje regularna struktura miejsc, w których drgania się wygaszają i wzmacniają, jak na il. 135, tabl. 33. Gdy jednak wytwarzane drgania mają różne częstości lub gdy wytwarzane są w sposób zupełnie przypadkowy, powstały ruch falowy nie wykazuje cech regularności. O falach omówionych wyżej mówimy odpowiednio, że są spójne lub niespójne. Jest to klasyfikacja uproszczona. Łatwo bowiem można sobie wyobrazić przypadki pośrednie, w których fala wypadkowa powstająca na powierzchni wody ma pewne cechy regularności niezbyt łatwe do obserwacji. Zwróćmy jeszcze uwagę, iż stwierdzenie, czy fale są spójne czy niespójne nie wiązało się dotąd z żadną wielkością mierzalną i oparte jest tylko na opisowym porównaniu obrazów interferencyjnych.

Spójność pierwszego rzędu

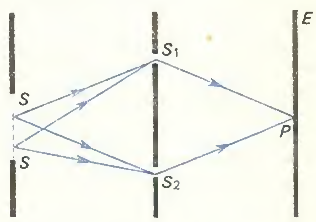
Doświadczenie interferencyjne, w którym nakładają się dwie fale świetlne i którego idea jest taka sama, jak doświadczenia z dwiema falami na wodzie, nosi nazwę doświadczenia Younga. Odegrało ono bardzo ważną rolę w rozwoju poglądów na naturę światła, gdyż w istocie jest dowodem na falową naturę światła. Przypomnijmy krótko warunki realizacji doświadczenia Younga. Światło jednobarwne ze źródła Z pada na dostatecznie mały otwór lub szczelinę S, a następnie na dwie szczeliny S_1 i S_2 w przesłonie,

doświadczenie Younga



Rys. 1. Światło wychodzące ze szczeliny S pada na szczeliny S_1 i S_2 i na ekranie E powstaje obraz interferencyjny

tak jak ilustruje to rys. 1. Światło wychodzące z S_1 i S_2 intereferuje i daje na ekranie E obraz interferencyjny, który obserwujemy bezpośrednio lub rejestrujemy na kliszy. Aby obraz ten był ostry, to znaczy, aby składał się z wyraźnych jasnych i ciemnych prążków, rozmiary szczeliny S muszą być dostatecznie małe. Powstanie miejsca jasnego lub ciemnego w danym punkcie P ekranu E zależy od różnicy dróg $S'S_1P$ i $S'S_2P$, gdzie przez S' oznaczyliśmy jakiś wybrany punkt szczeliny S (rys. 2). Gdy różnica ta



Rys. 2. Światło z punktów S' i S'' dociera do punktu P na ekranie

jest równa parzystej wielokrotności połowy długości fali, mamy miejsce wzmocnienia interferujących fal świetlnych, a gdy nieparzystej wielokrotności — miejsce wygaszenia. Powiększanie rozmiarów szczeliny S, przy ustalonej odległości od przesłony, prowadzi do znikania prążków. Wówczas bowiem do danego punktu P docierają promienie po drogach $S''S_1P$ i $S''S_2P$, dając w ogólności inny wynik interferencji niż dla promieni z punktu S'. Można jednak źródło o dowolnie dużych wymiarach (w istocie w danym wypadku jest nim szczelina S) odsunąć dostatecznie daleko od przesłony, tak że otrzyma się ostry obraz interferencyjny. Wówczas drogi $S'S_1P$ i $S''S_2P$ będą z bardzo dobrym przybliżeniem sobie równe.

Przy ustalonych wymiarach szczeliny S i jej odległości od przesłony stwierdza się, że dla pewnej odległości szczelin S_1 i S_2 i odległości większych obraz interferencyjny znika. Odległość ta określa wymiary obszaru spójności pola świetlnego wychodzącego ze źródła S i dochodzącego do przesłony. Gdy S_1 i S_2 leżą w obszarze spójności, obserwuje się prążki ciemne i jasne, a gdy S_1 i S_2 leżą za tym obszarem — prążków nie ma. Wielkość obszaru spójności zależy zatem od odległości i wymiarów źródła, jakim w danym wypadku jest szczelina S. Aby lepiej zilustrować powyższe zależności podajemy wybrane dane liczbowe, które uzyskujemy posługując się ścisłą teorią koherencji, a które potwierdzają pomiary. Dla źródła światła jednobarwnego o długości fali $\lambda = 600$ nm, którego średnica równa jest 1 mm, w odległości 20 m w płaszczyźnie równoległej do płaszczyzny źródła średnica obszaru spójności wynosi 3,8 mm. Oznacza to oczywiście, że aby otrzymać dobry obraz interferencyjny, szczeliny S_1 i S_2 powinny być w odległości mniejszej niż 3,8 mm. Wprowadzone tutaj pojęcie obszaru spójności wiąże się z pojęciem spójności przestrzennej. Chodzi o to, że wyznaczając kolejno

obszar spójności

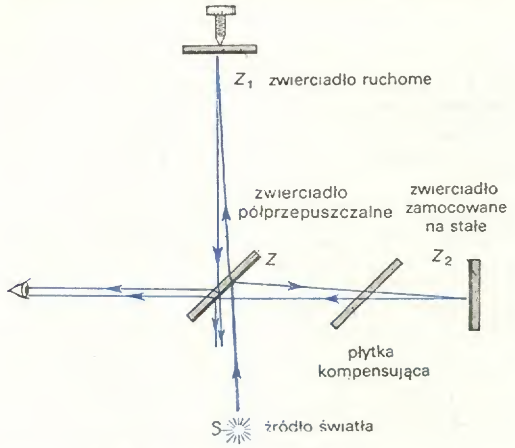
spójność przestrzenna

doświadczenie Michelsona

tego rodzaju powierzchnie, możemy określić w przestrzeni obszar o znanych powierzchniach, na których faza optycznej fali elektromagnetycznej jest taka sama.

spójność czasowa

Przejdźmy teraz do innego ważnego eksperymentu intereferencyjnego, który przeprowadził Michelson. Przyrząd, którym się posługiwał, nosi obecnie nazwę interferometru Michelsona. Za jego pomocą można zilustrować pojęcie spójności czasowej. Schemat zasadniczy doświadczenia podany jest na rys. 3. Za-



Rys. 3. Schemat interferometru Michelsona

znaczono na nim źródło światła jednobarwnego S, wiązkę światła dzielącą się na dwie na zwierciadle półprzepuszczającym Z, a następnie powstałe wiązki odbite od zwierciadeł Z_1 i Z_2, które nakładają się na siebie i na ekranie mogą dawać ostry obraz interferencyjny. Gdy odległości ZZ_1 i ZZ_2 są równe, powstałe wiązki nakładają się z takimi samymi fazami. Gdy jednak droga przebywana przez światło w jednym z ramion interferometru jest większa niż w drugim, wiązki są przesunięte w fazie względem siebie. Okazuje się, że istnieje graniczna wartość różnicy dróg optycznych w ramionach interferometru, dla której obraz interferencyjny znika. Oznaczymy tę wartość przez l_c. Nazywa się ją długością spójności. Wprowadza się również czas spójności zdefiniowany przez $\tau = l_c/c$, gdzie c oznacza prędkość światła. Z dobrym przybliżeniem można przyjąć, że τ określa przedział czasu, w jakim pola elektromagnetyczne wiązek świetlnych mają tę samą fazę. Jeśli więc doprowadzamy do interferencji dwu wiązek powstałych z jednej, jak np. w interferometrze Michelsona, i w miejscu nakładania się pola wiązek są przesunięte w czasie o mniej niż τ, to obserwuje się dobry obraz interferencyjny. Dla światła pochodzącego ze źródeł termicznych najdłuższy osiągalny czas koherencji jest rzędu 10^{-8} s, co odpowiada długości koherencji 3 m.

długość i czas spójności

W obydwu powyższych doświadczeniach interferencyjnych odnoszących się odpowiednio do spójności przestrzennej i czasowej, mieliśmy do czynienia z nakładaniem się dwu wiązek światła. Oznaczmy przez $\vec{E}_1$ i $\vec{E}_2$ wektory natężenia pól elektrycznych wiązek nakładających się i przyjmijmy dla uproszczenia, że kierunki wektorów pól są tak samo skierowane. Wartość natężenia pola wiązki powstałej w wyniku nałożenia wynosi więc w tym wypadku $E = E_1+E_2$. Można wykazać, że zaczernienie obrazu na kliszy zależy od odpowiednio uśrednionej gęstości energii pola elektrycznego wiązki. Istnieją dwa zasadnicze powody uśredniania: 1) częstość fali elektromagnetycznej w zakresie optycznym jest rzędu $10^{15}\,s^{-1}$ i klisza nie rejestruje zmian pola o tak dużej częstości; 2) fazy pola E mogą się bardzo szybko zmieniać w sposób przypadkowy. Drugi z wymienionych punktów wiąże się w istotny sposób z mechanizmem świecenia atomów źródła i powrócimy do niego dalej.

Ponieważ gęstość energii jest wprost proporcjonalna do wielkości E^2, zaczernienie zależy od wielkości $I = \langle E^2 \rangle$, gdzie ostre nawiasy oznaczają uśrednianie. Mamy więc

$$I = \langle E^2 \rangle = \langle E_1^2 \rangle + \langle E_2^2 \rangle + 2\langle E_1 E_2 \rangle.$$

Ponieważ $I_1 = \langle E_1^2 \rangle$, $I_2 = \langle E_2^2 \rangle$, więc

$$I = I_1 + I_2 + 2\langle E_1 E_2 \rangle. \quad (1)$$

Ostatni człon we wzorze (1) nosi nazwę członu interferencyjnego. Przeprowadzenie uśrednienia, które tutaj wprowadziliśmy, wymaga znajomości odpowiednich rozkładów prawdopodobieństwa. Można też, przy odpowiednich założeniach, rozumieć je jako uśrednienie po czasie. Problem równoważności obu sposobów uśredniania oraz kwestia kwantowego opisu zjawiska interferencji należą do współczesnej teorii koherencji światła. Dokładniejsza analiza wykazuje, że dla światła prawie monochromatycznego człon interferencyjny we wzorze (1) jest proporcjonalny do wielkości zwanej stopniem spójności interferujących wiązek. Gdy wiązki są niespójne, stopień spójności równa się zeru i człon interferencyjny znika, dla maksymalnej koherencji równy jest on jedności. Wartość stopnia spójności zawiera się w granicach od 0 do 1 i możliwe są sytuacje, w których wiązki światła są częściowo spójne. Stopień spójności można wyznaczyć, mierząc tak zwaną widzialność prążków, tj. stosunek $(I_{max}-I_{min}) : (I_{max}+I_{min})$, gdzie I_{max} i I_{min} są maksymalnymi i minimalnymi natężeniami prążków ciemnych i jasnych rejestrowanych na kliszy fotograficznej.

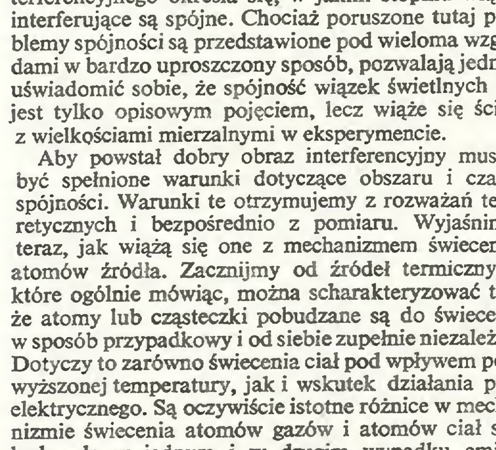
stopień spójności

W ten sposób na podstawie fotografii obrazu interferencyjnego określa się, w jakim stopniu wiązki interferujące są spójne. Chociaż poruszone tutaj problemy spójności są przedstawione pod wieloma względami w bardzo uproszczony sposób, pozwalają jednak uświadomić sobie, że spójność wiązek świetlnych nie jest tylko opisowym pojęciem, lecz wiąże się ściśle z wielkościami mierzalnymi w eksperymencie.

Aby powstał dobry obraz interferencyjny muszą być spełnione warunki dotyczące obszaru i czasu spójności. Warunki te otrzymujemy z rozważań teoretycznych i bezpośrednio z pomiaru. Wyjaśnimy teraz, jak wiążą się one z mechanizmem świecenia atomów źródła. Zacznijmy od źródeł termicznych, które ogólnie mówiąc, można scharakteryzować tak, że atomy lub cząsteczki pobudzane są do świecenia w sposób przypadkowy i od siebie zupełnie niezależny. Dotyczy to zarówno świecenia ciał pod wpływem podwyższonej temperatury, jak i wskutek działania pola elektrycznego. Są oczywiście istotne różnice w mechanizmie świecenia atomów gazów i atomów ciał stałych, ale w jednym i w drugim wypadku emisja spontaniczna (→ Lasery — podstawy działania) odgrywa najważniejszą i prawie jedyną rolę. Czas trwania impulsów wysyłanych przez pojedyncze atomy świecącego gazu wynosi około $\tau' \approx 10^{-8}$ s. Kształt takiego impulsu, który będziemy nazywać ciągiem falowym, ilustruje schematycznie rys. 4. Ciąg falowy

światło termiczne

ciąg falowy

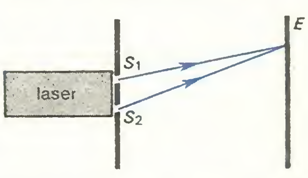


Rys. 4. Ciąg falowy o skończonym czasie trwania τ

nie jest ściśle monochromatyczny. Można go złożyć z wielu drgań sinusoidalnych, których częstości leżą w pewnym przedziale częstości $\Delta\nu$ nazywanym szerokością spektralną. Można udowodnić, że pomiędzy τ' i $\Delta\nu$ zachodzi związek $\Delta\nu \approx 1/\tau'$. Ze związku tego wynika oczywiście, że idealnie monochromatyczna

fala, dla której $\Delta\nu = 0$ miałaby czas trwania τ' nieskończenie długi. Dla światła ze źródeł termicznych można osiągnąć $\tau' \approx 10^{-8}$ s, stąd $\Delta\nu \approx 10^8\,s^{-1}$, a ponieważ w zakresie optycznym $\nu \approx 10^{15}\,s^{-1}$, więc $\Delta\nu/\nu \approx 10^{-7}$. Jak się przekonamy dalej, stosunek ten może być znacznie mniejszy dla światła laserowego. Wiedząc, że światło źródła termicznego powstaje w wyniku wysyłania ciągów falowych przez pojedyncze atomy, można zrozumieć warunki obserwacji obrazów interferencyjnych. W doświadczeniach Younga faza drgań pola elektrycznego, powstałego w wyniku nałożenia się pól ciągów falowych wysyłanych przez olbrzymią liczbę atomów źródła, jest przypadkowa. Oznacza to, że zmienia się ona nieregularnie i chaotycznie. Warunki doświadczenia są jednak tak dobrane, że zawsze na szczelinach S_1 i S_2 pola są takie same i ich fazy drgań sobie równe. W rezultacie na ekranie miejsca pełnego wygaszania i wzmocnienia są stałe w czasie, jeśli tylko, jak już dyskutowaliśmy rozpatrując punkty S' i S'' na rys. 2, rozmiary źródła są odpowiednio małe. W doświadczeniu Michelsona każdy ciąg falowy dzieli się na zwierciadle półprzepuszczalnym. Powstałe nowe dwa ciągi będą ze sobą interferować, jeśli różnica dróg optycznych w ramionach interferometru nie będzie większa niż $c\tau$. Zachodzi to dla wszystkich wysyłanych ciągów falowych i miejsca wygaszenia i wzmocnienia są stałe w czasie. Gdy powyższy warunek nie jest spełniony, interferencja pomiędzy ciągami z obydwu ramion interferometru ma charakter zupełnie przypadkowy i nie ma stałych miejsc wygaszenia i wzmocnienia. Z naszych rozważań wynika również, że wielkość czasu koherencji jest rzędu czasu trwania ciągu falowego $\tau \approx \tau'$, a stąd zachodzi $\tau \approx 1/\Delta\nu$. Mamy więc taką sytuację, że im mniejsza szerokość spektralna $\Delta\nu$, tym większy czas spójności τ. Oznacza to oczywiście, że fala ściśle monochromatyczna, dla której $\Delta\nu = 0$, miałaby czas spójności nieskończenie długi. Możemy więc teraz wyciągnąć ważny wniosek, że światło idealnie monochromatyczne jest całkowicie spójne. Znając mechanizm świecenia atomów w źródle termicznym, nie możemy oczekiwać, że uda się nam wytwarzać dowolnie długo trwające ciągi falowe, gdyż czas świecenia atomów ma zawsze pewną skończoną wartość. Ponadto dla tego rodzaju źródeł brak jest nam jakiejkolwiek kontroli aktów świecenia atomów, tak że faza drgań pola świetlnego zmienia się zupełnie przypadkowo. Zupełnie inna sytuacja jest w przypadku fal radiowych, których źródła kontrolujemy w taki sposób, że wytwarzają one drgania o dokładnie określonych przebiegach czasowych. Zbudowanie lasera, w którym atomy promieniują nie tylko w aktach emisji spontanicznej, jak dzieje się to w źródle termicznym, lecz głównie w aktach emisji wymuszonej, dało źródło światła o dużej spójności czasowej i przestrzennej. Dla światła laserowego jest możliwe przeprowadzenie eksperymentu Younga w sposób bezpośredni, przysuwając obydwie szczeliny do lasera, jak ilustruje to rys. 5, ponieważ powierzchnia spójności jest znacznie większa niż dla źródeł konwencjonalnych. Jest to konsekwencją tego, że fala świetlna wychodząca z lasera ma bardzo mały kąt rozbieżności i w pewnych warunkach może mieć tę samą fazę drgań na całym przekroju wiązki. Podobnie w interferometrze Michelsona, gdy źródłem światła jest odpowiedni laser, różnica dróg optycznych w obydwu ramionach może sięgać nawet kilku kilometrów, a otrzymywany obraz interferencyjny jest dobry. Ten ostatni fakt oznacza oczywiście, że światło lasera jest znacznie bardziej monochromatyczne niż światło ze źródeł termicznych. Stosunek szerokości spektralnej do częstości linii laserowej może osiągać $\Delta\nu/\nu \approx 10^{-11}$. W otrzymywaniu tych wyjątkowych właściwości światła laserowego istotną rolę odgrywa rezonator optyczny, wewnątrz którego znajdują się atomy promieniujące.

światło laserowe

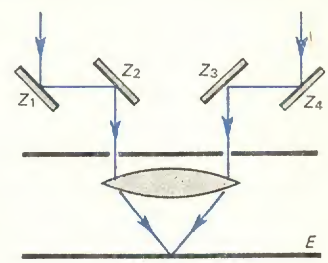


Rys. 5. Doświadczenie Younga wykazuje duży obszar spójności światła laserowego

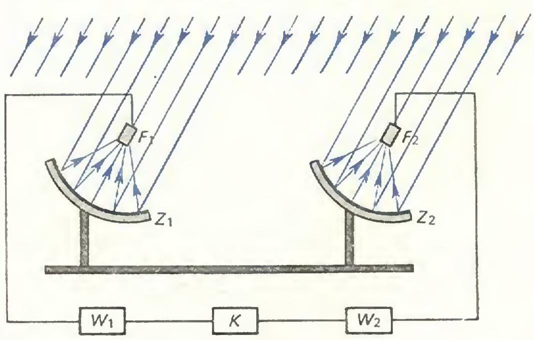
Spójność wyższego rzędu

Na przykładzie doświadczenia Younga i interferometru Michelsona zilustrowaliśmy istotne różnice w spójności światła termicznego i laserowego. Różnice te, polegające na tym, że czas i obszar spójności są różne dla obydwu rodzajów źródeł, dotyczą spójności pierwszego rzędu, a więc takich zjawisk, których opis wymaga uśrednienia kwadratów natężenia pola elektrycznego lub wyrażenia $\langle E_1 E_2 \rangle$, które występuje we wzorze (1). W tym miejscu nasuwają się dwa pytania: po pierwsze, czy istnieją inne różnice między światłem termicznym i laserowym, po drugie, w jakich doświadczeniach moglibyśmy te różnice wykryć? Pierwsze doświadczenia nie związane bezpośrednio ze zjawiskiem interferencji, w których można by stwierdzić różnice inne od tych, które omawialiśmy, przeprowadzono jeszcze przed zbudowaniem lasera. Zanim je krótko omówimy, przypomnijmy zasadę pomiarów średnic kątowych gwiazd, które po raz pierwszy przeprowadzał Michelson już w końcu ubiegłego stulecia. Ideę zasadniczą ilustruje rys. 6, na którym zaznaczono, w jaki sposób światło gwiazdy pada na układ zwierciadeł i następnie interferuje. Rozsuwając zwierciadła Z_1 i Z_2 można stwierdzić, że dla pewnej odległości między nimi obraz interferencyjny znika. Odległość ta określa wymiary obszaru spójności światła gwiazdy. Na podstawie znajomości średniej długości fali światła padającego i wymiarów obszaru koherencji ścisła teoria pozwala określić kąt, pod jakim widać odległe źródło, a więc określić kątową średnicę gwiazdy. W 1956 r. Hanbury-Brown i Twiss zmodyfikowali powyższe doświadczenie interferencyjne, tak jak pokazuje to schematycznie rys. 7. Światło gwiazdy pada na dwa fotodetektory umieszczone w ogniskach zwierciadeł parabolicznych. Posługując się odpowied-

pomiar średnic kątowych gwiazd



Rys. 6. Schemat interferometru Michelsona do pomiaru średnic kątowych gwiazd. Zwierciadła Z_1 i Z_4 można przesuwać



Rys. 7. Światło gwiazdy pada na zwierciadła paraboliczne Z_1 i Z_2, w których ogniskach znajdują się fotodetektory F_1 i F_2. Prądy fotoelektryczne są wzmacniane w układach W_1 i W_2. Korelator K mierzy uśrednioną wartość iloczynu $I_1 I_2$

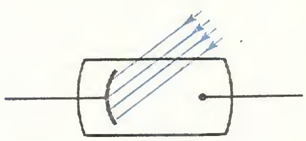
nim układem elektronicznym, bada się korelacje statystyczne natężeń fotoprądów, tzn. wielkość $\langle i_1 i_2 \rangle$, w zależności od odległości. Ponieważ natężenia prądów i_1 oraz i_2 są odpowiednio proporcjonalne do natężeń światła padającego $I_1' = E_1^2$ oraz $I_2' = E_2^2$, w istocie w doświadczeniu można wyznaczyć $\langle I_1' I_2' \rangle$. Przyjęto mówić, że eksperyment Hanbury-Browna i Twissa dotyczy koherencji drugiego rzędu. Wynika to stąd, że dla spójności pierwszego rzędu w doświadczeniach Younga i Michelsona badaliśmy wyrażenia typu $\langle E_1^2 \rangle$, $\langle E_1 E_2 \rangle$, a teraz badamy wyrażenia typu $\langle E_1^2 E_2^2 \rangle$. Na podstawie zależności $\langle I_1' I_2' \rangle$ od odległości między fotodetektorami Hanbury-Brown i Twiss zmierzyli średnicę gwiazd o małych wymiarach kątowych rzędu 0,006″ (i dostatecznie dużej jasności), co nie byłoby możliwe przy wykorzystaniu tradycyjnego interferometru Michelsona. W ten sposób powstały badania zwane interferometrią natężeniową.

interferometria natężeniowa

Lata późniejsze przyniosły liczne modyfikacje powyższego eksperymentu, którego celem było udzielenie odpowiedzi czy korelacje wyższych rzędów, a więc średnie wartości wyrażeń na przykład typu $\langle E_1^2 E_2^2 \rangle$, są różne dla światła termicznego i laserowego. Aby obliczyć wartość średnią $\langle I_1' I_2' \rangle = \langle E_1^2 E_2^2 \rangle$, trzeba znać odpowiednie rozkłady prawdopodobieństwa. W niektórych przypadkach ten sam wynik otrzymuje się dokonując uśrednienia po czasie, gdy znane są zależności E_1 i E_2 od czasu. Wyniki badań teoretycznych i doświadczalnych wykazują, że wiązki światła laserowego i termicznego spójne w sensie koherencji pierwszego rzędu (tzn. dla każdej z nich możemy obserwować ostre obrazy interferencyjne) charakteryzują się różnymi wartościami wyrażeń typu $\langle E_1^2 E_2^2 \rangle$. Oznacza to, między innymi, że rozkłady prawdopodobieństwa są różne dla światła termicznego i laserowego. Tak więc istota różnic między światłem termicznym i laserowym leży nie tyle w różnych wartościach czasu i obszaru koherencji, lecz w ich właściwościach statystycznych. Spośród wszystkich eksperymentów potwierdzających te różnice zbadanie statystyki fotonów optycznych należy do najbardziej oryginalnych i interesujących. Omówimy teraz zasadniczą ideę tego rodzaju doświadczeń.

Statystyczne własności światła

Załóżmy, że na fotokatodę pada światło prawie monochromatyczne o szerokości spektralnej $\Delta\nu$, tak jak schematycznie pokazuje to rys. 8. Z praw zjawiska



Rys. 8. Światło pada na katodę. Odpowiedni układ elektroniczny zlicza wybite elektrony

zjawisko fotoelektryczne

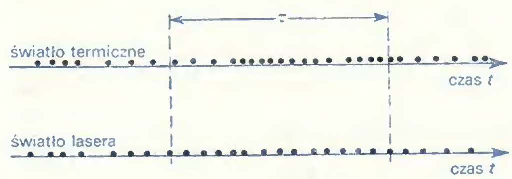
fotoelektrycznego wiemy, że fotoprąd (liczba elektronów wybijanych z katody w ciągu jednostki czasu) jest proporcjonalny do natężenia światła padającego, a więc do liczby fotonów, które w ciągu jednostki czasu padają na katodę. Cały proces ma charakter probabilistyczny, bowiem zjawiskiem wybicia elektronu z atomu rządzą prawa mechaniki kwantowej, które pozwalają obliczać jedynie prawdopodobieństwo wybicia elektronu. Ponadto światło padające może mieć różne właściwości statystyczne, to znaczy fluktuacje (statystyczne odchylenia od wartości średniej) natężenia światła mogą podlegać różnym rozkładom prawdopodobieństwa, co może w różny sposób wpływać na statystykę wybijanych elektronów. Oznacza to, że natężenie prądu płynącego w obwodzie z fotodetektorem zmienia się w sposób przypadkowy, istotnie jednak zależy od zmian natężenia światła.

Rozwój elektroniki przyniósł odpowiednio szybkie i czułe fotomnożniki oraz specjalne układy, które pozwalają zliczać fotoelektrony w ciągu dostatecznie krótkiego czasu T. Jeśli w odpowiednio dużej liczbie prób N, powiedzmy kilkunastu tysiącach, stwierdzimy, że zarejestrowano l_1 razy n_1 fotoelektronów, to prawdopodobieństwo tego wydarzenia $p(n_1, T) = l_1/N$. Przez $p(n_1, T)$ oznaczyliśmy tutaj prawdopodobieństwo zarejestrowania n_1 fotoelektronów w czasie T. Podobnie prawdopodobieństwo zarejestrowania n_2 fotoelektronów $p(n_2, T) = l_2/N$, gdzie l_2 oznacza ile razy spośród N pomiarów zliczono n_2 fotoelektronów. W ten sposób można wyznaczyć funkcję $p(n, T)$ dla różnych n. W ścisłej teorii dowodzi się, że jeśli czas T jest mniejszy niż czas spójności $\tau \approx 1/\Delta\nu$, to liczba zarejestrowanych fotoelektronów n jest proporcjonalna do liczby padających fotonów. Tak więc istnieje możliwość bezpośredniego zliczania fotonów i badania ich statystyki. Przeprowadzanie eksperymentów przy $T < \tau$ jest niemożliwe dla światła termicznego, dla którego $\tau \approx 10^{-8}$ s; nie jest to jednak trudne dla światła z lasera gazowego pracującego w sposób ciągły, dla którego na przykład można osiągnąć czas spójności $\tau \approx 10^{-2}$ s.

zliczanie fotonów

Można jednak otrzymać światło termiczne o dostatecznie długim czasie koherencji metodą „psucia spójności” światła laserowego. Również światło z lasera, który świeci przed progiem wzbudzenia właściwej akcji laserowej, ma właściwości światła termicznego o dostatecznie długim czasie spójności.

Eksperymenty, w których zlicza się fotony, przeprowadzane w latach sześćdziesiątych w wielu ośrodkach badań optycznych, należą do najważniejszych



Rys. 9. Rozkłady gęstości fotonów w świetle termicznym i laserowym, zliczone w czasie krótszym niż τ, są różne

osiągnięć optyki kwantowej. Wykazały one, że $p(n, T)$ dla $T < \tau$ jest różne dla światła termicznego i laserowego. Pomiary ściśle potwierdziły wyniki teorii, z której wynika, iż $p(n, T)$ dane jest dla światła termicznego wzorem zwanym w fizyce statystycznej wzorem Bosego–Einsteina, a $p(n, T)$ dla światła z lasera gazowego jednomodowego określa wzór Poissona. Nie podajemy tutaj tych wzorów, ale różnicę między nimi w uproszczony sposób ilustruje rys. 9. Różne statystyki fotonów zilustrowane są przykładem losowego rozrzucenia punktów na osi czasu. W świetle laserowym fotony są rozłożone całkowicie przypadkowo i w sensie statystycznym równomiernie. Dla światła termicznego rozkład nie jest równomierny, bowiem w sensie statystycznym pojawiają się miejsca zgęszczeń i rozrzedzeń. Pełna teoria opisuje również przypadki pośrednie, w których mamy do czynienia z nałożeniem się światła termicznego i laserowego. Ten fakt został również potwierdzony wynikami pomiarów.

statystyki fotonów światła termicznego i laserowego

Omówione tutaj doświadczenia doczekały się wielu bardzo różnych modyfikacji. Badano na przykład problem, jakie jest prawdopodobieństwo warunkowe zarejestrowania fotonu, jeśli wcześniej o czas Δt również zarejestrowano foton. Zależność tego prawdopodobieństwa od czasu Δt pozwala wykazać istnienie różnych właściwości światła termicznego i laserowego. Eksperymenty, w których zlicza się fotony są też sprawdzeniem teorii działania lasera gazowego, a w szczególności progu akcji laserowej, przy którym właśnie następuje zmiana właściwości statystycznych światła.

Do aktualnych problemów badawczych należy wyjaśnienie, dlaczego zjawiska optyczne nieliniowe, np. wytwarzanie drugiej harmonicznej, przejścia wielofo-

tonowe (→ Optyka nieliniowa) zależą od właściwości statystycznych światła, dla którego zjawiska te obserwujemy. Doświadczenia wykazały, że przebieg zjawisk nieliniowych może być różny dla światła laserowego i światła termicznego. Trzeba tutaj przypomnieć, że obserwacja zjawiska nieliniowego wymaga dużego natężenia światła padającego i z tego powodu jest nim prawie zawsze światło lasera.

Można jednak zmienić statystyczne cechy światła laserowego na cechy światła termicznego poprzez „psucie spójności" wskutek rozpraszania na szorstkich poruszających się powierzchniach lub poprzez wzbudzenie lasera do świecenia w warunkach pracy z nałożeniem się dużej liczby niezsynchronizowanych modów (→ Ultrakrótkie impulsy światła). Tak otrzymana wiązka świetlna ma dostatecznie duże natężenie do wywołania zjawisk nieliniowych, a jednocześnie jest podobna pod względem właściwości statystycznych do światła termicznego. Czasami nazywa się takie światło kwazitermicznym. Okazuje się, że przy obserwacji zjawisk nieliniowych skuteczność światła termicznego jest większa niż światła laserowego, np. wiązka światła laserowego wytwarza drugą harmoniczną słabiej niż wiązka światła termicznego o takim samym natężeniu. W bardzo uproszczony i jakościowy sposób można tłumaczyć to faktem, że fluktuacje gęstości fotonów, a więc i fluktuacje gęstości mocy (mierzonej w W/cm^2), są większe dla światła termicznego, jak to schematycznie ilustruje rys. 9. Ponieważ zjawiska nieliniowe silnie zależą od gęstości mocy, światło termiczne o większych zgęszczeniach okazuje się bardziej skuteczne.

światło kwazitermiczne

Na zakończenie zwróćmy uwagę na ważny fakt, że badania zliczeń fotonów optycznych dotyczą statystycznych właściwości światła z uwzględnieniem jego natury kwantowej. Doświadczenia, w których liczono fotony wykazały bowiem, iż dwie prawie zupełnie monochromatyczne wiązki światła, np. światła laserowego i termicznego, identyczne pod względem możliwości tworzenia ostrych obrazów interferencyjnych, wciąż istotnie różnią się między sobą „strukturą fotonową". We współczesnym opisie wyraża się to stwierdzeniem, że stany kwantowe pola tych wiązek są różne. Wytwarzanie światła o różnych właściwościach statystycznych lub, mówiąc dokładniej, o różnych cechach kwantowych zależy oczywiście od tego, jak zbudowane jest źródło, a ściślej — jaki jest mechanizm świecenia jego atomów.

natura kwantowa światła

W pierwszej części tego artykułu przy omawianiu zagadnień spójności posługiwaliśmy się jedynie falowymi cechami światła, które z bardzo dobrym przybliżeniem można opisać w ramach klasycznej teorii fal elektromagnetycznych bez wprowadzania pojęcia fotonu. Kwantowy opis zjawisk, w którym spójność — lub ogólniej właściwości statystyczne — gra istotną rolę, wymaga odpowiedniego aparatu matematycznego, jakim posługuje się elektrodynamika kwantowa. Nie poruszono tutaj zupełnie kwestii matematycznego formalizmu kwantowej teorii koherencji. Scharakteryzowano jedynie te najważniejsze wyniki eksperymentalne, które wzbogaciły naszą wiedzę o falowo-korpuskularnej naturze światła, a w szczególności istotnie rozszerzyły tradycyjne pojęcie spójności wiązek świetlnych stosowane przy opisie zjawisk interferencyjnych oraz stały się podstawą wielu kierunków badań współczesnej optyki.

H. Klejman *Lasery*, Warszawa 1974; J. L. Klimontowicz *Lasery i optyka nieliniowa*, Warszawa 1969; A. Kujawski, J. Mostowski *Promieniowanie elektromagnetyczne*, Warszawa 1977; A. Piekara *Nowe oblicze optyki*, Warszawa 1976; *Światło* (zbiór artykułów), Warszawa 1973.

Lasery – podstawy działania

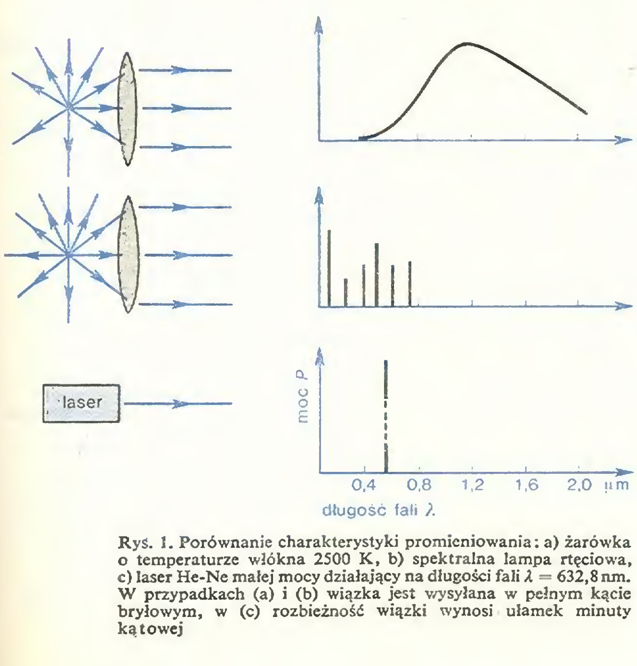
Tadeusz Skaliński

W 1960 r., po wieloletnich próbach, w kilku laboratoriach świata uruchomiono źródła promieniowania optycznego zupełnie nowego typu — lasery. Promieniowanie wysyłane przez te źródła charakteryzuje daleko posunięta monochromatyczność, kierunkowość rozchodzenia się, spójność i znaczna gęstość mocy w porównaniu z promieniowaniem otrzymywanym ze źródeł klasycznych (rys. 1 i tabela str. 357).

a)

b)

c)



Rys. 1. Porównanie charakterystyki promieniowania: a) żarówka o temperaturze włókna 2500 K, b) spektralna lampa rtęciowa, c) laser He-Ne małej mocy działający na długości fali $\lambda = 632{,}8$ nm. W przypadkach (a) i (b) wiązka jest wysyłana w pełnym kącie bryłowym, w (c) rozbieżność wiązki wynosi ułamek minuty kątowej

własności światła laserowego

Spójność promieniowania z lasera (→ Spójność światła) jest tak wielka, że przy jego użyciu można wykonać doświadczenie Younga, otrzymując bez wstępnej szczeliny piękny obraz interferencyjny. Doświadczenie z interferometrem Michelsona wskazuje na czasy spójności rzędu 1 μs (długość spójności, czyli różnica dróg optycznych interferujących wiązek wynosi kilkaset metrów). W laserach gazowych specjalnej konstrukcji jest możliwe uzyskanie szerokości linii widmowej około 2 Hz; odpowiada to spójności przy różnicy dróg optycznych 150 000 km! Czynnikami ograniczającymi spójność promieniowania są: niejednorodność ośrodka, w którym się rozchodzi promieniowanie, i drgania mechaniczne, na które narażone są poszczególne elementy układu laserowego.

Perspektywa praktycznych i badawczych zastosowań laserów sprawiła, że prace w tej dziedzinie podjęły setki laboratoriów, a minione lata przyniosły olbrzymi postęp zarówno w konstruowaniu tych urządzeń, jak i w zrozumieniu podstawowych procesów związanych z ich działaniem i wykorzystaniem. Prowadzone są prace nad objęciem przez promieniowanie z laserów coraz szerszych obszarów widmowych (aż do obszaru rentgenowskiego włącznie), nad uzyskaniem coraz większych gęstości mocy promieniowania, coraz krótszych impulsów itp. Należy tu jednak podkreślić, że wiele z tych wymienionych właściwości ma charakter przeciwstawny (np. bardzo silnemu zwężeniu spektralnemu linii widmowej towarzyszy silny spadek mocy promieniowania, krańcowym zaś skróceniom czasu impulsu — duże rozszerzenie obszaru widmowego).

Fizyczne zasady działania laserów

Najprostszy układ kwantowy — mikroukład (atom lub cząsteczka), w którym mogą zachodzić procesy absorpcji i emisji promieniowania, ma dwa dyskretne stany energetyczne: stan podstawowy 1 o energii W_1 i stan wzbudzony 2 o energii W_2, przy tym $W_2 > W_1$.

kwantowy układ dwupoziomowy

Przyjmujemy, że przejścia promieniste między tymi stanami są dozwolone przez reguły wyboru, a częstość promieniowania z nimi związana wynosi $\nu_0 = \Delta W/h = (W_2 - W_1)/h$ (→ Spektroskopia atomowa). Takie uproszczenie można przyjąć z dobrym przybliżeniem dla każdego układu kwantowego o większej liczbie poziomów, jeśli oddziaływanie wybranych dwóch poziomów z wszystkimi pozostałymi jest tak niewielkie, że można je pominąć. W ośrodku złożonym z bardzo wielkiej liczby N (w 1 cm³) identycznych mikroukładów rozkład obsadzeń w warunkach równowagi termodynamicznej między stan 1 (N_1) i 2 (N_2) jest opisany zależnością $N_2/N_1 = e^{-\Delta W/kT}$ (prawo Boltzmanna); $k = 1{,}38 \cdot 10^{-23}$ J/K oraz $N_1 \gg N_2$.

prawo Boltzmanna

Zgodnie z tym prawem w parze sodu w temperaturze 400 K rozkład obsadzeń między stan podstawowy atomów ($3S$) i rezonansowy ($3P$) o energii 2,09 eV jest określony stosunkiem $N_2/N_1 \approx 10^{-27}$, w temperaturze zaś 2100 K stosunek ten jest równy 10^{-5}. Widać że w stanie równowagi termodynamicznej w niezbyt wysokiej temperaturze obsadzenie wzbudzonych stanów atomowych jest znikome.

W 1917 r. Einstein, analizując wyrażenie na rozkład natężeń w widmie ciała doskonale czarnego, doszedł do wniosku, że aby uzyskać zgodność opisu teoretycznego z doświadczeniem, należy oprócz absorpcji i emisji samorzutnej promieniowania wprowadzić pojęcie emisji wymuszonej — procesu symetrycznego w stosunku do absorpcji i zależnego od natężenia promieniowania znajdującego się w obszarze, w którym się znajduje atom wzbudzony; częstość promieniowania wymuszającego musi być równa możliwej częstości przejścia atomu. Bilans obsadzeń stanów w przypadku równowagi można zapisać następująco:

emisja wymuszona

$$B_{12} N_1 u(\nu)_{\nu=\nu_0} = B_{21} N_2 u(\nu)_{\nu=\nu_0} + A_{21} N_2.$$

$B_{12} = B_{21}$ są to współczynniki Einsteina absorpcji i emisji wymuszonej, A_{21} jest współczynnkiem emisji samorzutnej, $u(\nu)_{\nu=\nu_0}$ jest gęstością promieniowania o częstości ν_0. Lewa strona równości przedstawia więc (odniesioną do 1 cm³) liczbę procesów absorpcji kwantów o energii $h\nu_0$, prawa — liczbę procesów emisji wymuszonej i samorzutnej łącznie (rys. 2).

Z analizy przeprowadzonej przez Einsteina wynika, że promieniowanie wymuszone winno być zgodne w fazie i kierunku rozchodzenia się z promieniowaniem wymuszającym (powinno być z nim spójne). Tej własności nie ma promieniowanie samorzutne o izotropowym rozkładzie przestrzennym i chaotycznie rozłożonych fazach drgania. W zwykłych warunkach — wobec $N_1 \gg N_2$ i bardzo niewielkich gęstości mocy promieniowania $u(\nu)$ w wąskim przedziale częstości obejmującym ν_0 — udział emisji wymuszonej jest znikomy i może być pominięty.

Gdy promieniowanie (monochromatyczne i tworzące równoległą wiązkę) przechodzi przez ośrodek pochłaniający, jego natężenie maleje w miarę przenikania promieniowania w głąb ośrodka. Tę zmianę natężenia opisuje prawo Lamberta $I = I_0 e^{-\varkappa x}$; I_0 jest natężeniem początkowym, I — natężeniem po przebyciu warstwy grubości x, $\varkappa$ jest współczynnikiem absorpcji promieniowania w ośrodku (w ogólności $\varkappa = \varkappa(\nu)$ — zależy on od częstości fali). W zwykłych warunkach $\varkappa \sim N_1$. Gdy jednak uwzględnimy istnienie emisji wymuszonej (a emisję samorzutną pominiemy, bo przy jej izotropowym rozkładzie jej udział w określonym kierunku jest znikomy), to wówczas

prawo absorpcji (Lamberta)

$$\varkappa(\nu) \sim N_1 (1 - N_2/N_1).$$

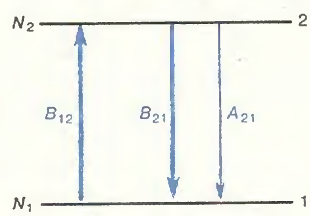
Jeśli więc będziemy powiększali obsadzenie stanu o większej energii, udział emisji wymuszonej będzie wzrastał, a $\varkappa(\nu)$ będzie malało. Gdyby się nam udało uzyskać tak znaczne obsadzenie stanu 2, by $N_2/N_1 > 1$ wówczas $\varkappa(\nu)$ stałoby się mniejsze od zera i zamiast osłabienia wiązki nastąpiłoby jej wzmocnienie. Wiązka miałaby przy tym właściwości, jakie Einstein przewidział dla promieniowania wymuszonego. Oczywiście nie ma tu sprzeczności z zasadą zachowania energii, bo uzyskanie stanu odwrócenia obsadzeń (tak nazywamy stan, w którym obsadzenie poziomu o wyższej energii jest większe od obsadzenia poziomu o niższej energii) wymaga dostarczenia energii na przepompowanie atomów z 1 do 2. Łatwo zauważyć (w równaniu bilansu w stanie równowagi $B_{12} = B_{21}$ i $A_{21} > 0$), że pompowanie w układzie dwupoziomowym nie może doprowadzić do odwrócenia obsadzeń. Jest to możliwe w układach bardziej złożonych, których rozważenie zaproponowali Ch. H. Townes i A. L. Schawlow w 1958 r.

stan odwrócenia obsadzeń

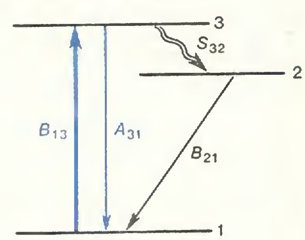
Zatem warunki, jakie winny być spełnione, by uzyskać możliwość wytwarzania promieniowania spójnego w ośrodku czynnym, są następujące: wytworzenie w ośrodku czynnym stanu odwrócenia obsadzeń (wystarczającego na to, by wzmocnienie wiązki przewyższyło możliwe straty jej natężenia) oraz wstępne naświetlenie ośrodka czynnego promieniowaniem wymuszającym o gęstości mocy dostatecznej na to, by efektywnie spowodować emisję wymuszoną w określonym, celowo wybranym kierunku. Pierwszy warunek realizuje się przez pompowanie ośrodka czynnego, drugi — przez zastosowanie komory rezonansowej (rezonatora).

Metoda, którą się uzyskuje stan odwrócenia obsadzeń, nawiązuje do technik stosowanych i sprawdzonych w mikrofalowych generatorach kwantowych — maserach (→ Spektroskopia mikrofalowego rezonansu rotacyjnego). Niezbędne jest w tym celu użycie układu mającego co najmniej 3 dyskretne poziomy energetyczne (rys. 3), co jest typowe dla znacznej liczby substancji fluoryzujących (są to albo substancje zawierające centra utworzone przez domieszki jonów metali w osnowach krystalicznych i szkliwach, albo roztwory barwników organicznych). Błysk lampy wyładowczej przenosi przeważającą część centrów czynnych ze stanu 1 do 3 (absorpcja promieniowania), z którego oprócz powrotu promienistego do stanu 1 możliwe jest bardzo szybkie — w porównaniu z przejściem 3 → 1 — przejście bezpromieniste do stanu 2, do którego zostaje przeniesiona przeważająca część atomów wzbudzonych (energia uwolniona w przejściach bezpromienistych zamienia się w energię ciepl-

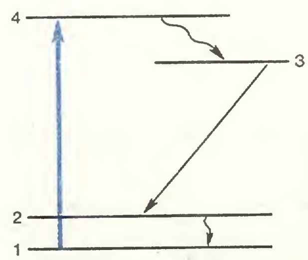
pompowanie optyczne



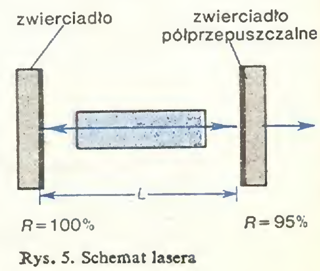
Rys. 2. Przejścia w układzie dwupoziomowym



Rys. 3. Przejścia w układzie trójpoziomowym



Rys. 4. Przejścia w układzie czteropoziomowym



Rys. 5. Schemat lasera

ną). Ze stanu 2 następuje przejście 2 → 1 (fluorescencja) z niewielką szybkością, zależną od natury centrów, np. w różowym rubinie (domieszka 0,05% Cr^{3+} w osnowie kryształu Al_2O_3), który jest typowym materiałem laserowym, czas życia poziomu 2 wynosi 3 ms. W ten sposób, przy dostatecznej mocy błysku, możliwe jest uzyskanie odwrócenia obsadzeń między stanami 2 i 1. Tego typu układy działają w sposób impulsowy. Efektywność działania można zwiększyć przez zastosowanie substancji o układzie 4 poziomów (np. jony uranu lub samaru w osnowie kryształu fluorytu), gdyż w takim wypadku fluorescencja jest wywołana przejściem między stanami 3 i 2, a z dolnego stanu 2 następuje szybkie bezpromieniste przejście do stanu podstawowego (rys. 4). W takim układzie można uzyskać stacjonarne odwrócenie stanów 3 i 2.

rezonator

Ośrodek czynny lasera w postaci walca umieszcza się w rezonatorze, którego zasadniczą częścią jest układ dwóch zwierciadeł (płaskich lub sferycznych). Zwierciadła zmuszają promieniowanie zapoczątkowane w pompowanym ośrodku czynnym do rozchodzenia się w nim wzdłuż osi walca wielokrotnie tam i z powrotem. Promieniowanie to wymusza przejścia promieniste wzbudzonych centrów i w ten sposób się wzmacnia (rys. 5). Jedno ze zwierciadeł maksymalnie odbija padające na nie promieniowanie (współczynnik odbicia $R \approx 100\%$), drugie — ma zdolność odbijającą nieco mniejszą ($R \approx 95\%$) i niewielką transmisję ($T \approx 5\%$), co pozwala na wyprowadzenie części promieniowania na zewnątrz. W rzeczywistości przy każdym odbiciu zachodzą pewne straty i wzmocnienie przez napompowany ośrodek czynny winno być wystarczająco duże, by pokryło zarówno straty przy odbiciu, jak i odpływ energii promieniowania na zewnątrz rezonatora.

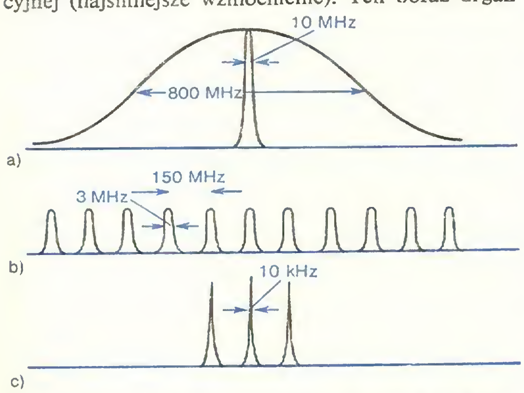
Aby działanie wymuszające promieniowanie było jak największe, rezonator winien być dostrojony do wzmacnianej długości fali (oprócz wyjustowania zwierciadeł ściśle prostopadle do osi rezonatora).

warunek rezonansu

Dostrojenie do rezonansu jest wtedy, gdy w obszarze między zwierciadłami rezonatora zawiera się całkowita liczba połówek długości fali światła: $2L/\lambda = n$. Długość rezonatora L jest rzędu kilkudziesięciu cm, a λ jest ułamkiem µm, a więc liczba całkowita n jest ogromna (ok. 10^6) i w szerokości linii fluorescencji (wynoszącej np. w rubinie ok. 0,2 nm, tj. ok. 100 GHz, a w laserze gazowym He-Ne ok. 0,0015 nm, tj. 0,8 GHz) znajdzie się przy określonym L kilka, a czasem kilkanaście możliwych długości fali spełniających warunek rezonansu (rys. 6), można też dla nich wzbudzić kilka lub kilkanaście rodzajów drgań,

mody drgań

czyli modów (łac. *modus* 'rodzaj, sposób'; → Ultrakrótkie impulsy światła). Najsilniej wzbudzają się drgania w pobliżu maksimum rozkładu linii absorpcyjnej (najsilniejsze wzmocnienie). Ten obraz drgań



Rys. 6. a) Linia emisyjna Ne (λ = 632,8 nm), szerokość dopplerowska linii 800 MHz, szerokość naturalna linii 10 MHz; b) rozkład częstości rezonansowych komory (przyjęto L = 1 m); szerokości instrumentalne rezonansów (ok. 3 MHz) zależne są od współczynników odbicia zwierciadeł komory i dokładności wykończenia ich powierzchni; c) oscylacje wzbudzają się tam, gdzie przypada obszar maksimum fluorescencji, i mogą zachodzić na kilku częstościach rezonansowych

w rezonatorze jest bardzo uproszczony z powodu uwzględnienia tylko osiowych modów drgania. Ponieważ w rzeczywistości drgania zachodzą w trójwymiarowym rezonatorze o symetrii cylindrycznej, obraz jest o wiele bardziej złożony.

Przejście od fluorescencji do wymuszonej generacji promieniowania przejawia się w skokowej zmianie rozbieżności wiązki wysłanej z ośrodka czynnego: od rozkładu izotropowego (równomiernego we wszystkich kierunkach) do postaci niemal doskonale równoległego promienia. Przejściu temu towarzyszy bardzo silne zwężenie linii widmowej (rys. 6c).

próg generacji

Przy pompowaniu metodą impulsową oscylacje ośrodka czynnego rozpoczynają się z chwilą, gdy zostaje przekroczony próg generacji (wzmocnienie większe niż straty), a trwają one tak długo, jak długo pompowanie utrzymuje niezbędne do tego odwrócenia obsadzeń. Natomiast w układzie o pompowaniu ciągłym ustala się stan równowagi, w którym wzmocnienie uzyskane w ośrodku pompowanym kompensuje straty (w które wliczamy energię wyprowadzonej na zewnątrz wiązki promieniowania).

Działanie wybranych typów laserów

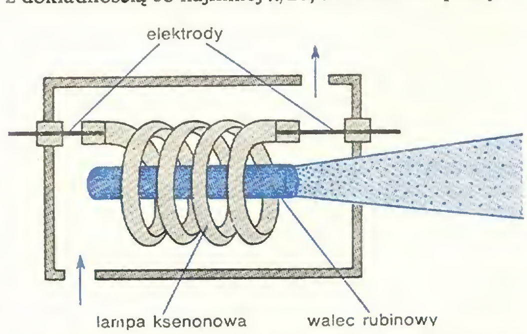
Mimo że od uruchomienia pierwszych laserów nie upłynęło jeszcze 20 lat, technika laserowa poczyniła w tym czasie olbrzymie postępy. Skonstruowano dziesiątki typów różnych laserów, uzyskano efekt laserowy w setkach różnych ośrodków czynnych (gazowych, ciekłych i stałych), a liczba zidentyfikowanych przejść laserowych wynosi zapewne kilkanaście tysięcy. W dalszym ciągu artykułu omówimy najbardziej powszechne typy laserów, a mianowicie lasery krystaliczne — rubinowy i neodymowy, laser gazowy He-Ne i lasery barwnikowe, wspomnimy także o laserach chemicznych. Lasery półprzewodnikowe omówione są w artykule „Optoelektronika półprzewodnikowa", natomiast otrzymywanie impulsów gigantycznych oraz bardzo krótkich opisane jest w artykule „Ultrakrótkie impulsy światła".

Laser rubinowy i neodymowy

ośrodki czynne laserów

Ośrodek czynny lasera krystalicznego ma postać cylindrycznego pręta o średnicy ok. 1 cm i długości do kilkunastu cm ustawionego między zwierciadłami.

W laserze rubinowym (rys. 7) osnową jest kryształ szafiru (Al_2O_3) domieszkowany jonami Cr^{3+}, w neodymowym — szkło lub kryształy CaF_2, $CaWO_4$, YAG (granat itrowo-glinowy) i inne domieszkowane jonami Nd^{3+}. Monokryształy stanowiące osnowę muszą wykazywać wielką doskonałość struktury. Pręty wycina się tak, że oś walca tworzy z osią optyczną kryształu określony kąt, zadany warunkami doświadczenia. Ścianki czołowe pręta są oszlifowane z dokładnością co najmniej $\lambda/20$, a nierzadko pokryte



Rys. 7. Schemat lasera rubinowego; strzałki oznaczają obieg gazu chłodzącego

są warstwą odbijającą tworzącą rezonator (w pewnych typach doświadczeń konieczne jest jednak używanie zwierciadeł zewnętrznych).

Pompowanie ośrodka czynnego prowadzi się za pomocą wyładowczych lamp błyskowych wypełnionych ksenonem. Błysk uzyskuje się rozładowując przez lampę baterię kondensatorów o pojemności od kilkuset do kilku tysięcy μF przy napięciu kilku kV. W celu jak najlepszego wykorzystania strumienia świetlnego lampie nadaje się kształt linii śrubowej owijającej pręt i dodatkowo otacza się ją osłoną pokrytą chromem odbijającym promieniowanie. W innych układach stosuje się oświetlacze skupiające na pręcie promieniowanie wysyłane przez podłużną lampę wyładowczą. Mimo tych zabiegów zaledwie drobna część energii elektrycznej dostarczonej do lampy zostaje efektywnie przemieniona w energię wypromieniowaną przez laser (mniej niż 0,1%); przeważająca część energii zamienia się w ciepło, co narzuca konieczność silnego chłodzenia układu.

laser rubinowy

Układ poziomów energetycznych jonu Cr^{3+} w rubinie odpowiada schematowi z rys. 3. Korzystną sytuację stwarza okoliczność, że absorpcja promieniowania pompującego zachodzi w stosunkowo szerokim obszarze widmowym (poziom 3 należy zastąpić szerokim ciągłym pasmem). Mimo to, obszar widmowy absorpcji pokrywa zaledwie drobną część widma wysyłanego przez lampę. Po przekroczeniu progu generacji zostaje wypromieniowany w błysku krótszym niż 1 ms impuls laserowy odpowiadający fluorescencji R_1 rubinu (λ = 694,3 nm). Przebieg czasowy wzbudzonej akcji laserowej jest na ogół bardzo złożony, składa się z szeregu kolejnych rozbłysków odpowiadających różnym modom drgań rezonatora (w miarę dopompowywania ośrodka czynnego przez lampę).

laser neodymowy

Pompowanie lasera neodymowego zachodzi w warunkach dogodniejszych — wg schematu na rys. 4. Wobec tego, że różnica energii między poziomami 2 i 1 równa 2000 cm^{-1} (tj. $4 \cdot 10^{-6}$J) w temperaturze pokojowej (300 K) jest rzędu kT, uzyskanie odwrócenia obsadzeń między poziomami 3 i 2 jest znacznie łatwiejsze niż w rubinie, a ochłodzenie ośrodka czynnego do temperatury ciekłego azotu (77 K) pozwala na uzyskanie pracy ciągłej tego lasera na długości fali 1,05 μm przy niezbyt wysokiej mocy (ok. 1 mW) Laser Nd^{3+} w YAG pozwala na uzyskanie w impulsie olbrzymich mocy. W podobny sposób jak laser neodymowy działają lasery, w których w różnych osnowach krystalicznych centrami są jony metali ziem rzadkich i uranu.

Lasery gazowe

Odwrócenie obsadzeń poziomów jako przygotowanie do akcji laserowej w gazach może być uzyskane przez wyładowanie elektryczne. Ogromne znaczenie mają wówczas atomy w stanach metatrwałych (→ Spektroskopia atomowa), ich energia może być przekazana w zderzeniach atomom lub cząsteczkom właściwego ośrodka laserującego. Tak jest właśnie w laserze He-Ne, w którym ciałem roboczym jest mieszanina helu i neonu o ciśnieniu cząstkowym helu ok. 130 Pa i neonu ok. 13 Pa.

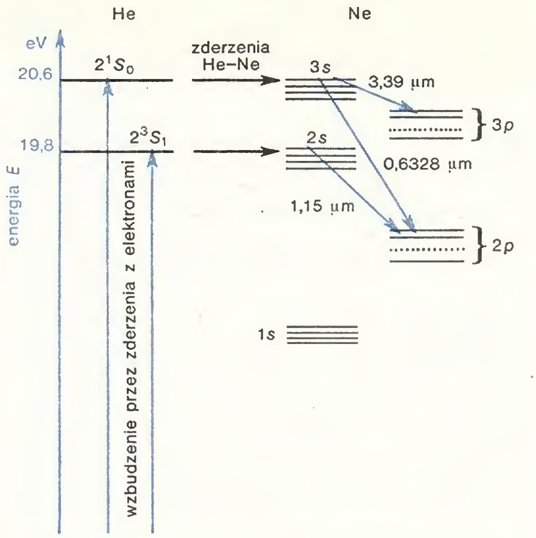
laser helowo-neonowy

Wyładowanie elektryczne prowadzone w tej mieszaninie wzbudza atomy helu i neonu do różnych stanów. Najważniejsze jednak dla uzyskania akcji laserowych jest wzbudzenie atomów helu do stanów metatrwałych 2^1S_0 (o energii 20,61 eV) oraz 2^3S_1 (o energii 19,82 eV). Energia pierwszego z wymienionych stanów niemal dokładnie odpowiada energii wzbudzenia stanu $3s_2$ neonu, drugiego — energii stanu $2s_2$ (rys. 8). Wobec tego wzbudzenie może być bardzo efektywne w zderzeniach przebiegających według symbolicznych równań:

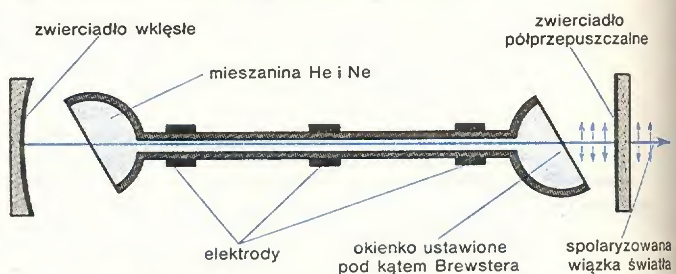
$$He(2^1S_0)+Ne(1^1S_0) \rightarrow He(1^1S_0)+Ne(3s_2)$$

$$He(2^3S_1)+Ne(1^1S_0) \rightarrow He(1^1S_0)+Ne(2s_2).$$

Ten rodzaj pompowania prowadzi do bardzo skutecznego wytworzenia nadwyżki obsadzenia stanu $3s_2$ nad $3p_4$ i $2p_4$ oraz stanu $2s_2$ nad $2p_4$ i do możliwości uzyskania między nimi akcji laserowej.



Rys. 8. Schemat wzbudzenia w laserze He-Ne (stosuje się skrótowe oznaczenia poziomów neonu)



Rys. 9. Schemat lasera He-Ne

Budowę lasera He-Ne przedstawia schematycznie rys. 9. Rura laserowa (szklana lub kwarcowa) zamknięta jest doskonale płasko-równoległymi okienkami nachylonymi do osi rury pod kątem Brewstera (w celu minimalizacji strat przy odbiciu); jej typowe wymiary: długość — kilkanaście cm do kilku m, średnica wewnętrzna — kilka do kilkunastu mm. Do rury wlutowane są elektrody, do których przykłada się napięcie powodujące wyładowanie. Rezonator tworzą zewnętrzne zwierciadła (płaskie lub sferyczne w ustawieniu współogniskowym), z których jedno ma pewną, niewielką przepuszczalność, co umożliwia wyprowadzenie wiązki laserowej na zewnątrz.

Zastosowanie zwierciadeł dielektrycznych o dużej zdolności odbijającej w wąskim pasmie widmowym umożliwia wybór jednej z licznych możliwych akcji laserowych. Inna metoda selekcji polega na zamknięciu rury pryzmatem i takim ustawieniu zwierciadła rezonatora, by wracało do rury tylko promieniowanie o wybranej długości fali. W tych warunkach łatwo można otrzymać jedną z trzech bardzo silnych akcji laserowych o długościach fali λ_1 = 3,39 μm, λ_2 = = 1,15 μm i λ_3 = 0,6328 μm. Liczba zidentyfikowanych przejść w laserze He-Ne przekracza już 70. Obejmują one obszar od 0,63 μm do pośredniej podczerwieni.

laser o pracy ciągłej

W czasie trwania akcji laserowej wyładowanie stale podtrzymuje różnicę obsadzeń, otrzymuje się zatem akcję laserową o działaniu ciągłym. Moc laserów tego typu wynosi, zależnie od konstrukcji i wymiarów, od kilkunastu do stu mW.

Laser He-Ne pracuje na ogół w układzie, w którym

jest wzbudzonych jednocześnie wiele modów. Często jednak w pracach badawczych, zwłaszcza wówczas, gdy istotne jest uzyskanie wysokiego stopnia spójności promieniowania, należy doprowadzić laser do pracy jednomodowej. Można to uzyskać m.in. przez odpowiednie skrócenie lasera. Rysunek 6 obrazuje rozkład modów w laserze metrowym. Odległość między modami wynosi 150 MHz. Można obliczyć, że przy skróceniu lasera do 10 cm odległość sąsiadujących modów wzrośnie do 1,5 GHz i w praktyce można oczekiwać wówczas wzbudzenia tylko jednego modu. Równocześnie jednak skrócenie ośrodka czynnego wywołuje znaczne zmniejszenie wzmocnienia i moc lasera maleje (poniżej 100 µW).

laser jonowy

Oprócz laserów gazowych, w których ciałem roboczym jest gaz szlachetny (lub mieszanina gazów szlachetnych), zbudowano wiele typów laserów jonowych. Lasery na jonach gazu szlachetnego (wśród nich jest szczególnie ważny ze względu na zastosowanie laser na jonach argonu Ar^+) pokrywają swym promieniowaniem obszar widmowy od nadfioletu do bliskiej podczerwieni. W każdym gazie szlachetnym (przeważnie przy jednokrotnej jonizacji) zidentyfikowano po kilkadziesiąt przejść laserowych. Warto wymienić również lasery na jonach metali ziem alkalicznych (Mg^+, Ca^+, Ba^+, Sr^+), na jonach grupy kadmu (Cd^+, Hg^+ i Zn^+), ważne wreszcie znaczenie mają lasery gazowe, w których ciałem roboczym są jony chlorowców (F_2^+, Cl_2^+, Br_2^+, J_2^+).

lasery molekularne

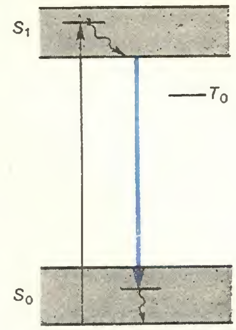
Na zakończenie przeglądu laserów gazowych wspomnimy jeszcze o dwóch, ważnych ze względu na zastosowanie, laserach molekularnych: azotowym i na CO_2. Pierwszy, wzbudzany wyładowaniem elektrycznym, pracuje w bliskim nadfiolecie i w obszarze promieniowania widzialnego (przejścia zachodzą między wzbudzonym stanem elektronowo-oscylacyjno-rotacyjnym i jednym ze stanów oscylacyjno-rotacyjnych stanu elektronowo nie wzbudzonego); drugi — pracuje zarówno w sposób ciągły, jak i impulsowy, a emituje promieniowanie w pośredniej (9,6–10,6 µm) podczerwieni (przejścia zachodzą między wzbudzonymi stanami oscylacyjno-rotacyjnymi elektronowo nie wzbudzonymi). Dzięki dużemu zagęszczeniu poziomów molekularnych sprawność pompowania jest większa w laserach molekularnych niż w atomowych i jonowych, a dodatek azotu (N_2) do dwutlenku węgla powiększa efektywność pompowania lasera na CO_2 przez przekazywanie energii wzbudzenia oscylacyjnego od cząsteczek N_2 do CO_2 w trakcie ich zderzeń.

Laser azotowy służy do pompowania laserów barwnikowych, a laser CO_2 jest źródłem promieniowania podczerwonego wielkiej mocy (w pracy impulsowej moc szczytowa wynosi do 50 kW w impulsach trwających ok. 150 ns przy częstości powtórzeń 400 Hz, a w pracy ciągłej — 500 W).

Lasery barwnikowe

Opisane dotąd lasery obejmują wprawdzie szeroki obszar widmowy, jednakże możliwości ich przestrajania w sposób ciągły są bardzo ograniczone. Tymczasem takie przestrajanie jest warunkiem licznych zastosowań laserów w spektroskopii. Właściwość tę można uzyskać w laserach barwnikowych, tj. takich, w których ośrodek czynny stanowi roztwór silnie fluoryzującego barwnika (np. fluoresceiny, rodaminy, zob. il. 12 i 13, tabl. 4).

Typowy układ poziomów barwnika w roztworze przedstawia rys. 10. Ciągłe pasma energetyczne S_0 i S_1 powstały z układów poziomów oscylacyjno-rotacyjnych barwnika w wyniku oddziaływań z cząsteczkami rozpuszczalnika. W zwykłych warunkach obsadzone są najniższe poziomy pasma S_0, a pasmo S_1 (fluorescencyjne) jest próżne. Po wzbudzeniu (zależnie od energii kwantu pochłoniętego) do jednego z wyższych poziomów pasma wzbudzonego S_1 następuje najpierw szybkie (w czasie 10^{-12} s) bezpromieniste przejście do dna pasma S_1, a następnie, po czasie rzędu 10^{-9} s, emisja kwantu fluorescencji i przejście do jednego z poziomów oscylacyjno-rotacyjnych



Rys. 10. Układ poziomów barwnika w roztworze; S_0 i S_1 pasmo podstawowe i fluorescencyjne, T_0 metatrwały poziom trypletowy

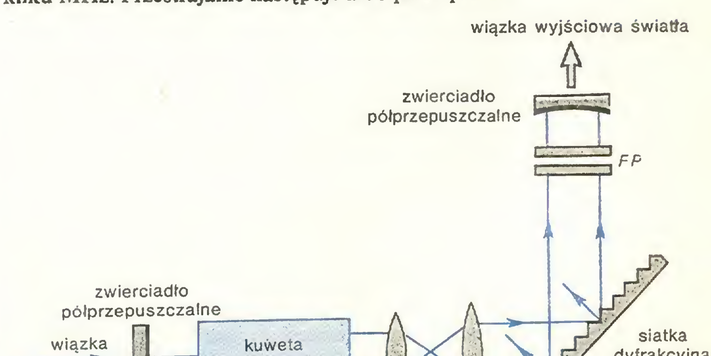


Rys. 11. Symetria zwierciadlana widm absorpcji i fluorescencji cząsteczek barwnika

S_0. Z niego następuje szybki powrót na dno pasma S_0. W ten sposób widmo absorpcji (przy wzbudzeniu źródłem o widmie ciągłym) wykazuje zwierciadlaną symetrię w stosunku do widma fluorescencji (rys. 11). Widać również, że przy silnym wzbudzeniu cząsteczek barwnika można uzyskać znaczne odwrócenie obsadzenia dna pasma S_1 względem górnego obszaru pasma S_0.

warunki akcji laserowej

Uzyskanie akcji laserowej w takim układzie wymaga nie tylko umieszczenia kuwety z roztworem barwnika w rezonatorze i intensywnego pompowania optycznego, lecz również wyposażenia rezonatora w element rozszczepiający, który umożliwia wybranie odpowiednio wąskiego pasma widma do zainicjowania i podtrzymania akcji laserowej na określonej długości fali. Stosuje się tu różne układy (rys. 12), od pryzmatu począwszy, a na kombinacji siatki dyfrakcyjnej schodkowej z interferometrem Fabry'ego–Pérota skończywszy. Ten ostatni układ pozwala na otrzymanie wiązki promieniowania o szerokości widmowej kilku MHz. Przestrajanie następuje albo przez powol-



Rys. 12. Układ optyczny lasera barwnikowego strojonego siatką dyfrakcyjną i interferometrem Fabry'ego–Pérota

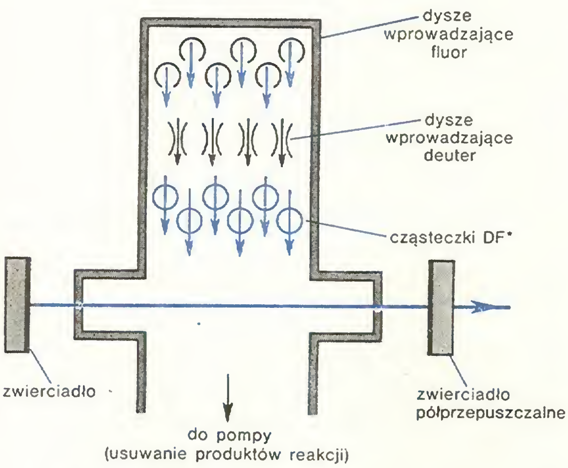
przestrajanie lasera

ną zmianę drogi optycznej w interferometrze (przestrajanie precyzyjne), albo przez niewielkie zmiany nachylenia siatki (przestrajanie zgrubne). Obszar przestrajania wynosi kilkadziesiąt nanometrów, a użycie zestawu kilku barwników pozwala na otrzymanie szerokiego obszaru widmowego. Z rys. 11 widać również, że warunki uzyskania promieniowania laserowego w obszarze nakrywania się widm absorpcji i fluorescencji są niekorzystne, absorpcja powiększa bowiem straty w układzie.

Lasery barwnikowe pompuje się bądź lampą błyskową (podobnie jak laser rubinowy), bądź też odpowiednio dostrojonym laserem argonowym lub azotowym.

Lasery chemiczne

Uzyskanie efektu laserowego jest możliwe w układach molekularnych, w których reakcje chemiczne prowadzą bezpośrednio do powstania znacznej liczby



Rys. 13. Schemat lasera chemicznego

swobodnych wzbudzonych atomów lub cząsteczek. Tego typu mechanizm wzbudzenia daje duże szanse uzyskania stanu odwrócenia obsadzeń między poziomami atomowymi lub molekularnymi.

reakcje w laserze chemicznym

Do chwili obecnej jedną z najlepiej zbadanych kombinacji składników lasera chemicznego jest fluor z wodorem lub z deuterem. Schemat tego typu lasera o działaniu ciągłym pokazany jest na rys. 13. W laserze takim zachodzą reakcje (Q — energia):

$F_2+Q = 2F$ (dysocjacja cząsteczki fluoru),
$F+D_2 = DF^*+D$ (reakcja z deuterem dająca wzbudzoną cząsteczkę DF^*),
$DF^*+nh\nu \rightarrow DF+(n+1)h\nu$ (emisja wymuszona cząsteczki DF^*).

Często do fluoru i deuteru dodaje się jeszcze CO_2 (jako składnik, który przejmuje wzbudzenie od fluoru) oraz gazy szlachetne. Wówczas (doświadczenie prowadzono przy stosunku ciśnień cząstkowych p_{D_2}: p_{F_2}:p_{CO_2}:p_{He} = 1:1:4:5 i ciśnieniu całkowitym p = = około 1000 hPa) wzbudzenie przekazane zostaje w zderzeniach cząsteczkom CO_2, które w akcji laserowej wysyłają promieniowanie o długości fali ok. 10 µm.

Lasery chemiczne nasuwają jak dotąd wiele trudnych do rozwiązania problemów technologicznych. Podczas reakcji w obszarze reaktora wydziela się bardzo wiele ciepła, które należy szybko odprowadzić, a poza tym dynamika takiego układu bardzo się komplikuje. Innymi zagadnieniami czekającymi na rozwiązanie są: proste i skuteczne przeprowadzenie dysocjacji składników reakcji na atomy; uzyskiwanie wzbudzenia molekularnego do określonych grup poziomów oscylacyjno-rotacyjnych cząsteczki (najlepsze by było wprowadzenie układu laserującego w wysokoenergetyczne stany oscylacyjne); objęcie akcją laserową obszaru widzialnego widma. Próby prowadzi się w dwóch kierunkach: jednym jest użycie jako cząsteczek laserujących wzbudzonych tlenków metali z mieszaniny Me_2+O_2, drugim — poszukiwanie przejść ze zmianą liczby kwantowej stanu oscylacyjnego większą niż o jeden; jak dotąd próby te nie dały pozytywnego wyniku.

Bibliografia → Lasery — zastosowanie.

Lasery – zastosowanie

Jacek Chrostowski

Po okresie pionierskim lasery pod koniec lat sześćdziesiątych weszły w etap dojrzałych technologicznie konstrukcji i stały się niezastąpionymi narzędziami w pracach badawczych oraz w technice. W stosunku jednak do ogromnych możliwości pod tym względem lasery są wprowadzane do techniki opornie. Poza badaniami naukowymi, w których się stały normalnym narzędziem pracy, główne zastosowanie znalazły one w obróbce termicznej materiałów, w pomiarze odległości oraz wytyczaniu kierunków. Wykorzystanie bardziej subtelnych cech promieniowania laserowego (jak w telekomunikacji laserowej) nie wyszło poza dziedzinę wojskowości lub kosmonautyki. (Tabela: Porównanie podstawowych typów laserów.)

W artykule tym omówimy tylko ważniejsze lub ciekawsze zastosowania laserów, nie omówione w innym miejscu. Zastosowanie laserów w holografii, która się rozwinęła w samodzielną dziedzinę, oraz laserów półprzewodnikowych omówione jest w artykułach „Holografia" i „Optoelektronika półprzewodnikowa". Duże znaczenie, m.in. ze względu na prace nad uzyskaniem kontrolowanej reakcji termojadrowej, ma wytwarzanie impulsów gigantycznych, omówione w artykule „Ultrakrótkie impulsy światła". Natomiast ogólna zasada działania laserów i podstawowe ich typy przedstawione są w artykule „Lasery — podstawy działania".

Telekomunikacja optyczna

Współczesne systemy łączności wykorzystują fale elektromagnetyczne z zakresu częstości radiowych i mikrofalowych. Przez modulację jednego z parametrów fali nośnej można przekazywać wiele niezależnych sygnałów; im większa jest częstość fali nośnej, tym więcej niezależnych sygnałów, zwanych kanałami, można przekazać (tabela) w jednym łączu. (Łączem nazywa się zestaw urządzeń rozmieszczonych w przestrzeni od miejsca, w którym informacja została wprowadzona i zakodowana w fali nośnej, do miejsca, gdzie została ona przekazana odbiorcy.)

kanał informacji

Idea zastosowania promieniowania optycznego do przenoszenia informacji nie jest nowa, sięga zamierzchłych czasów (najprostszym urządzeniem wykorzystującym światło do przenoszenia informacji jest lustro, odbijające promieniowanie słoneczne; taki słoneczny „zajączek" to tylko jeden kanał informacji). Można przypuszczać, że sprzężenie elektroniki z optyką spowoduje przeobrażenia telekomunikacji w niedalekiej przyszłości w większym stopniu, niż kiedyś zastosowanie techniki mikrofalowej (→ Optyka fourierowska).

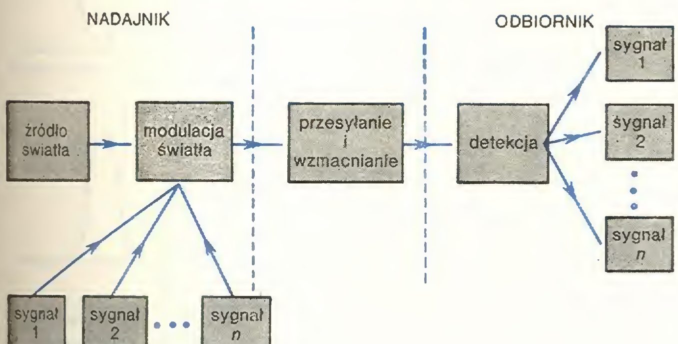
Podstawowe elementy łącza optycznego przedstawiono na rys. 1. Poniżej omówimy poszczególne jego elementy.

Porównanie podstawowych typów laserów

Typ lasera	Długość fali, nm	Rodzaj pracy, długość impulsu	Energia J	Moc W	Gęstość strumienia w ognisku W/cm²	Zastosowanie
Laser rubinowy	694,3	impulsowa ($30-3\cdot10^5$ ns)	$1-10^2$	10^3-10^9	10^8-10^{13}	technologiczne, spawanie, topienie, wiercenie, dentystyka, biologia
Laser neodymowy	1 060	ciągła lub impulsowa (15 ns)	$10^{-1}-10^2$	$10-10^3$	10^7-10^{15}	telekomunikacja, laserowe układy śledzące, kontrolowane reakcje jądrowe
Laser półprzewodnikowy GaAs	800–900	ciągła lub impulsowa (10^2 ns)	$10^{-5}-10^{-3}$	$10^{-3}-10$		telekomunikacja
Laser barwnikowy	przestrajany w zakresie 200–800	ciągła lub impulsowa ($2-2\cdot10^3$ ns), pompowany laserem argonowym lub azotowym	zależna od lasera pompującego	1 (ciągła), zależna od lasera pompującego		spektroskopia, rozdzielanie izotopów, biologia
Laser gazowy He-Ne	632,8	ciągła	—	$10^{-3}-10^{-1}$	10^2-10^4	metrologia, interferometria, holografia, geodezja
Laser argonowy jonowy	488–514,5	ciągła lub impulsowa (10^3 ns)		$1-10^3$ W	10^5-10^8	chirurgia, spektroskopia
Laser azotowy	337,1	impulsowa (10 ns)	0,01	10^6 W	10^{11}	spektroskopia, reakcje fotochemiczne
Laser CO_2	10 600	ciągła lub impulsowa ($10^2-5\cdot10^4$ ns)	$1-10^3$	$10-10^4$	10^6-10^{15}	laserowe układy śledzące, chirurgia, dentystyka, obróbka materiałów, cięcie i spawanie metali, kontrolowane reakcje jądrowe, rozdzielanie izotopów

Porównanie maksymalnej liczby kanałów telefonicznych i telewizyjnych transmitowanych na różnych falach nośnych

Rodzaj fali nośnej	Zakres częstości	Wykorzystane pasmo	Maksymalna liczba kanałów	
			telefonicznych	TV
Fale długie	30 kHz–300 kHz	10%	3	—
Fale średnie	300 kHz–3 MHz	10%	25	—
Fale krótkie	3 MHz–30 MHz	10%	200	—
UKF	30 MHz–300 MHz	10%	4000	1
UHF	300 MHz–3000 MHz	10%	10000	10
Mikrofale	3000 MHz–10^{12} Hz	10%	100000	100
Pasmo optyczne	$5\cdot10^{13}-10^{15}$ Hz	0,1%	10[illegible]	10^5



Rys. 1. Schemat blokowy łącza optycznego

laser jako źródło światła

Idealne źródło światła do celów telekomunikacyjnych powinno mieć dużą moc wyjściową, dużą sprawność przetwarzania energii, długą żywotność, wysoką stabilność częstości, dużą spójność przestrzenną i czasową, łatwość modulacji i wzbudzenia oraz małe wymiary. Różne typy laserów w różny sposób realizują te cechy idealnego źródła. Lasery półprzewodnikowe mają małe moce i niską spójność, za to są wielkości łebka od szpilki, a przy tym łatwo modulować je sygnałami o częstościach aż do 10^{10} Hz (np. przez modulację stałego prądu zasilania). Lasery gazowe mają dużą spójność emitowanego światła, wysoką stabilność częstości (względna zmiana częstości $\Delta\nu/\nu$ jest rzędu 10^{-8}) i możliwa jest modulacja światła wewnątrz lasera, ale wymagają delikatnej obsługi, są mało sprawne w przetwarzaniu energii elektrycznej na światło i stosunkowo szybko się psują (laser He-Ne ma czas życia rzędu 10 000 h, a laser argonowy — 1000 h). Moce wyjściowe laserów argonowych są rzędu dziesiątków watów, a laserów CO_2 — kilowatów (przy pracy ciągłej). Lasery stałe (m.in. neodymowe) mogą pracować w sposób ciągły ze stosunkowo dużą mocą wyjściową, lecz mają niską spójność emitowanego światła i nie mogą być modulowane wewnętrznie. Ich czas życia zależy od trwałości lampy pompującej (żywotność lamp nie przekracza obecnie 1000 godzin).

Bardzo dobre własności, zbliżone do idealnych, zwłaszcza z uwagi na zastosowanie w telekomunikacji dalekiego zasięgu i w kosmosie, ma laser neodymowy z kryształem YAG pracujący w układzie zsynchronizowanych modów, dający ciąg impulsów pikosekundowych w odstępach rzędu 1 ns. Lasery argonowe emitujące światło zielone, słabo pochłaniane przez wodę morską, mogą być zastosowane w łączności podwodnej. Do łączności naziemnej, jak też i kosmicznej, laser gazowy na CO_2 konkuruje z laserem YAG, gdyż emitowana przez niego fala długości 10,6 μm jest słabo pochłaniana w powietrzu. Natomiast wydaje się, że ze względu na małe wymiary i niskie koszty produkcji — do łączności naziemnej bliskiego zasięgu najlepsze będą lasery półprzewodnikowe w układzie ze światłowodami.

modulacja fali ciągłej

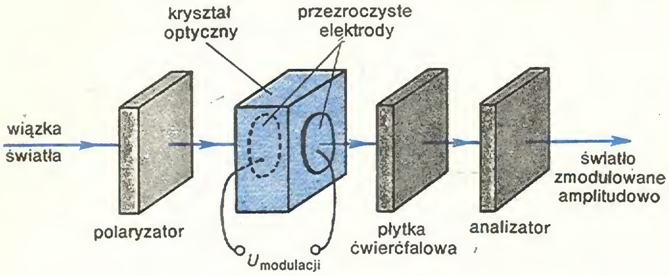
Wektor natężenia pola elektrycznego promieniowania wychodzącego z lasera można przedstawić w postaci:

$$\vec{E} = \vec{e}E_0 \sin(2\pi\nu t - \varphi),$$

gdzie E_0 jest amplitudą fali, ν — częstością, φ — fazą, $\vec{e}$ — wektorem polaryzacji, t — czasem. Zmieniając w takt przekazywanego sygnału amplitudę E_0, częstość lub fazę, otrzymuje się odpowiednio zmodulowaną falę elektromagnetyczną. Promieniowanie laserowe umożliwia jeszcze inny sposób modulacji — modulację polaryzacji. Polega ona na tym, że w takt sygnału modulującego zmienia się kierunek wektora $\vec{e}$ przy nie zmienionych innych parametrach. Do modulacji wykorzystuje się zmianę parametrów ośrodka, przez który przechodzi światło (np. zmianę współczynnika pochłaniania ośrodka, grubości środka czy

modulacja polaryzacji

współczynnika załamania), wywołaną zmianą zewnętrznego pola elektrycznego lub magnetycznego. Najczęściej wykorzystuje się zjawiska elektrooptyczne (rys. 2). W niektórych kryształach (np. w krysztale KH_2PO_4, zw. KDP) pod wpływem pola elektrycznego pojawia się wymuszona dwójłomność (zjawisko Kerra): w krysztale rozchodzą się dwie fale świetlne o różnych prędkościach i o polaryzacji liniowej wzajemnie prostopadłej. Różnica faz obu fal po opuszczeniu kryształu zależy od przyłożonego napięcia. Proporcjonalność w pewnym zakresie różnicy faz do natężenia pola elektrycznego pozwala zrealizować modulację polaryzacji lub amplitudy wiązki światła. Gdy różnica faz między obiema falami (przy ich równym natężeniu) $\Gamma = \pi/2$, fala na wyjściu modulatora jest kołowo spolaryzowana, gdy $\Gamma = \pi$, występuje polaryzacja liniowa. Ustawienie za modulatorem polaroidu (analizatora) powoduje zamianę modulacji polaryzacyjnej w amplitudową, polaroid przepuszcza bowiem światło zgodnie ze wzorem $I = I_0 \sin^2(\Gamma/2)$, jeśli płaszczyzny polaryzacji światła padającego i polaroidu są wzajemnie prostopadłe (I = natężenie fali opuszczającej modulator, I_0 — natężenie fali padającej). Właściwości kryształów elektrooptycznych pozwalają na modulację światła laserowego sygnałami, których szerokość pasma wynosi nawet kilka tysięcy MHz.

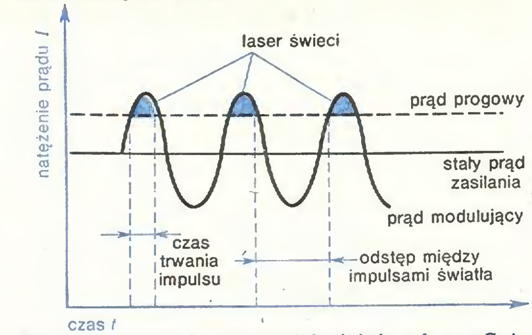


Rys. 2. Schemat elektrooptycznego modulatora światła

Omówiona wyżej metoda polegająca na modulacji ciągłej fali laserowej ma kilka wad. Mianowicie, do przekazania informacji z reguły nie jest potrzebna praca ciągła lasera, a urządzenie impulsowe jest bardziej sprawne energetycznie. Sygnały ciągłe przekazywane na duże odległości ulegają znacznym zniekształceniom, zmieniającym się w czasie i powodującym przekłamania w przekazywaniu informacji. Dochodzi do tego jeszcze fakt, że coraz powszechniej informacja jest transmitowana w postaci binarnej (dane komputerowe, a nawet rozmowy telefoniczne). Dlatego w systemach telekomunikacji optycznej stosuje się modulacje impulsowo-kodowe. W systemach tych można rozróżnić dwa etapy: wytwarzanie ciągu impulsów (laser pracujący w specjalnych warunkach) i kodowanie (modulację). Wytwarzanie impulsów w stałych odstępach czasu najczęściej realizuje się przez synchronizację modów w laserach YAG czy helowo-neonowych. W laserach półprzewodnikowych ciąg impulsów można uzyskać przez periodyczną zmianę dobroci rezonatora, również przez synchronizację modów lub zmiany natężenia prądu zasilania (rys. 3). Gdy przez laser przepływa prąd o natężeniu trochę mniejszym od wartości progowej, laser nie promieniuje spójnie. Jeśli w pewnych chwilach spowodujemy wzrost tego prądu do wartości większej niż progowa, laser zacznie świecić. Periodyczna zmiana prądu powoduje więc powstanie impulsów światła spójnego. Częstość powtarzania impulsów zależy od częstości prądu modulującego. Zakreskowane pola na rysunku odpowiadają impulsom światła wychodzącego z lasera.

modulacja impulsowo-kodowa

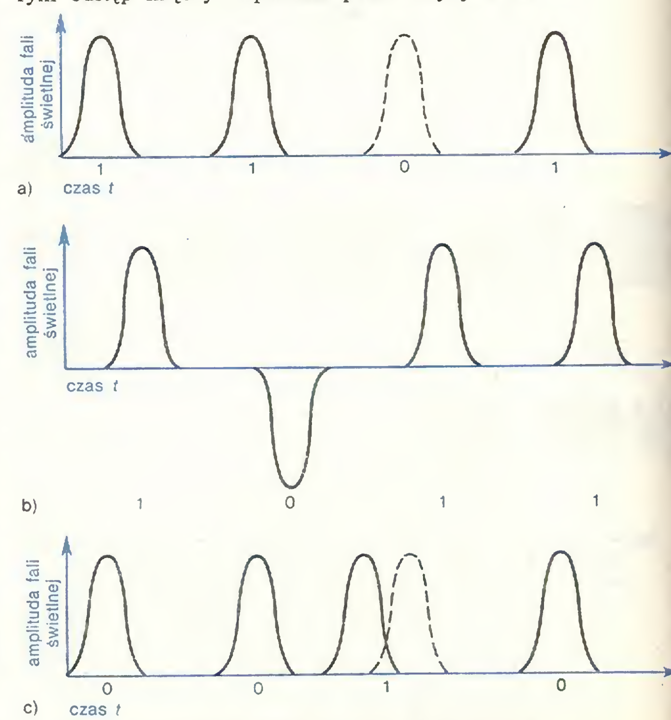
Drugi etap — kodowanie informacji — polega na modyfikowaniu ciągu impulsów z lasera za pomocą zwykłych modulatorów elektrooptycznych. Modulacja impulsowo-kodowa jest z natury binarna, gdyż rola detektora polega wyłącznie na określeniu, czy dany impuls dotarł, czy nie dotarł do odbiornika. Informacja o amplitudzie nie jest istotna, co ma ogromne znaczenie przy nie najlepszych warunkach propagacji fali świetlnej na Ziemi.



Rys. 3. Zasada wytwarzania impulsów światła w laserze GaAs przez modulację prądu zasilania

modulacje PCM-IM, PCM-PM i PPM

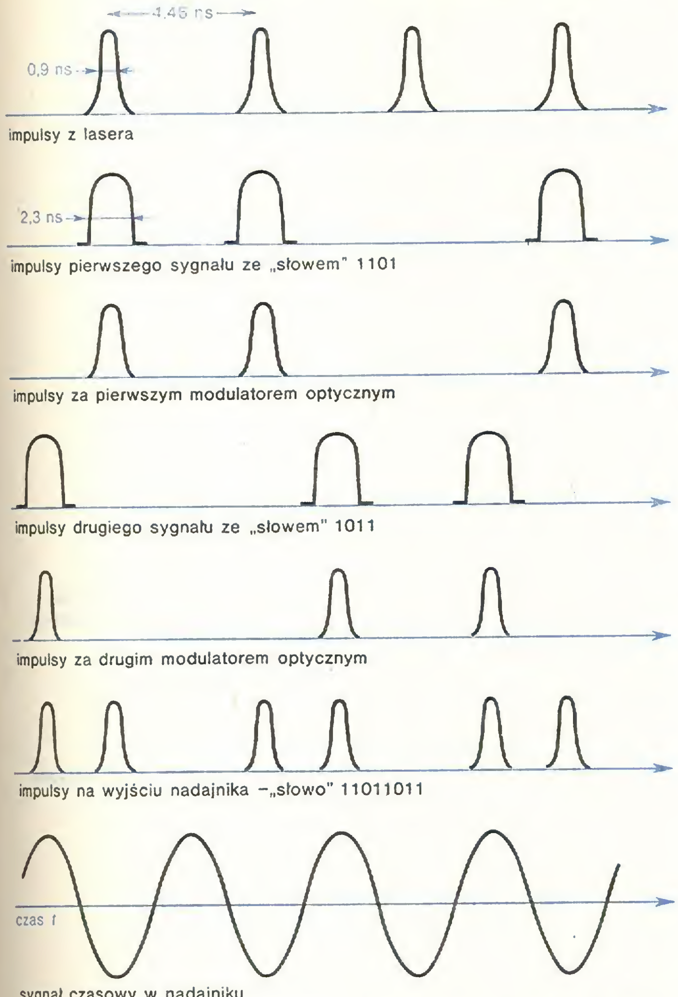
Podobnie jak w modulacji fali ciągłej, w tej metodzie modulacji również istnieje kilka sposobów zmiany parametrów impulsów świetlnych. Najczęściej stosuje się trzy z nich: modulację natężenia, modulację polaryzacji i modulację położenia impulsu (rys. 4a, b, c). Zasadę modulacji impulsowo-kodowej z modulacją natężenia (tzw. PCM-IM) wyjaśnia rys. 4a. Modulator zatrzymuje dany impuls lub go przepuszcza i daje w odbiorniku informację 1, co oznacza, że jest impuls, lub 0, świadcząc, że go nie ma. Przy modulacji impulsowo-kodowej z modulacją polaryzacji (tzw. PCM-PM) modulator przepuszcza impuls bez zmiany polaryzacji, albo też zmienia jego polaryzację o 180°. Może to być również zmiana na polaryzację lewoskrętną (1) i prawoskrętną (0). Modulacja położenia impulsu (PPM) przesuwa impuls względem położenia początkowego. Możliwe jest rozwiązanie w postaci analogowej (położenie impulsu zmienia się w sposób ciągły) lub rozwiązanie, w którym odstęp między impulsami podzielony jest na



Rys. 4. Zasada modulacji: a) impulsowo-kodowej z modulacją natężenia (PCM-IM); b) impulsowo-kodowej z modulacją polaryzacji (PCM-PM); c) położenia impulsu (PPM)

szereg dyskretnych przedziałów czasu. Jeśli rozwiązanie jest takie, jak przedstawiono na rys. 3, możliwa jest prosta modulacja PPM przez modulację częstości sinusoidalnego prądu (co odpowiada przesunięciu impulsów światła na osi czasu względem siebie).

Zasadę przesyłania informacji w ciągu impulsów przedstawia rys. 5. Dwie niezależne informacje —



Rys. 5. Zasada zwielokrotniania informacji przesyłanej w jednym łączu optycznym z wykorzystaniem modulacji impulsowo-kodowej

słowa 1101 oraz 1011 — przekazywane są razem jako jedno słowo 11011011. Sygnał czasowy w odbiorniku pozwala określić, czy dany impuls pochodzi z pierwszego słowa (rejestracja danego impulsu odpowiada ujemnej części sinusoidy na rysunku) czy z drugiego (dodatniej części sinusoidy).

przesyłanie sygnałów

Wydaje się, że największym problemem w obecnej chwili jest przenoszenie informacji z nadajnika do odbiornika. Transmisja w wolnej przestrzeni na dalsze odległości jest praktycznie możliwa tylko w przestrzeni kosmicznej (systemy łączności satelitarnej lub międzyplanetarnej). Atmosfera ziemska powoduje zniekształcenia czoła fali i duże straty mocy; wynoszą one — w zależności od warunków atmosferycznych — ok. 3–8 dB/km (2–6,31 razy) przy deszczu, 3–10 dB/km (2–10 razy) przy mgle i 3–20 dB/km (2–100 razy) przy padającym śniegu. Z tego względu nie stosuje się swobodnej propagacji fali na odległość ponad 20 km; moce laserów konieczne do zapewnienia trwałej łączności musiałyby być olbrzymie. Przy transmisji na większe odległości stosuje się światłowody, które zachowują stałość parametrów przez długi czas.

światłowody

Światłowody mogą być dwóch rodzajów. Jeden z nich zawiera prowadnicę oraz ustawione w niej soczewki. Soczewki szklane lub gazowe ustawia się tak, że odległości między nimi są równe podwojonej ogniskowej, dzięki czemu wiązka światła jest prowadzona wzdłuż osi światłowodu. Gdy soczewki są oddalone od siebie o kilkadziesiąt metrów, straty wynoszą ok. 1 dB/km (1,26 razy). Światłowody tego typu są jednak drogie (ze względu na precyzję ustawienia soczewek) i mimo bardzo dobrych własności ich zastosowanie ograniczone jest do małych odległości i do systemów teletransmisji o dużych pojemnościach informacyjnych. Znacznie prostsze i tańsze w produkcji są światłowody zbudowane z włókien szklanych grubości kilku mikrometrów i o małym współczynniku załamania. Światło się odbija wielokrotnie od ścianek włókien (całkowite wewnętrzne odbicie) i rozchodzi ze stosunkowo małymi stratami. Splecione z wielu takich cienkich włókien giętkie przewody grubości cienkich kabli elektrycznych mają straty rzędu kilkunastu dB/km (kilkadziesiąt razy). Dla porównania podamy, że w litym szkle straty mocy wynoszą 100 dB/km (10^{10} razy).

wzmacnianie sygnału

Jak widać, straty w światłowodach są znaczne i porównywalne ze stratami w wolnej przestrzeni. Konieczne jest więc wzmacnianie sygnału na drodze do odbiornika. Najczęstszą metodą jest wstawienie na drodze zmodulowanej wiązki światła materiału laserującego o odwróconych obsadzeniach poziomów energetycznych. Foton, który się dostanie w obszar wypełniony atomami wzbudzonymi, spowoduje emisję wymuszoną określonej liczby fotonów dokładnie takich samych jak padający. W rezultacie sygnał opuszczający materiał laserujący zostanie wzmocniony.

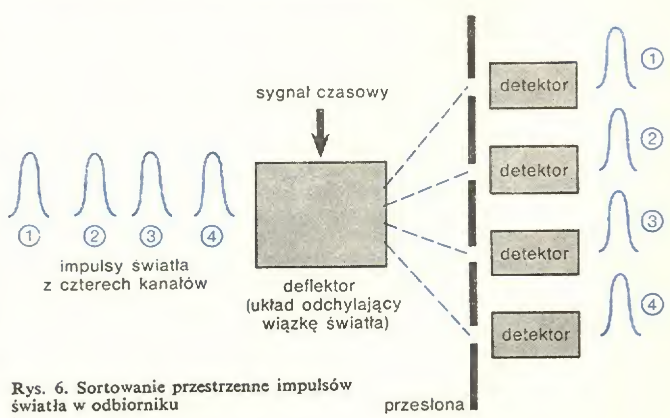
Po dotarciu do odbiornika zmodulowany sygnał świetlny przechodzi przez układ soczewek zbierających oraz przez filtr pasmowy (dla zmniejszenia szumów), dostrojony do częstości fali nośnej, i skupia się na detektorze. Rolą detektora jest oddzielenie sygnału informacyjnego od fali nośnej. W pasmie optycznym stosowane są dwa rodzaje detekcji.

detekcja niekoherentna

Detekcję niekoherentną (bezpośrednią) stosuje się wówczas, gdy stosunek mocy sygnału do mocy szumów jest większy od jedności. Polega ona na przetworzeniu obwiedni natężenia wiązki świetlnej na prąd elektryczny w fotopowielaczu (prąd fotoelektronów jest proporcjonalny do natężenia padającego światła) lub w fotodiodzie półprzewodnikowej (prąd w kierunku zaporowym jest modulowany padającym światłem). Skończony czas przejścia elektronów między elektrodami umożliwia detekcję sygnałów zmiennych tylko do częstości 10^8 Hz.

detekcja koherentna

Detekcja koherentna (heterodynowa lub homodynowa) polega na mieszaniu wiązki zmodulowanej z inną, niezmodulowaną wiązką spójną; jest stosowana przy sygnałach bardzo słabych, na poziomie



Rys. 6. Sortowanie przestrzenne impulsów światła w odbiorniku

niższym od poziomów szumów. W wyniku złożenia obu wiązek — zmodulowanej, niosącej sygnał, i nie zmodulowanej — odtwarzamy samą obwiednię wiązki sygnałowej. Dalszy proces rozdzielenia wielu sygnałów niesionych razem odbywa się metodami elektronicznymi.

Przy modulacji impulsowo-kodowej detektor daje tylko informację o dotarciu danego impulsu do odbiornika, a więc mniejsza jest tu możliwość przekłamania informacji w wyniku zakłóceń. Inna jest w takim wypadku technika rozdzielania informacji z wielu kanałów. Rysunek 6 przedstawia schemat układu rozdzielającego poszczególne kanały przez sortowanie impulsów przychodzących w danych odstępach czasu od odbiornika. Deflektorem (układem odchylającym wiązkę światła) jest kryształ, w którym pod wpływem fali akustycznej lub napięcia elektrycznego wiązka światła zmienia kierunek rozchodzenia się.

stosowanie impulsów

Zastosowania technologiczne

Z punktu widzenia zastosowań technologicznych bardzo ważną własnością wiązki laserowej jest możliwość uzyskiwania olbrzymiej ilości energii w sposób kontrolowany; ponadto wiązka laserowa może być skupiona do rozmiarów poprzecznych porównywalnych z długością fali świetlnej, tj. rzędu 1 μm. Obie te własności sprawiają, że działanie wiązki laserowej ogranicza się do małego obszaru, miejsca dalej położone nie doznają wpływu promieniowania.

spawanie i zgrzewanie

Spawanie polega na stopieniu brzegów oddzielnych części materiału (lub różnych materiałów) i połączeniu ich w jedną całość, a następnie — ostudzeniu. Palnik acetylenowy dostarcza do spawanego miejsca 1000 W/cm^2, laser — co najmniej kilkaset razy więcej. Pewną trudność przy spawaniu laserem sprawia fakt, że stopiony metal ma mniejszy współczynnik odbicia światła, tak że w chwili, gdy się materiał topi, następuje wzrost energii kumulowanej w spoinie; aby przy topieniu uniknąć odparowania materiału, trzeba w tym momencie zmniejszyć moc lasera, co jest zabiegiem dość trudnym.

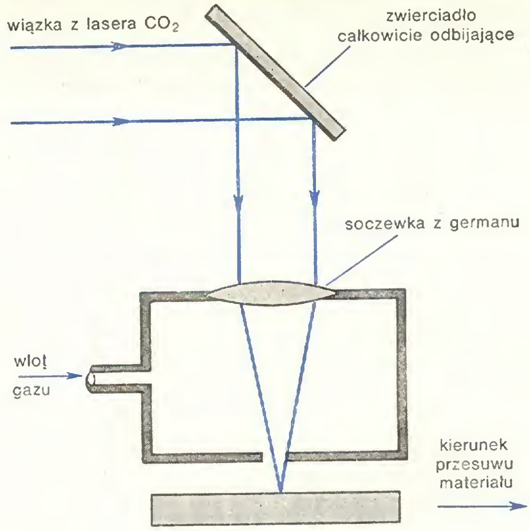
Do spawania najlepiej się nadają impulsowe lasery neodymowe i rubinowe. Szczególnie wygodne i celowe jest użycie lasera do zgrzewania elementów przy produkcji układów scalonych i innych urządzeń o małych wymiarach (il. 136, tabl. 34). Odpowiedni dobór czasu trwania impulsów i skupienie wiązki na reperowanym fragmencie umożliwiło dostarczenie dużej energii na małym obszarze bez zbytniego nagrzania szkła. Jest to praktycznie jedyny sposób usuwania uszkodzeń w drogich urządzeniach.

wiercenie otworów

Lasery neodymowe i rubinowe dostarczają energii tak dużej i dobrze skupionej, że możliwe jest wykonywanie otworów w materiałach bardzo twardych (jak diament, różne ceramiki, tantal, wolfram, molibden), których nie można obrabiać metodami tradycyjnymi. Do wiercenia używa się takich samych układów laserowych jak do spawania, jedyna różnica polega na zwiększeniu mocy lasera. Wiercenie za pomocą lasera jest szczególnie efektywne przy wykonywaniu otworów bardzo małych, trudnych do wykonania inną techniką. Można w ten sposób uzyskiwać otwory o średnicy rzędu 1 μm, przy czym stosunek głębokości otworu do jego średnicy w pewnych materiałach wynosi 20:1.

cięcie laserem

Duże szybkości cięcia, niewytwarzanie kurzu czy opiłków metalu, a jednocześnie względna cisza przy cięciu — to podstawowe zalety cięcia laserem. Najlepiej do tego celu nadaje się laser CO_2 (il. 15, 17, tabl. 4). Schemat noża laserowego przedstawia rys. 7. Przez odpowiedni dobór mocy lasera (kilkaset watów) i prędkości przesuwu wiązki względem materiału uzyskuje się tylko lokalne odparowanie materiału. Cięcie może się odbywać w atmosferze obojętnego azotu lub tlenu. Tlen w wyniku egzotermicznej reakcji z materiałem przyspiesza obróbkę. W trakcie spawania gaz wydmuchuje stopiony materiał i zostawia ostrą krawędź o tak doskonałych właściwościach, że nie ma potrzeby wygładzania jej.



Rys. 7. Schemat noża laserowego

Noże laserowe stosowane są głównie w przemyśle lotniczym i kosmonautycznym, ale używa się ich również w innych przemysłach (cięcie materiałów na ubrania lub tektury na pudełka).

korekcja oporników

Lasery impulsowe sprzężone z układem sterującym i pomiarowym stosuje się do kontroli i wzorcowania oporników; skupiona wiązka laserów odparowuje lokalne obszary materiału oporowego, zmieniając jego opór. Urządzenie takie może wykonać w zautomatyzowany sposób nawet 10 000 kalibracji na godzinę przy dokładności dopasowania oporu do nominalnej wartości nawet do 0,01%.

trasowanie

W przemyśle półprzewodnikowym lasery stosuje się do trasowania. W jednej płytce materiału półprzewodnikowego wytwarza się naraz bardzo wiele układów scalonych, które po wykonaniu należy rozdzielić i zamknąć w obudowach. Rozdzielanie metodą tradycyjną przebiega w ten sposób, że się nacina linie diamentem i łamie płytkę na małe fragmenty. Laser od razu tnie płytkę na części. Trasowanie linii laserem jest bardzo szybkie, typowa prędkość przesuwu wynosi 15 cm/s, a przy tym krawędzie poszczególnych układów scalonych są ostrzejsze.

Żyroskop laserowy

Żyroskop jest to szybko obracające się ciało sztywne, którego oś obrotu może zmieniać swoją orientację względem obudowy dzięki odpowiedniemu zawieszeniu (np. Cardana). Przyłożenie do osi żyroskopu momentu sił powoduje jego precesję, której prędkość kątowa jest tym mniejsza, im większy jest moment pędu żyroskopu $J = I\omega$, gdzie I — moment bezwładności wirującej z prędkością kątową ω masy. Oś szybko obracającego się żyroskopu niemal nie zmienia swej orientacji pod wpływem krótkotrwałych sił zewnętrznych i dzięki temu może służyć za czujnik zmiany kierunku w przestrzeni poruszającego się urządzenia, zmiany prędkości kątowych lub postępowych. Żyroskopy używane są do automatycznego sterowania ruchem rakiet i samolotów, do stabilizacji statków na morzu oraz w nawigacji samolotów, statków i pojazdów kosmicznych (wskaźniki kursu, sztuczny horyzont, określanie stron świata).

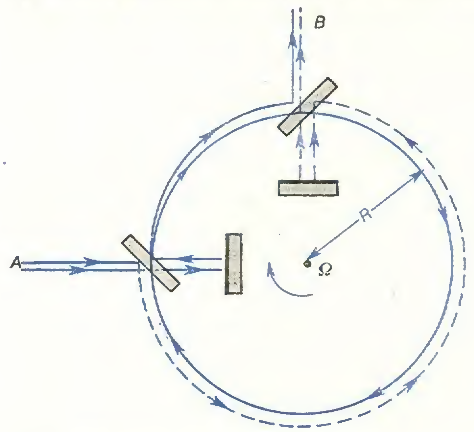
żyroskop konwencjonalny

Żyroskop laserowy różni się od tradycyjnych tym, że nie ma wirującej masy ani żadnej części ruchomej. Zasadę działania żyroskopu laserowego najlepiej się

zrozumie po zapoznaniu się z działaniem interferometru pierścieniowego Sagnaca (1913 r.).

interferometr pierścieniowy bierny

W interferometrze pierścieniowym biernym światło wchodzi w punkcie A i przechodzi przez płytkę dzielącą, by następnie w postaci dwóch wiązek obiec obwód zamknięty w przeciwnych kierunkach (dla uproszczenia obwód jest przedstawiony w postaci koła, rys. 8). Wiązki łączą się ze sobą na płytce dzielącej. Kiedy interferometr jest nieruchomy, czas obiegu obwodu przez oba promienie jest taki sam i wynosi $t = 2\pi R/c$, gdzie c jest prędkością światła. Jeżeli interferometr obraca się wokół osi prostopadłej do jego powierzchni ze stałą prędkością kątową Ω, czas obiegu obwodu koła przez oba promienie ulegnie zmianie,



Rys. 8. Interferometr pierścieniowy Sagnaca

gdyż po pewnym czasie płytka dzieląca znajdzie się w punkcie B i wobec tego promień biegnący zgodnie z kierunkiem obrotu przemierzy dłuższą drogę niż ten, który biegnie w przeciwną stronę. Prędkość światła jest oczywiście stała. Wobec tego

$$2\pi R \pm X_{\pm} = ct_{\pm},$$

gdzie droga dodatkowa przebyta przez każdy z promieni wynosi $X_{\pm} = R\Omega t_{\pm}$. Czas przejścia obu fal przez interferometr wynosi więc

$$t_{\pm} = \frac{2\pi R}{c \mp R\Omega}.$$

Znak „+" dotyczy fali biegnącej w kierunku obrotu, a znak „—" — fali biegnącej w kierunku przeciwnym. Różnica czasów w pierwszym przybliżeniu wynosi:

$$\Delta t = t_{+} - t_{-} = \frac{4\pi\Omega R^2}{c^2},$$

a różnica dróg optycznych:

$$\Delta L = c \cdot \Delta t = \frac{4\pi R^2 \Omega}{c} = \frac{4A\Omega}{c},$$

gdzie $A = \pi R^2$ jest powierzchnią zakreśloną przez promienie. Wzór na ΔL można stosować w wypadku dowolnej powierzchni oraz ośrodka o współczynniku załamania różnym od jedności. Za pomocą interferometru pierścieniowego A. H. Michelson i H. G. Gale zmierzyli (1925) szybkość obrotu Ziemi. Interferometr miał kształt prostokąta o wymiarach 610×335 m, a mierzona różnica dróg optycznych wynosiła 130 nm, czyli bardzo mało w porównaniu z wymiarami interferometru i długością fali.

interferometr pierścieniowy aktywny

Trudność użycia interferometru pierścieniowego jako praktycznego urządzenia wynika z jego bardzo małej czułości. Użycie lasera jako źródła zewnętrznego nic nie pomaga. Sytuacja ulega natomiast radykalnej zmianie, gdy się użyje aktywnego interferometru pierścieniowego. Warunek wzbudzenia drgań odpowiadających najniższemu modowi poprzecznemu zarówno w laserze liniowym, jak i pierścieniowym (rys. 9) jest taki sam: „na długości rezonatora L powinna się zmieścić całkowita liczba połówek fali". W laserze liniowym poszczególne rodzaje drgań (mody) odpowiadają dwóm przeciwnie biegnącym falom, które w sumie dają falę stojącą. Amplitudy i częstości fal biegnących muszą być jednakowe. Natomiast w laserze pierścieniowym każdy mod poprzeczny odpowiada falom biegnącym w przeciwnych kierunkach, są one niezależne w tym sensie, iż mogą mieć różne częstości i amplitudy.

Warunek wzbudzenia drgań można zapisać następująco:

$$\frac{m}{2}\lambda_{\pm} = L_{\pm} \quad \text{lub} \quad \nu_{\pm} = \frac{mc}{2L_{\pm}},$$

gdzie m oznacza liczbę modów (typowe wartości m wynoszą 10^5–10^6). Małe zmiany drogi optycznej L powodują różnicę częstości między falami $\Delta\nu$, przy czym

$$\frac{\nu_{+} - \nu_{-}}{\nu_{\pm}} = \frac{\Delta\nu}{\nu} = \frac{\Delta L}{L}.$$

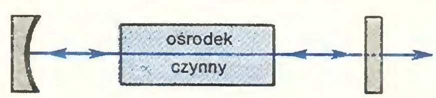
Gdy różnica długości ΔL wywołana jest obrotem, wówczas otrzymujemy:

$$\Delta\nu = \frac{4A\Omega}{L\lambda}.$$

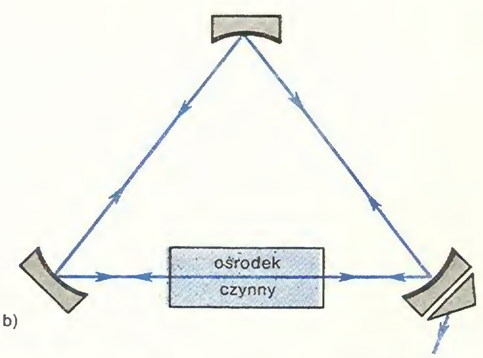
W wypadku lasera w kształcie trójkąta o boku 13,2 cm, wysyłającego falę długości $\lambda = 0{,}633$ μm i przy prędkości obrotu 10° na godzinę różnica częstości wynosi 5,9 Hz. Częstość taką przy zastosowaniu techniki heterodynowej można dość łatwo zmierzyć.

wyprowadzanie informacji

Rysunek 10 przedstawia zasadę wyprowadzania informacji z lasera pierścieniowego. Niewielki procent (przeważnie mniej niż 0,1%) energii obu fal jest



a)



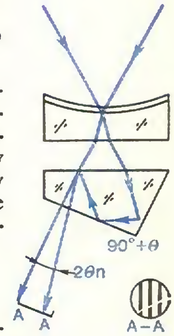
b)

Rys. 9. Konfiguracja laserów: a) liniowego, b) pierścieniowego

transmitowany przez jedno ze zwierciadeł dielektrycznych na zewnątrz. Dzięki pryzmatowi doprowadza się obie wiązki do interferencji. Prążki interferencyjne są miarą chwilowej różnicy faz między biegnącymi w przeciwnych kierunkach falami. Gdy natężenia obu fal są jednakowe, a wiązki wychodzące z pryzmatu prawie równoległe (o rozbieżności kątowej ε), rozkład prążków dany jest wzorem:

$$I = I_0[1 + \cos(2\pi\varepsilon \cdot X/\lambda + \Delta\omega \cdot t + \varphi),$$

gdzie $\Delta\omega = 2\pi\Delta\nu$ jest częstością różnicową, λ — długością fali, φ — stałą fazą. Gdy laser się nie obra-



Rys. 10.

ca, wtedy $\Delta\omega = 0$ i rozkład prążków jest nieruchomy. Podczas obrotowego ruchu lasera prążki przesuwają się z prędkością zależną od $\Delta\omega$. Odstęp między prążkami wynosi λ/ε, gdzie $\varepsilon = 2n\theta$, n — współczynnik załamania światła w pryzmacie, a kąt θ — to odchylenie kąta łamiącego pryzmatu od 90°. Kiedy np. $\theta = 15''$ i $\lambda = 0{,}633$ µm, odstęp między prążkami wynosi ok. 3 mm. Jeśli więc w miejscu pojawiania się prążków ustawi się detektor o wymiarach znacznie mniejszych od szerokości prążka (np. fotodiodę półprzewodnikową), otrzyma się bezpośrednio informację o częstości różnicowej.

Kierunek ruchu prążków określa jednocześnie kierunek obrotu lasera, tak więc ustawiając drugi detektor w odległości 1/4 prążka od pierwszego i łącząc oba detektory z układem logicznym, otrzymamy tak dodatnie, jak i ujemne zliczenia impulsów. W ten sposób żyroskop laserowy daje scałkowaną informację cyfrową w postaci:

$$N = \frac{4A}{\lambda L}\,\psi,$$

gdzie liczba zliczeń $N = \int_0^t \Delta\nu\,dt$, a kąt odchylenia od położenia początkowego $\psi = \int_0^t \Omega\,dt$. W podanym przykładzie liczbowym odnoszącym się do lasera He-Ne jedno zliczenie odpowiada obrotowi o 1,7 sekundy kątowej, tak więc pełny obrót o 360° daje $0{,}76\cdot10^6$ zliczeń.

żyroskop laserowy

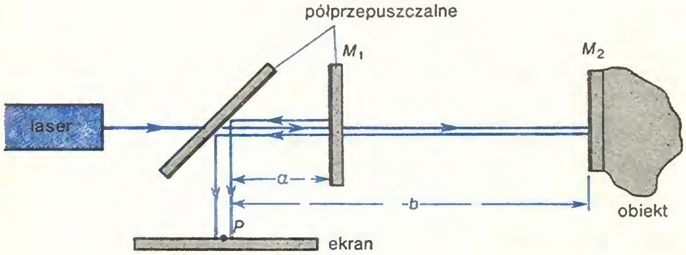
Żyroskop laserowy obejmuje trzy inteferometry pierścieniowe, ustawione w 3 płaszczyznach prostopadłych do siebie (il. 137, tabl. 35). Urządzenia takie w chwili obecnej rywalizują jeszcze z rozwiązaniami tradycyjnymi. Podstawowe niedogodności żyroskopu laserowego to: dość duża wielkość (10×5 cm) w porównaniu z wymiarami żyroskopu zwykłego (8××4 cm), zmiany parametrów w czasie (dryfy — nie pozwalające na razie na zastosowanie żyroskopów laserowych przy precyzyjnych pomiarach kosmicznych), stosunkowo krótki czas życia rzędu 5000 h pracy.

Główne zalety to: niski koszt produkcji, duża szybkość reakcji, stabilność i wytrzymałość na duże przyspieszenia i bardzo szybkie obroty, niezawodność (brak części ruchomych) Ze względu na te zalety omówiona technika pomiaru obrotów będzie najprawdopodobniej stosowana w nawigacji i w urządzeniach kontrolnych.

Zastosowanie w metrologii i geodezji

wytyczanie kierunków

Najprostszym zastosowaniem lasera w geodezji jest użycie go jako doskonałej linijki. Mała rozbieżność wiązki umożliwia wytyczanie kierunku w przestrzeni na bardzo dużej odległości. Wiązka lasera może np. służyć za oś toru wodnego w porcie, ułatwiającą prowadzenie statków o dużej wyporności w wąskich kanałach; wiązki używanych do tego celu laserów He-Ne mają średnicę kilku centymetrów po przebyciu jednego kilometra. Na lądzie promień lasera prowadzi koparkę lub maszynę górniczą w kopalni.



Rys. 11. Pomiar odległości w interferometrze Michelsona

Zainstalowany w postaci układu fotodiod czujnik, na który pada promieniowanie laserowe pozwala zdalnie sterować maszyną lub rejestrować odchylenia od linii prostej szybko i z dużą dokładnością.

pomiar małych odległości

Przy pomiarze odległości wykorzystuje się tradycyjne interferometry wielowiązkowe z laserem jako źródłem światła. Najprostsze są interferometry Michelsona (rys. 11). Po opuszczeniu lasera wiązka światła pada na zwierciadło półprzepuszczalne M_1, gdzie ulega podziałowi. Jedna z wiązek, po odbiciu od M_1, dochodzi bezpośrednio do punktu P na ekranie, druga odbywa dłuższą drogę do zwierciadła M_2 i po odbiciu dociera też do punktu P. Zwierciadło M_2 jest przymocowane do ruchomego przedmiotu albo jest po prostu jego powierzchnią. W płaszczyźnie ekranu obserwuje się prążki interferencyjne światła o natężeniu

$$I = I_0 \cos^2\left[\frac{2\pi}{\lambda}|b-a|\right].$$

Jeśli w płaszczyźnie ekranu ustawimy mały detektor — fotodiodę, to prąd płynący przez fotodiodę będzie się zmieniał wraz ze zmianą odległości b zgodnie z powyższym wzorem. Maksimum natężenia przypadnie na punkty określone zależnością

$$2\,|b-a| = N\lambda,$$

gdzie N jest liczbą naturalną. Ze wzoru tego można wyznaczyć wartość b, znając rząd interferencji N. Metoda powyższa jest szczególnie wygodna do pomiaru różnicy odległości, np. jest wykorzystywana w technologii elektronicznej do pomiaru grubości cienkich warstw, równej różnicy dróg optycznych światła odbitego od podłoża i od powierzchni warstwy. Podstawowa dokładność pomiaru odległości w interferometrze dwuwiązkowym wynosi $\lambda/2$, a w rozbudowanych interferometrach, tzw. kalibratorach laserowych, można — przy użyciu lasera helowo-neonowego — uzyskać pomiar z dokładnością $\lambda/300$, czyli ok. 2 nm. Stosowanie tej metody ogranicza trudność związana ze zliczaniem dużej liczby N prążków (zależnej od $b-a$), które się przesuwają w trakcie pomiaru na ekranie, np. pomiarowi grubości 1 cm odpowiada zliczenie 30 000 prążków. Stąd też kalibratory laserowe wyposażone są w małe komputery liczące i analizujące ruch prążków.

kalibratory laserowe

Interferometryczne kalibratory laserowe znajdują zastosowanie np. w kontroli jakości dużych elementów optycznych, w pomiarach weryfikacyjnych wzorców długości i grubości, pomiarach w trakcie produkcji elementów mikroelektronicznych, w bardzo dokładnym ustawieniu obrabiarek (np. stół obrabiarki można ustawić lub przesunąć o kilka metrów z dokładnością do 1 µm).

pomiar dużych odległości

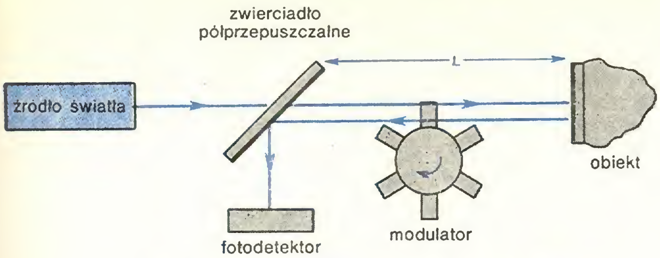
Precyzyjne pomiary dużych odległości (rzędu 1 km i większych) polegają na pomiarze czasu przejścia impulsu laserowego do elementu odbijającego i z powrotem do detektora (il. 138, tabl. 35). Czas ten pomnożony przez prędkość światła w powietrzu daje podwojoną odległość od obiektu. Można także mierzyć różnicę faz światła wychodzącego i powracającego do detektora.

Omówimy tylko jedną z metod, która stanowi rozwinięcie metody Fizeau, użytej niegdyś do pomiaru prędkości światła. Światło, które przeszło przez modulator (rys. 12) i zostało odbite od badanego obiektu, dotrze do detektora, jeżeli akurat modulator elektrooptyczny będzie otwarty (w doświadczeniu Fizeau była to przesłona mechaniczna). Przez zmianę częstości napięcia modulującego w modulatorze można tak dobrać czas τ między kolejnymi zasłonięciami modulatora, by światło wróciło do odbiornika, zatem

$$2L/v = N\tau,$$

gdzie v — prędkość światła w powietrzu, liczba N mówi, ile razy modulator otworzył się do momentu powrotu impulsu świetlnego do przyrządu. Mierząc

$N\tau$ — całkowity czas między pierwszym a drugim przejściem impulsu świetlnego przez modulator (pomiar elektroniczny) — łatwo jest obliczyć odległość



Rys. 12. Metoda Fizeau pomiaru czasu przejścia sygnału

L dalmierza od przedmiotu. Dokładność pomiaru odległości tą metodą na odcinkach rzędu 1 km wynosi $\Delta L/L = 10^{-7}-10^{-8}$, czyli bezwzględny błąd pomiaru jest rzędu 1 µm.

pomiar prędkości

Jeśli jedno ze zwierciadeł interferometru Michelsona (rys. 11) porusza się wzdłuż osi optycznej z prędkością v, to natężenie światła w płaszczyźnie ekranu zmienia się z częstością $\nu = 2v/\lambda$. Gdy długość fali $\lambda = 630$ nm, częstość 1000 prążków/s odpowiada prędkości zwierciadła 0,3 mm/s. Ruch prążków na ekranie oczywiście oznacza, że światło odbite od zwierciadła zmienia wskutek efektu Dopplera częstość o wartość $2v/\lambda$. Poruszające się zwierciadło można zastąpić przy odpowiedniej mocy lasera chropowatą powierzchnią, a nawet poruszającym się płynem. W taki to sposób mierzy się od niedawna prędkość wiatru w czasie huraganów.

Laserowe układy śledzące

Celem działania układów śledzących jest określenie odległości, prędkości i kątów, tj. podanie zupełnej informacji o obiektach ruchomych. Pojęcie „układ śledzący" jest jednak wieloznaczne, gdyż wymagania dotyczące skali odległości i czasu mogą się różnić w zasadniczy sposób. Może być np. potrzebna bieżąca informacja o locie rakiety czy samolotu, a także wieloletnie pomiary odchyleń położenia Księżyca od orbity wokół Ziemi. Przy śledzeniu startującej rakiety mierzona odległość jest rzędu 10^4 m, a prędkość kątowa — 10^5 sekund kątowych na sekundę. Natomiast pomiary orbity Księżyca z Ziemi wymagają określenia odległości $3{,}8\cdot10^8$ m i prędkości kątowej 14,5 sekundy kątowej na sekundę. Również żądana dokładność jest różna: w systemach obrony powietrznej wystarczy informacja, czy dany obiekt jest w pewnej objętości przestrzeni powietrznej, przy kontroli zaś lotu niektórych satelitów czy statków kosmicznych odległość 10^6 m musi być określona z dokładnością do 1 cm. W wielu sytuacjach wystarczy dany obiekt uznać za punkt w przestrzeni, gdy w innych wymagana jest znajomość kształtu lecącego ciała. Z racji tak różnych wymagań istnieje wiele odmian układów śledzących, a wszystkie stanowią bardzo rozbudowane urządzenia, wyposażone w wyspecjalizowane komputery.

układy laserowe a radarowe

Ogólna zasada pracy układów laserowych jest taka sama jak układów radarowych mikrofalowych: pomiaru odległości dokonuje się przeważnie przez pomiar czasu powrotu odbitego od obiektu impulsu światła; informacje o prędkości uzyskuje się przez pomiar zmian częstości fali świetlnej (efekt Dopplera). Z reguły układy te pracują w podczerwieni ze względu na mniejsze zakłócenia atmosferyczne i niewidzialność wiązki laserowej. Zaletą użycia laserów jest zmniejszenie wymiarów urządzeń, obniżenie kosztów, większa dokładność pomiarów oraz utrudnienie w zakłócaniu pracy urządzenia (ważne przy zastosowaniu militarnym). Korekta wyników pomiarów, której konieczność narzuca wpływ warunków atmosferycznych na prędkość światła w powietrzu, jest łatwiejsza niż w pasmie mikrofalowym. Wadę użycia laserów stanowią trudności w pracy podczas padającego śniegu lub deszczu czy mgły, gdyż wiązka laserowa jest silnie tłumiona w takich warunkach atmosferycznych. Trudność sprawia szybkie nakierowanie wąskiej wiązki laserowej na cel, co wymaga stosowania skomplikowanych układów mechanicznych „przemiatających" przestrzeń powietrzną.

zastosowanie w Kosmosie i na Ziemi

Z tych to powodów zastosowania laserowych układów śledzących dotyczą dziś głównie przestrzeni kosmicznej; są to pomiary ruchu satelitów, systemy do lądowania na Księżycu i łączenia pojazdów kosmicznych, pomiary odległości Księżyca od Ziemi. Z układów bliższych Ziemi wymienić można układ ostrzegawczy przed przeszkodami dla helikopterów oraz radar laserowy używany przez milicję do pomiaru prędkości samochodów.

Do pomiarów dalekiego zasięgu używane są przede wszystkim lasery neodymowe i CO_2. W układach bliższego zasięgu, jak w układzie NASA (posługiwano się nim w początkach lat siedemdziesiątych przy łączeniu pojazdów kosmicznych), źródłem światła jest zbiór diod laserowych GaAs, których światło jest skolimowane przez układ optyczny. Możliwości tych układów ilustruje tabela.

Podstawowe parametry układów NASA

Wielkość	Zasięg	
	daleki (3–120 km)	bliski (0–3 km)
Dokładność pomiaru odległości	0,5‰	0,1 m
Pomiar prędkości obiektu	50–120 m/s	0,3–50 m/s
Dokładność pomiaru prędkości	0,2%	0,03‰
Dokładność pomiaru położenia kątowego	0,1°	0,1°
Dokładność pomiaru prędkości kątowej	0,5 mrad/s	0,05 mrad/s

retroreflektory

Przy pomiarach bardzo dużych odległości konieczne jest — w celu zwiększenia natężenia światła wracającego do odbiornika — instalowanie na badanych obiektach specjalnych zwierciadeł, tzw. retroreflektorów, do odbijania światła w kierunku, z którego padła wiązka. Wiązka laserowa jest mało rozbieżna (kąt rozwarcia jest nawet mniejszy od 1 µrad), mimo to po przebyciu setek tysięcy czy milionów kilometrów ma znikome natężenie i trudno jest wyodrębnić sygnał z szumów (świecenie gwiazd).

Takie właśnie zwierciadła zainstalowali kosmonauci amerykańscy na Księżycu podczas wypraw „Apolla". Umożliwiły one przeprowadzenie w Lick Laboratory w Kalifornii pomiaru odległości między Ziemią i Księżycem. Źródłem światła był laser rubinowy dający impulsy o energii 8 J i o czasie trwania 10 ns. Aby maksymalnie skupić wiązkę, kierowano ją na Księżyc przez 120-calowy teleskop — na Księżycu miała średnicę ok. 1600 m. Każdy z wysłanych impulsów zawierał 10^{20} fotonów, z których przeciętnie tylko 25 wracało do odbiornika na Ziemi. Odległość Ziemia–Księżyc zmierzono tą metodą z dokładnością do ±15 cm.

Zastosowanie w medycynie i biologii

cięcie tkanek

W tych dziedzinach używa się lasera głównie dlatego, że jego promieniowanie ma dużą gęstość energii skupioną w małej objętości. Najbardziej przydatne do cięcia tkanek (miękkich i twardych) są lasery CO_2 i neodymowe. Połączenie lasera z chirurgicznym mikroskopem pozwala dokonywać precyzyjnych zabiegów, m.in. na tkance mózgowej (np. odparowywanie nowotworów mózgu). Wiązka laserowa daje ostro zarysowane cięcie z delikatnym obrzeżem, głębokość

cięcia można regulować przez zmianę mocy lub energii lasera, krwawienie jest blokowane mikroskrzepami, a rany po zabiegu goją się szybko.

koagulator laserowy

Jednym z pierwszych urządzeń medycznych, w którym zastosowano laser (rubinowy) jest koagulator laserowy (→ Fizyka medyczna). Laser rubinowy doskonale nadaje się do „przyklejania" siatkówki do naczyniówki (odklejenie siatkówki jest dość częstym uszkodzeniem oka), gdyż wiązka laserowa rozchodzi się bez dużych strat w elementach przezroczystych oka, a jest silnie pochłaniana przez nabłonek siatkówki; impuls lasera, odpowiednio skupiony, wywołuje w siatkówce odczyn zapalny, w wyniku czego powstaje zrost (il. 16, tabl. 4). Laserem rubinowym można usuwać tatuaż lub zabarwienia skóry w miejscach różniących się współczynnikiem absorpcji od miejsc sąsiednich.

mikropunkcje

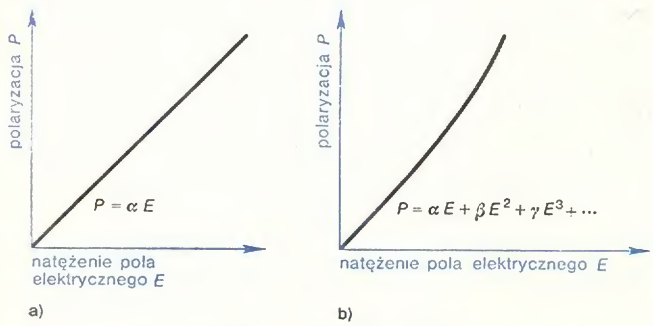
W badaniach medycznych i biologicznych lasery służą m.in. do mikropunkcji komórek i organelli. Głowica laserowa sprzężona z mikroskopem pozwala precyzyjnie zniszczyć, a nawet odparować dany fragment komórki. Wobec krótkiego czasu ekspozycji — rzędu 0,1 ms — można naświetlać nawet poruszające się elementy. Takim zabiegom na żywych komórkach nie towarzyszy upośledzenie procesów życiowych, co jest bardzo ważne przy tego typu badaniach.

J. Martin *Przyszłość telekomunikacji*, Warszawa 1975; E. R. Mustiel, W. N. Parygin *Metody modulacji światła*, Warszawa 1974; H. Kaczmarek *Wstęp do fizyki laserów*, Warszawa 1978; H. Klejman *Lasery* Warszawa 1979; A. Piekara *Nowe oblicze optyki*, Warszawa 1976; *Światło* (zbiór artykułów), Warszawa 1973

Optyka nieliniowa

Andrzej Graja

W początkach lat pięćdziesiątych wydawało się, że podstawowe problemy optyki fizycznej zostały już dostrzeżone i rozwiązane, a sama dziedzina stanowi przykład nauki pięknej, zwartej i bliskiej doskonałości. W tej idealnej konstrukcji pojawiły się jednak rysy, gdy w II poł. lat pięćdziesiątych teoretycznie wykazano, że niektóre własności materii mogą zależeć od natężenia padającego na nią światła. Okazało się mianowicie, że polaryzacja ośrodka dielektrycznego, wywołana polem elektrycznym fali świetlnej, może wzrastać szybciej aniżeli proporcjonalnie (liniowo) do natężenia pola; może być kwadratową lub wyższego rzędu funkcją tego natężenia (rys. 1), czyli może wzrastać nieliniowo wraz z natężeniem pola.

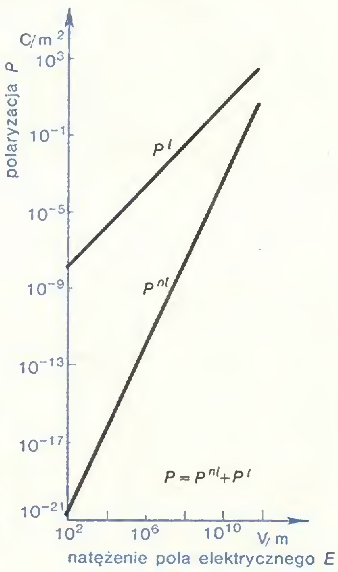


Rys. 1. Zależność polaryzacji P od natężenia przyłożonego pola elektrycznego E: a) dielektryk liniowy, b) dielektryk nieliniowy. Współczynniki α, β i γ oznaczają odpowiednio: makroskopową polaryzowalność liniową, polaryzowalność drugiego rzędu oraz polaryzowalność trzeciego rzędu

zależność polaryzacji od natężenia pola

To na pozór proste zjawisko pociągnęło za sobą całkowitą przebudowę podstaw optyki fizycznej; trzeba było bowiem znaleźć miejsce na zjawiska optyczne, których przebieg zależy od natężenia promieniowania, i które zachodzą bez spełnienia zasady superpozycji fal. W ten sposób na początku lat sześćdziesiątych pojawiła się nowa dziedzina optyki — optyka nieliniowa, obejmująca obecnie wszystkie zjawiska z zakresu częstości optycznych związane z nieliniowością materii, czyli zależnością jej własności od natężenia przyłożonego pola elektromagnetycznego. Początki optyki nieliniowej przypadają nieprzypadkowo na okres, w którym pojawiły się źródła światła zdolne do ujawnienia nieliniowości materii, a więc do weryfikacji wcześniejszych prac. Źródłami tymi były generatory kwantowe światła czyli lasery.

Aby zrozumieć, dlaczego dopiero lasery umożliwiły zaobserwowanie pierwszych zjawisk optyki nieliniowej, porównajmy wkład wnoszony do całkowitej polaryzacji ośrodka przez pierwsze dwa wyrazy αE i βE^2 występujące w wyrażeniu wiążącym polaryzację dielektryka (→ Dielektryki) z natężeniem przyłożonego pola elektrycznego fali świetlnej (rys. 2).



Rys. 2. Zależność składowej liniowej P^l i nieliniowej P^{nl} polaryzacji typowego nieliniowego ośrodka dielektrycznego od natężenia pola elektrycznego fali świetlnej (uwaga: skala logarytmiczna)

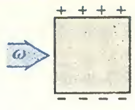
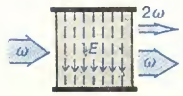
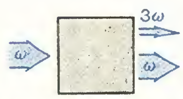
polaryzacja nieliniowa

Wyraz αE opisuje polaryzację liniową P^l ośrodka, natomiast βE^2 jest pierwszym przybliżeniem polaryzacji nieliniowej P^{nl}. W polach elektrycznych o natężeniu rzędu 10^2 V/m polaryzacja nieliniowa ośrodka jest o kilkanaście rzędów wielkości mniejsza od jego polaryzacji liniowej (mała wartość współczynnika β). W silnych polach elektrycznych fali świetlnej o natężeniu rzędu 10^{12} V/m polaryzacja nieliniowa zaczyna odgrywać pewną rolę, pozostając jednak nadal mniejszą od polaryzacji liniowej. Nikła rola polaryzacji nieliniowej w szerokim zakresie natężeń przyłożonych pól elektrycznych jest związana z małą wartością współczynnika β w omawianych materiałach. Współczynniki stojące przy wyrazach, w których natężenie pola elektrycznego fali świetlnej występuje w jeszcze wyższych potęgach, są znacznie mniejsze niż β i dlatego nieliniowość wyższego niż drugiego rzędu ujawnia się tylko w polach najsilniejszych.

Teoretyczne rozważania dotyczące ośrodków anizotropowych, czyli takich, których własności fizyczne uzależnione są od kierunku obserwacji, ukazują ogromną różnorodność zjawisk nieliniowych wywołanych falą świetlną. Najprostsze z tych zjawisk przedstawia tabela.

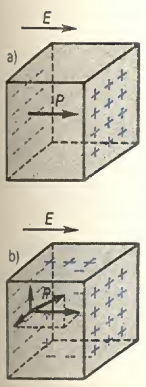
W przeciwieństwie do ośrodka izotropowego, po-

Wybrane zjawiska nieliniowe wywołane polem elektromagnetycznym fali świetlnej o częstości ω

Opis zjawiska	Schemat
stała polaryzacja dielektryka wywołana nieliniowym oddziaływaniem z falą świetlną	
wytwarzanie drugiej harmonicznej światła w kryształach bez środka symetrii	
wytwarzanie drugiej harmonicznej światła wymuszonej stałym polem elektrycznym E w kryształach ze środkiem symetrii	
wytwarzanie trzeciej harmonicznej światła	

polaryzacja ośrodka anizotropowego

laryzującego się w kierunku przyłożonego pola elektrycznego (rys. 3a), ośrodek anizotropowy polaryzuje się w innym kierunku (wektor polaryzacji nie jest równoległy do wektora natężenia pola elektrycznego), co można traktować jako polaryzację w trzech wzajemnie prostopadłych kierunkach (rys. 3b). Ogólnie, każda ze składowych pola elektrycznego wywołuje polaryzację ośrodka anizotropowego we wszystkich trzech kierunkach układu współrzędnych i dla scharakteryzowania ośrodka trzeba podać dziewięć liczb. Taki zespół liczb tworzy tensor. Inaczej mówiąc, współczynnik proporcjonalności między wektorem polaryzacji i wektorem natężenia pola nie jest liczbą, lecz wielkością tensorową, zwaną tensorem polaryzowalności liniowej. Podobnie polaryzacja nieliniowa ośrodka związana jest z natężeniem pola elektrycznego przez współczynniki tensorowe, zwane tensorami polaryzowalności kwadratowej, sześciennej lub wyższego



Rys. 3. Polaryzacja w polu elektrycznym o natężeniu E: a) ośrodek izotropowy, b) ośrodek anizotropowy

Przegląd kryształów stosowanych do wytwarzania drugiej harmonicznej światła

Kryształ	Długość fali, nm		Średnia wartość		Kąt dopasowania fazowego
	podstawowej	drugiej harmonicznej	elementów tensora b_{ijk}	długości spójności, μm	
Kwaśny fosforan potasu	694,3	347,1	1,00	9,25	50°,4
Kwaśny arsenian rubidu	694,3	347,1	0,64	158	80°
Kwaśny fosforan amonu	694,3	347,1	1,02	8,8	51°,9
Siarczan trójglicyny	694,3	347,1	0,004	4	48°
Tytanian baru	1 058,2	529,1	30	2,0	—
Niobian litu	1 058,2	529,1	29		
Tellur	10 600	5 300	12 700	40	14°,8
Arsenek indu	10 600	5 300	780	53	

tensory polaryzowalności

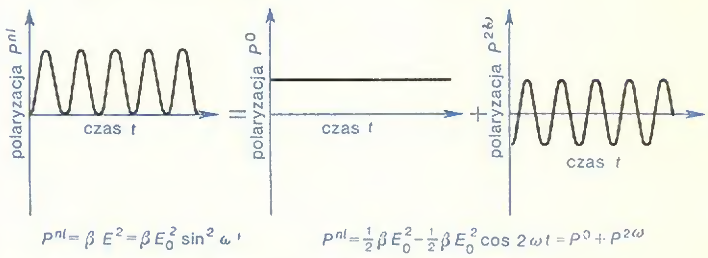
rzędu. Tensory polaryzowalności liniowej i sześciennej mają elementy różne od zera nawet dla ciał izotropowych, podczas gdy tensor polaryzowalności kwadratowej staje się zerem dla ośrodków mających środek symetrii.

Wartości elementów tensora nieliniowej polaryzowalności optycznej decydują o efektywności zjawisk nieliniowych zachodzących w rozpatrywanych ośrodkach. Tensor polaryzowalności kwadratowej b_{ijk} określa np. wydajność wytwarzania drugiej harmonicznej światła w kryształach. Kryształy bardzo się między sobą różnią wartościami elementów tego tensora, a więc ich przydatność do podwajania częstości optycznych jest zróżnicowana (tabela).

Wytwarzanie drugiej harmonicznej światła

druga harmoniczna światła

Wytwarzanie drugiej harmonicznej światła, czyli podwajanie częstości optycznych jest najprostszym i najwcześniej zaobserwowanym efektem nieliniowym (1961). Zjawisko to może występować w kryształach pozbawionych środka symetrii; kryształy takie w normalnych warunkach są dwójłomne. Jeżeli pole elektryczne fali świetlnej padającej na nieliniowy ośrodek zmienia się periodycznie z częstością ω (co opisuje funkcja $\sin \omega t$), to periodycznie zmienia się również składowa polaryzacji nieliniowej ośrodka zależna od kwadratu natężenia pola elektrycznego padającej fali. Składową tę (rys. 4) można przedstawić jako sumę dwóch funkcji: stałej P_0 i zależnej od czasu $P_{2\omega}$. Taka prosta analiza graficzna funkcji opisującej



Rys. 4. Analiza graficzna funkcji opisującej polaryzację zależną od kwadratu natężenia pola elektrycznego. Polaryzację tę można rozłożyć na składowe o częstościach 2ω i 0

polaryzację zależną od kwadratu natężenia pola elektrycznego padającej fali o częstości ω pokazuje więc, że w ośrodku nieliniowym pojawia się składowa zmienna, reprezentująca polaryzację o częstości 2ω, oraz składowa stała, opisująca polaryzację niezmienną. Pojawienie się stałej polaryzacji jest optycznym prostowaniem.

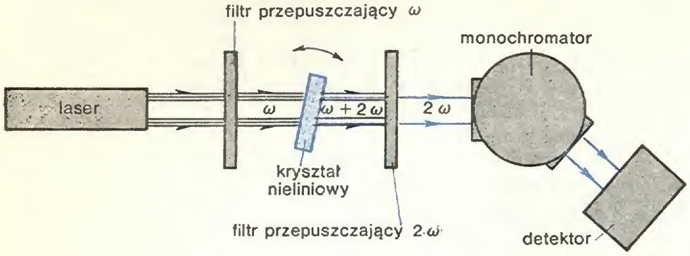
analogia z diodami krystalicznymi

Zarówno powielanie częstości jak i prostowanie fal świetlnych w krysztale nieliniowym jest pełnym analogiem dobrze znanych zjawisk występujących w diodach krystalicznych pod wpływem mikrofal i fal radiowych. Dzięki charakterystyce nieliniowej detektora krystalicznego można używać go nie tylko do prostowania fal elektromagnetycznych, ale również do mieszania oraz powielania ich częstości.

Powróćmy jednak do optyki i do składowej polaryzacji zmieniającej się periodycznie z częstością 2ω. Takie zmiany polaryzacji ośrodka to nic innego jak periodyczne zmiany makroskopowego momentu dipolowego, a co za tym idzie — emisja promieniowania elektromagnetycznego. Ośrodek nieliniowy może więc — pod wpływem fali świetlnej o częstości ω — stać się źródłem fali świetlnej o częstości 2ω. Innymi słowy, ośrodek nieliniowy może emitować promieniowanie o częstości dwukrotnie większej niż częstość fali padającej, czyli może wytwarzać drugą harmoniczną światła. Zjawisko takie nie mieści się w kategoriach optyki liniowej, można je tłumaczyć jedynie oddziaływaniem nieliniowym promieniowania elektromagnetycznego z ośrodkiem.

Uproszczony układ doświadczalny służący do wytwarzania drugiej harmonicznej światła (rys. 5) składa

się z silnego źródła spójnego światła monochromatycznego jakim jest laser (zwykle impulsowy), zestawu filtrów oraz kryształu nieliniowego, spełniającego



Rys. 5. Schemat układu doświadczalnego służącego do generacji drugiej harmonicznej światła

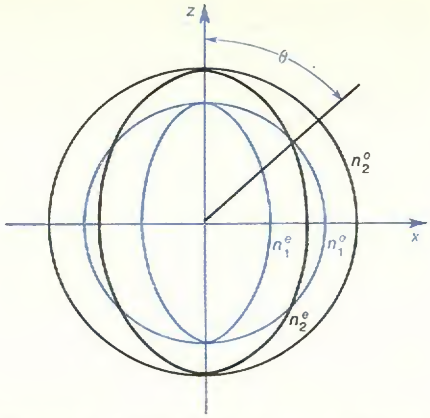
funkcję generatora fali świetlnej o podwojonej częstości. Dla przykładu, jeśli na kryształ nieliniowy skierować czerwoną wiązkę świetlną z lasera rubinowego (długość fali $\lambda_1 = 694{,}3$ nm), to otrzymuje się drugą harmoniczną niewidzialną gołym okiem, leżącą w ultrafioletowej części widma ($\lambda_2 = 347{,}1$ nm). Natomiast druga harmoniczna światła emitowanego przez laser neodymowy ($\lambda_1 = 1058$ nm, czyli w zakresie podczerwieni) ma barwę zieloną ($\lambda_2 = 529$ nm), a więc można ją obserwować wzrokowo.

W krysztale nieliniowym naświetlonym wiązką promieniowania o częstości ω pojawia się fala polaryzacji o częstości 2ω, która jest źródłem promieniowania świetlnego o częstości 2ω. Początkowo, polaryzacja i wywołana przez nią fala elektromagnetyczna są w fazach zgodnych. Później jednak powstaje między nimi różnica faz będąca wynikiem różnych prędkości rozchodzenia się tych fal. Oznacza to, że psują się warunki do „przepompowywania" energii z fali polaryzacji do fali drugiej harmonicznej. Droga optyczna w krysztale, potrzebna do tego, by fazy fal harmonicznej o częstości 2ω i polaryzacji o częstości 2ω różniły się o 180° nazywa się długością spójności. Długość spójności jest więc wielkością charakteryzującą długość drogi optycznej, na której następuje efektywne „przepompowywanie" energii z fali polaryzacji do fali drugiej harmonicznej. Jak widać z podanego wyżej przeglądu kryształów nieliniowych, długość spójności jest zwykle niewielka, więc i „przepompowywanie" energii w dowolnym kierunku kryształu jest małe. Optymalna sytuacja powstałaby wtedy, gdyby podczas całej drogi optycznej w krysztale następowało „przepompowywanie" energii, tak jak to dzieje się w wypadku przepuszczalności promieniowania przez ośrodek bez zmiany jego częstości. Okazuje się, że podobną sytuację można urzeczywistnić w krysztale nieliniowym w kierunku, w którym wektory falowe $\vec{k}_1$ (dla promieniowania o częstości ω) i $\vec{k}_2$ (dla promieniowania o częstości 2ω) spełniają równanie $\vec{k}_1+\vec{k}_1 = \vec{k}_2$. W kierunku tym prędkości rozchodzenia się fal o częstościach ω i 2ω, a co za tym idzie i współczynniki załamania światła dla tych częstości, muszą być jednakowe, czyli dopasowane. Z tego też względu kierunek najefektywniejszego „przepompowywania" energii nazywa się kierunkiem dopasowania fazowego.

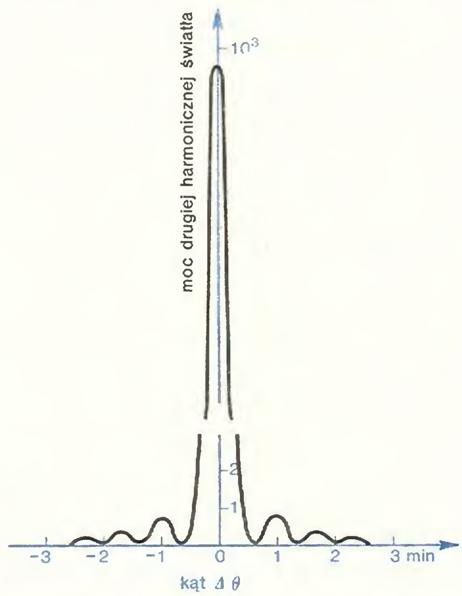
długość spójności

kierunek dopasowania fazowego

W krysztale anizotropowym kierunek wyznaczony przez przecięcie się powierzchni współczynnika załamania promienia zwyczajnego dla częstości podstawowej i promienia nadzwyczajnego (elipsoidy obrotowe) oraz dla drugiej harmonicznej (indeks 2) czyli kierunek, w którym odpowiednie współczynniki załamania są jednakowe, wyznacza położenie kierunku dopasowania fazowego. Przekrój przecinających się powierzchni współczynników załamania typowego kryształu nieliniowego, a mianowicie kwaśnego fosforanu amonu (ADP), przedstawia rys. 6. Widać z niego, że kierunek dopasowania fazowego do wytwarzania drugiej harmonicznej światła o długości fali $\lambda_1 = 694{,}3$ nm tworzy z osią optyczną kryształu kąt θ bliski 52°. Kierunek dopasowania do wytwarzania harmonicznej promieniowania podczerwonego z lasera neodymowego tworzy z osią kryształu kąt θ zbliżony do 42°. Znając wartości kąta dopasowania fazowego, można wyciąć z badanego kryształu próbkę w ten sposób, by oś optyczna kryształu tworzyła kąt θ z prostą prostopadłą do powierzchni próbki.



Rys. 6. Przekrój przez powierzchnie współczynników załamania promienia zwyczajnego (kule, indeks *o*) i nadzwyczajnego (elipsoidy obrotowe, indeks *e*) dla częstości podstawowej (indeks *1*) i drugiej harmonicznej (indeks *2*) kryształu ADP. Kierunek dopasowania fazowego tworzy z kierunkiem osi optycznej kryształu, czyli z osią *z*, kąt $\theta = 52°$

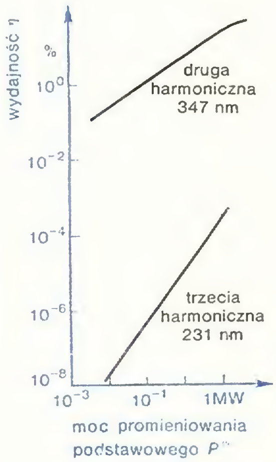


Rys. 7. Zależność mocy drugiej harmonicznej światła (w jednostkach względnych) od kąta odchylenia osi wiązki laserowej od kierunku dopasowania fazowego

Jak już powiedziano, wytwarzanie drugiej harmonicznej światła jest w zasadzie możliwe w dowolnym kierunku krystalograficznym, ale wydajność przetwarzania częstości jest wtedy przeciętnie tysiąc razy mniejsza niż w kierunku dopasowania fazowego (rys. 7). W pobliżu dopasowania fazowego zależność ta jest tym bardziej krytyczna, im mniejsza jest rozbieżność wiązki promieniowania podstawowego i im większa jest grubość kryształu nieliniowego; zależy ona również od jego jakości optycznej. Występowanie defektów sieci krystalicznej powoduje bowiem powstawanie w krysztale naprężeń i związanych z nimi lokalnych zmian współczynnika załamania. Efektem tego jest poszerzenie głównego maksimum na krzywej opisującej zależność mocy drugiej harmonicznej od

orientacji kryształu. Tak więc, do wytwarzania drugiej harmonicznej światła winno używać się kryształów możliwie najwyższej jakości optycznej.

Moc drugiej harmonicznej światła wytwarzanej w krysztale nieliniowym silnie zależy od mocy promieniowania podstawowego. Ponieważ moc promieniowania elektromagnetycznego jest proporcjonalna do kwadratu natężenia pola elektrycznego odpowiedniej fali elektromagnetycznej w wypadku efektu kwadratowego, jakim jest wytwarzanie drugiej harmonicznej światła, przeto moc promieniowania o częstości 2ω winna zależeć od kwadratu mocy promieniowania



Rys. 8. Zależność wydajności wytwarzania drugiej i trzeciej harmonicznej światła w kalcycie od mocy wiązki lasera rubinowego

podstawowego o częstości ω. Zależność ta jest dobrze spełniana w szerokim zakresie wartości mocy promieniowania (rys. 8.).

metoda proszkowa

Do badania doświadczalnego zjawiska wytwarzania drugiej harmonicznej w monokryształach próbki muszą być bardzo dokładnie wycięte z badanego monokryształu i obrobione z najwyższą dokładnością optyczną. Niekiedy jednak trudności z otrzymaniem odpowiedniej jakości i dużych monokryształów uniemożliwiają pomiary metodą tradycyjną. Stosuje się wówczas próbkę z proszku krystalicznego o przypadkowo zorientowanych ziarnach, których rozmiary są porównywalne z długością spójności dla badanego materiału. W niektórych ziarnach wytwarza się druga harmoniczna w kierunku dopasowania, w innych natomiast warunek dopasowania nie jest spełniony. Wydajność przetwarzania częstości metodą proszkową nie jest duża, ale za to taki sposób badania nieliniowych materiałów pozwala uniknąć trudności technologicznych z hodowlą, orientacją i obróbką monokryształów. Badanie drugiej harmonicznej w sproszkowanych kryształach umożliwia oszacowanie najważniejszych parametrów charakteryzujących materiał nieliniowy, a więc pozwala określić jego przydatność w optyce nieliniowej. Wygląd aparatury służącej do takich badań przedstawiono na il. 139, tabl. 35).

zastosowanie do badania kryształów

Duże zainteresowanie zjawiskiem wytwarzania drugiej harmonicznej światła jest spowodowane dwiema przyczynami: po pierwsze, jest to metoda badania materii dostarczająca nowych informacji o niej; po drugie, jest to źródło spójnego krótkofalowego promieniowania świetlnego. Badanie zjawiska powielania częstości optycznych dostarcza informacji o zmianach struktury krystalicznej i mechanizmie przejść fazowych w kryształach nieliniowych, a także informacji o ilości defektów sieci krystalicznej, o ich symetrii oraz rozkładzie. Jest to metoda szczególnie cenna w badaniach kryształów ferroelektrycznych, zwłaszcza w obszarze ich przejścia fazowego. Wytwarzanie drugiej harmonicznej światła stanowi niekiedy najwygodniejszy sposób otrzymywania spójnego promieniowania krótkofalowego, a w tym również promieniowania ultrafioletowego o bardzo dużej mocy, potrzebnych do badania optycznych własności ośrodków.

wytwarzanie spójnego promieniowania

polaryzacja kwadrupolowa

Stwierdzono poprzednio, że elementy tensora polaryzowalności kwadratowej, opisującego wytwarzanie drugiej harmonicznej światła, są równe zeru dla kryształów mających środek symetrii. Jest to słuszne w pierwszym przybliżeniu, tzn. wtedy, gdy uwzględniamy jedynie polaryzację dipolową dielektryka. Ogólnie, do polaryzacji dielektryka dochodzi jeszcze przyczynek związany z przestrzennymi zmianami pola elektrycznego (przyczynek kwadrupolowy), który odgrywa rolę w kryształach mających środek symetrii oraz w ośrodkach izotropowych. Ponieważ polaryzacja kwadrupolowa jest ok. tysiąc razy słabsza od polaryzacji dipolowej, wytwarzanie drugiej harmonicznej dzięki mechanizmowi kwadrupolowemu jest niezwykle mało skuteczne i można je zaobserwować tylko w niektórych materiałach (np. w kalcycie) i w szczególnych warunkach doświadczalnych.

wymuszone wytwarzanie drugiej harmonicznej

Przyłożenie dostatecznie silnego stałego pola elektrycznego do ośrodka izotropowego powoduje odkształcenie jego sieci krystalicznej, a co za tym idzie — utratę symetrii. Wymuszona w ten sposób anizotropia ośrodka umożliwia wytwarzanie w nim drugiej harmonicznej światła. Zjawisko wymuszonego wytwarzania drugiej harmonicznej obserwowano w niektórych kryształach centrosymetrycznych, w cieczach organicznych, w roztworach wielkocząsteczkowych związków organicznych; obserwowano je także w niektórych gazach.

procesy wielofotonowe

Na zjawisko wytwarzania drugiej harmonicznej, podobnie jak i na naturę światła, można również spojrzeć nieco inaczej. W opisie kwantowym zjawisko podwajania częstości jest najprostszym przykładem procesu wielofotonowego. W elementarnym akcie wytwarzania drugiej harmonicznej biorą bowiem udział trzy fotony: dwa fotony o jednakowych energiach równych $\hbar\omega$ oraz foton drugiej harmonicznej o energii $2\hbar\omega$. Ogólnie, wynikiem oddziaływania dwóch fal świetlnych o częstościach ω_1 (energia fotonów $\hbar\omega_1$) i ω_2 (energia $\hbar\omega_2$) w ośrodku nieliniowym jest fala o częstości ω_3 (energia $\hbar\omega_3$) spełniająca zasadę zachowania energii $\omega_3 = m_1\omega_1 \pm m_2\omega_2$, gdzie m_1 i m_2 są liczbami całkowitymi. Suma tych liczb równa się wykładnikowi najwyższej potęgi pola elektrycznego w równaniu opisującym polaryzację nieliniową (rys. 1b). Jeśli ograniczymy się do polaryzacji kwadratowej, a więc do wypadku, gdy $m_1+m_2 = 2$, uzyskamy cztery nieliniowe zjawiska optyczne drugiego rzędu:

— wytwarzanie drugiej harmonicznej wiązki pierwszej o częstości $2\omega_1$ (w tym wypadku $m_1 = 2$, $m_2 = 0$).

— wytwarzanie drugiej harmonicznej wiązki drugiej o częstości $2\omega_2$ ($m_1 = 0$, $m_2 = 2$).

— wytwarzanie wiązki o częstości sumacyjnej $\omega_1+\omega_2$ ($m_1 = m_2 = 1$).

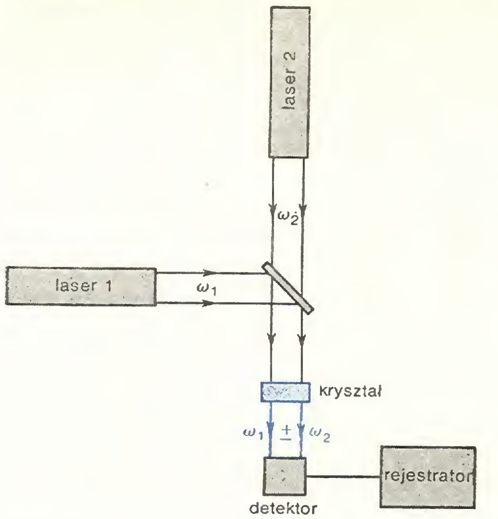
— wytwarzanie wiązki o częstości różnicowej $\omega_1-\omega_2$ ($m_1 = -m_2 = 1$).

Mieszanie wiązek świetlnych

układy doświadczalne

Mieszanie dwóch wiązek światła monochromatycznego można przeprowadzić w układzie doświadczalnym, którego uproszczony schemat przedstawia rys. 9. Wiązki promieniowania spójnego emitowane przez dwa lasery są kierowane na kryształ nieliniowy poprzez odpowiednio ustawione zwierciadło półprzepuszczalne. Zapewnia się w ten sposób współbieżność promieni ulegających zmieszaniu w krysztale nieliniowym. Pierwsze doświadczenia nad mieszaniem częstości optycznych przeprowadzono w 1962 r. z wiązkami świetlnymi pochodzącymi od dwóch laserów rubinowych pracujących w różnych temperaturach, a więc nieznacznie się różniących długością emitowanej fali. Później podobne doświadczenia powtarzano w innych warunkach i z rozmaitymi materiałami nieliniowymi. Mieszano np. wiązki promieniowania spójnego lasera rubinowego ($\lambda = 694{,}3$ nm) i lasera kry-

stalicznego emitującego światło o długości fali $\lambda = 1058,2$ nm. Zdołano zmieszać wiązki promieniowania dwóch różnych laserów gazowych pracujących

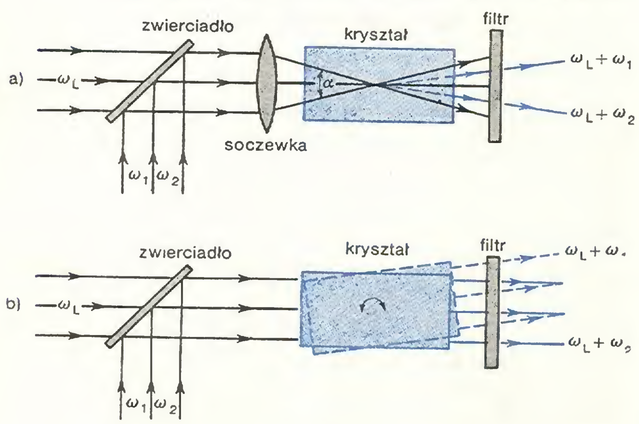


Rys. 9. Schemat układu do otrzymywania i rejestracji optycznych częstości sumacyjnych i różnicowych

w odległych zakresach widmowych, np. wiązkę podczerwoną z wiązką widzialną, przeprowadzono również badania nad mieszaniem fal podczerwonych i mikrofal. Pomyślnym wynikiem zakończyły się również próby mieszania spójnego promieniowania laserowego z promieniowaniem niespójnym emitowanym przez klasyczne źródło światła.

Podobnie jak w zjawisku wytwarzania drugiej harmonicznej światła, dużą wydajność sumowania częstości optycznych otrzymuje się jedynie w kryształach nieliniowych dwójłomnych zorientowanych w kierunku dopasowania fazowego. Wydajność przemiany częstości jest bardzo czuła na niewielkie nawet odchylenie od kierunku dopasowania fazowego. Znalazło to praktyczne zastosowanie w spektrografach i monochromatorach nieliniowych (rys. 10). W obydwu wypadkach współbieżność promieniowania laserowego o częstości ω_L i badanego światła o częstościach ω_1 i ω_2 osiąga się dzięki zastosowaniu zwierciadła przepuszczalnego dla promieniowania laserowego i silnie odbijającego promieniowanie analizowane. W spektrografach nieliniowych (rys. 10a) wiązkę światła ogniskuje się na krysztale nieliniowym. Dzięki temu promieniowanie o częstościach ω_L, ω_1 i ω_2 rozchodzi się w krysztale we wszystkich kierunkach wewnątrz kąta bryłowego α. Jeśli kierunki dopasowania fazowego mieszczą się w rozbieżności wiązki, to w kierunkach tych zachodzi efektywna generacja częstości sumacyjnych $\omega_L+\omega_1$ i $\omega_L+\omega_2$. W ten sposób uzyskuje się kątowe rozdzielenie wiązek o różnych częstościach.

spektrograf nieliniowy



Rys. 10. Zasada działania: a) nieliniowego spektrografu optycznego, b) nieliniowego monochromatora

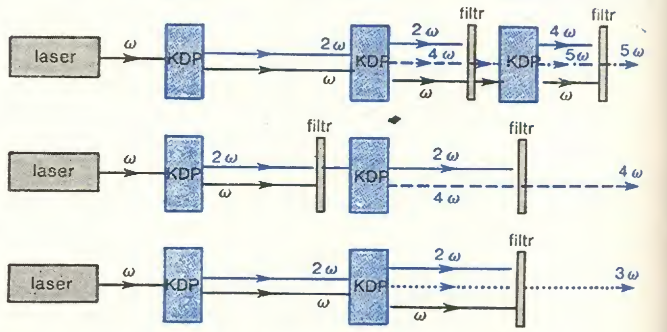
W monochromatorach nieliniowych (rys. 10b) równoległą wiązkę światła, zawierającą promieniowanie o częstościach ω_L, ω_1 i ω_2, kieruje się wprost na kryształ. Przy określonej orientacji kryształu mogą być spełnione warunki dopasowania fazowego dla procesu sumowania tylko dwóch częstości, np. $\omega_L+\omega_2$, a więc promieniowanie o takiej częstości może przejść przez kryształ nieliniowy. Dostrojenie do warunków dopasowania dla innych częstości można otrzymać przez obrót kryształu.

monochromator nieliniowy

Ze względu na trudności w otrzymywaniu spójnego promieniowania ultrafioletowego, duże znaczenie praktyczne mają metody otrzymywania wyższych harmonicznych światła. Otrzymywanie trzeciej harmonicznej światła można zrealizować dwiema metodami. Pierwsza — to wykorzystanie sześciennej polaryzowalności nieliniowych ośrodków (polaryzacja jest proporcjonalna wówczas do E^3). Druga metoda — to podwajanie częstości, a następnie sumowanie jej z promieniowaniem o częstości ω. Analiza skuteczności potrajania częstości tymi dwoma sposobami wskazuje na przewagę ostatniego. W taki też sposób, zwany kaskadowym, otrzymuje się nie tylko trzecią ale i czwartą oraz piątą harmoniczną światła. Schemat powielania kaskadowego częstości optycznych przedstawia rys. 11. Trzeba podkreślić, że wszystkie metody prowadzące do otrzymania harmonicznych zapewniają wystarczająco dużą wydajność przemiany częstości jedynie w kierunku dopasowania fazowego właściwym dla danego procesu. Niestety, realizacja dopasowania dla promieniowania o długości fali krótszej od ok. 200 nm jest niemożliwa ze względu na brak kryształów nieliniowych, mających odpowiednią dwójłomność w tym zakresie widmowym.

trzecia harmoniczna światła

wyższe harmoniczne światła



Rys. 11. Metody kaskadowego powielania częstości optycznych przy użyciu odpowiednio zorientowanych płytek płaskorównoległych wyciętych z kwaśnego fosforanu amonu (KDP)

Równoległe z mieszaniem i wydzielaniem częstości sumacyjnych, zainteresowanie badaczy jest skierowane na otrzymywanie częstości różnicowych ($\omega_1-\omega_2$). Metody wydzielania częstości różnicowych są szczególnie ważne dla spektrofotometrii w podczerwieni. Przez niewielkie przestrojenie częstości promieniowania widzialnego czy też bliskiej podczerwieni można otrzymać promieniowanie spójne z pogranicza dalekiej podczerwieni oraz mikrofal. Zmieszanie dwóch składowych wiązki lasera neodymowego daje np. promieniowanie o długości fali ok. 0,1 mm, natomiast zmieszanie składowych emitowanych przez laser rubinowy pozwala otrzymać impulsy promieniowania mikrofalowego o długości fali 0,34 mm. W wyniku zmieszania wiązek stabilizowanych laserów gazowych można natomiast otrzymać promieniowanie elektromagnetyczne o długości fali ok. 30 m, a więc należące do pasma częstości radiowych. Tak więc, przez mieszanie wiązek laserowych rozszerza się znacznie zakres widmowy obejmowany przez źródła promieniowania spójnego.

częstości różnicowe

Generatory parametryczne

Omówione wyżej metody przemiany częstości światła spójnego umożliwiają otrzymanie tylko pewnych, ściśle określonych wartości częstości, które są zdeterminowane mechanizmami zachodzących procesów. Natomiast w przezroczystych, optycznie nieliniowych ośrodkach może zachodzić tzw. parametryczne oddziaływanie fal świetlnych, umożliwiające płynne przestrajanie częstości źródeł spójnych światła. W ślad za pracami teoretycznymi, wskazującymi na możliwość zbudowania parametrycznych generatorów światła, przeprowadzono w 1965 r. doświadczenia realizujące ten pomysł.

generacja parametryczna

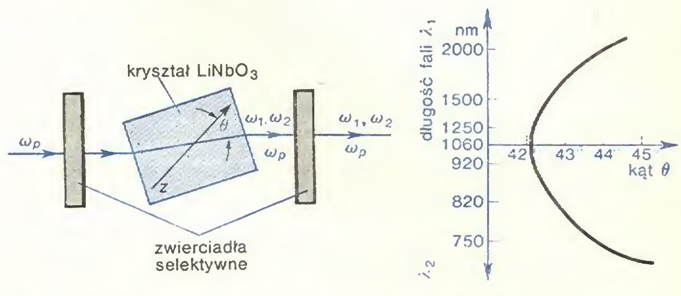
Okazuje się, że w ośrodku nieliniowym, w którym polaryzacja jest kwadratową funkcją natężenia pola elektrycznego fali świetlnej ($P = \alpha E + \beta E^2$), energia fali, zwanej pompującą, o częstości ω_p może być przekazana słabym drganiom o częstościach ω_1 i ω_2 występującym w układzie jako składowe „szumu" optycznego. Częstości te, zwane sygnałowymi muszą spełniać zasadę zachowania energii, którą można w tym wypadku zapisać w postaci $\omega_p = \omega_1 + \omega_2$ (warunek synchronizacji czasowej). Silna fala pompująca o częstości ω_p zmienia stan ośrodka, w którym się rozchodzi. W krysztale nieliniowym zmiany te polegają na modulacji przenikalności elektrycznej; mamy więc do czynienia ze zmianą parametru układu. Oznacza to równocześnie, że w takim ośrodku nieliniowym o zmiennym parametrze fale o częstości ω_1 i ω_2 nie rozchodzą się niezależnie od siebie, tak jak to byłoby w ośrodku liniowym o stałych parametrach, ale oddziałują wzajemnie. Oddziaływanie to jest maksymalne wtedy, gdy trzy rozchodzące się w ośrodku fale spełniają warunek dopasowania fazowego $\vec{k}_p = \vec{k}_1 + \vec{k}_2$ (warunek synchronizacji przestrzennej), gdzie $\vec{k}_p$, $\vec{k}_1$ i $\vec{k}_2$ oznaczają wektory falowe promieniowania o częstościach odpowiednio ω_p, ω_1 i ω_2. Przy spełnieniu tego warunku amplitudy fal o częstościach ω_1 i ω_2 narastają wykładniczo w polu fali pompującej ω_p, czyli ośrodek o zmiennym parametrze staje się wzmacniaczem fal ω_1 i ω_2. Jeśli fale sygnałowe rozchodzą się w rezonatorze nastrojonym na częstości ω_1 i ω_2, to może dojść do wzbudzenia drgań o tych właśnie częstościach. W ten sposób promieniowanie spójne pojawia się jako wynik selektywnego wzmocnienia niespójnego spontanicznego promieniowania o podanych wyżej częstościach, które może występować w układzie jako „szum" optyczny.

dopasowanie fazowe

Emisja promieniowania spójnego o częstościach ω_1 i ω_2 może zachodzić jedynie wtedy, gdy moc wiązki pompującej przekroczy wartość progową (zależną przede wszystkim od dobroci rezonatora dla częstości sygnałowych). W typowych układach parametrycznych gęstość progowa mocy jest rzędu 1 MW/cm^2, a więc można ją łatwo osiągnąć używając laserów impulsowych.

Przy ustalonej częstości ω_p promieniowania pompującego, częstości fal sygnałowych ω_1 i ω_2 — wzmacniane bądź wytwarzane w układzie parametrycznym — mogą być zasadniczo dowolne. Oczywiście muszą one spełniać podane wcześniej warunki synchronizacji czasowej ($\omega_p = \omega_1 + \omega_2$) i przestrzennej ($\vec{k}_p = \vec{k}_1 + \vec{k}_2$). Ten ostatni warunek dopasowania przy ustalonych częstościach ω_p, ω_1 i ω_2 spełniony jest w krysztale anizotropowym tylko w ściśle określonym kierunku. Oznacza to, że orientację kryształu można wykorzystać do przestrajania generatora parametrycznego. Widać to znakomicie, jeśli warunek dopasowania fazowego jest zapisany w postaci zależności między współczynnikami załamania dla odpowiednich częstości w kierunku tworzącym kąt θ z osią optyczną kryształu: $n_p(\theta) = {}^1/_2\,[n_1(\theta) + n_2(\theta)]$. W praktyce, przy zadanej częstości pompującej ω_p, wybór kąta θ determinuje częstości ω_1 i ω_2. Na przykład generator parametryczny (rys. 12) zbudowany na monokrysztale niobianu litu i pompowany promieniowaniem o długości fali 528 nm (druga harmoniczna promieniowania lasera neodymowego) można przestrajać od ok. 680 do 2350 nm przez niewielki obrót kryształu nieliniowego. Przy optymalnym wyborze struktury czasowo-przestrzennej wiązki promieniowania pompującego oraz przy dużej dobroci rezonatora, moc wiązek emitowanych przez optyczne układy parametryczne wynosi ok. kilkuset kW.

przestrajanie generatora parametrycznego przez obrót kryształu



Rys. 12. Układ do przestrajania i charakterystyka generatora parametrycznego pracującego na krysztale niobianu litu. Przestrajanie można zrealizować przez obrót nieliniowego kryształu umieszczonego w rezonatorze optycznym o dużej dobroci dla częstości ω_1 i ω_2

W innej metodzie przestrajania generatora parametrycznego wykorzystuje się temperaturową zależność współczynnika załamania kryształu. Wynikiem zmiany temperatury jest przesuwanie się położenia kierunku dopasowania dla danej pary częstości sygnałowych. Innymi słowy, zmiany temperatury nieruchomego kryształu powodują, że warunek dopasowania w ustalonym kierunku krystalograficznym spełniają coraz to inne pary częstości ω_1 i ω_2. W generatorze parametrycznym pracującym na krysztale niobianu litu i pompowanym promieniowaniem o długości fali 528 nm uzyskuje się — przy zmianie temperatury kryształu od 50°C do 60°C — przestrajanie generatora w zakresie od ok. 970 nm do ok. 1150 nm.

przestrajanie przez zmianę temperatury

Generator parametryczny można również płynnie przestrajać, wykorzystując zależność współczynnika załamania kryształu od natężenia stałego pola elektrycznego (metoda elektrooptyczna). Zależność ta jest szczególnie duża w pobliżu ferroelektrycznego przejścia fazowego i dlatego w temperaturze zbliżonej do temperatury przejścia uzyskuje się tą metodą największe przestrajanie częstości.

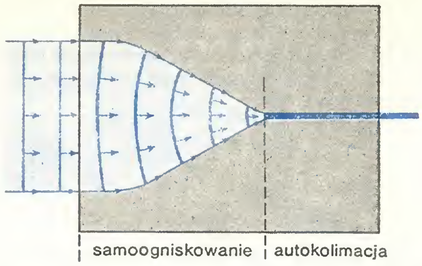
elektrooptyczna metoda przestrajania

Generatory parametryczne emitujące promieniowanie spójne można przestrajać w bardzo szerokim zakresie widmowym od dalekiej podczerwieni (powyżej 25 µm) do fioletu (ok. 300 nm). Krótkofalową granicę można oczywiście obniżyć, powielając częstość fali świetlnej wytwarzanej przez układ parametryczny. Tak więc, generatory parametryczne mogą być stosowane jako źródła spójnego promieniowania świetlnego o częstości, którą można dokładnie dostosować do potrzeb doświadczenia. Ma to szczególne znaczenie w badaniach optycznych zjawisk rezonansowych.

Samoogniskowanie i autokolimacja

Omawiając nieliniowe oddziaływanie światła z materią, w którego wyniku następuje zmiana częstości promieniowania, zakładaliśmy milcząco, że mamy do czynienia z falami płaskimi, przestrzennie nieograniczonymi. Założenie to przestaje być jednak słuszne, gdy wiązka świetlna o ograniczonej średnicy rozchodzi się w ośrodku nieliniowym. W takim wypadku natężenie pola elektrycznego fali świetlnej zmienia się w płaszczyźnie prostopadłej do kierunku rozchodzenia się, a płaszczyzna ta przestaje być powierzchnią stałej fazy.

Ponieważ wartość współczynnika załamania $n(E)$ ośrodka nieliniowego zależy od natężenia pola elektrycznego fali świetlnej: $n(E) = n_0 + n_2E^2 + \ldots$, zatem



Rys. 13. Samoogniskowanie i autokolimacja silnej wiązki świetlnej w ośrodku nieliniowym. Widać stopniową deformację powierzchni fazowych w ośrodku

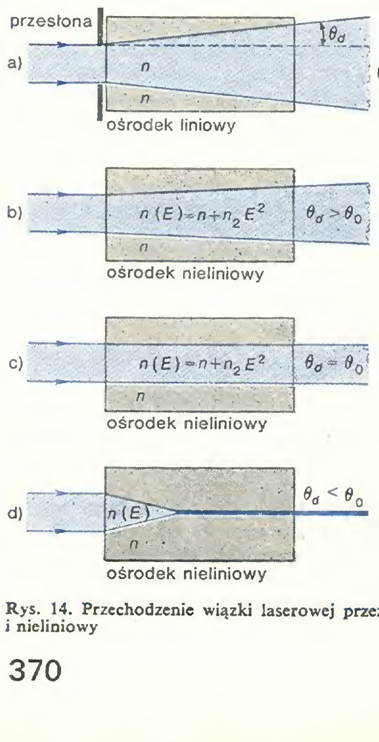
wiązka intensywnego światła rozchodzi się w obszarze o zmienionym współczynniku załamania, większym od współczynnika n_0 mierzonego dla światła o niewielkim natężeniu. Ośrodek nieliniowy działa więc podobnie do soczewki skupiającej, koncentrując strumień energii świetlnej w pobliżu osi wiązki. Ilustruje to rys. 13, na którym schematycznie przedstawiono ten mechanizm ogniskowania światła.

Wobec tego, że promień odchyla się w kierunku ośrodka o większym współczynniku załamania, obserwuje się dodatkowy wzrost natężenia pola elektrycznego wzdłuż osi wiązki, a co za tym idzie — wzrost współczynnika załamania. W związku z tym zjawisko ogniskowania samo się potęguje, przyjęto więc określać je jako samoogniskowanie światła.

samo-ogniskowanie

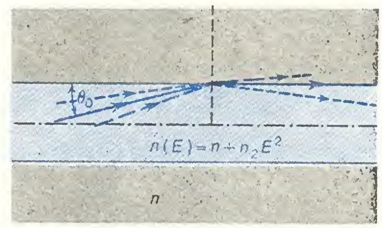
Geometryczna analogia między ogniskowaniem światła za pomocą soczewki i samoogniskowaniem w ośrodku nieliniowym kończy się po skupieniu wiązki. Za ogniskiem soczewki skupiającej wiązka świetlna staje się rozbieżna, podczas gdy światło, które uległo samoogniskowaniu rozchodzi się dalej w postaci bardzo cienkiej, skolimowanej nici. Zjawisko takie, zwane autokolimacją, może wystąpić w ośrodku nieliniowym przy dostatecznie dużych mocach wiązki świetlnej. Elementarną analizę zjawiska autokolimacji światła można przeprowadzić, korzystając z praw optyki geometrycznej oraz wspomnianej wyżej zależności nieliniowej współczynnika załamania od natężenia pola elektrycznego fali świetlnej. Wiązka świetlna,

autokolimacja



Rys. 14. Przechodzenie wiązki laserowej przez ośrodek liniowy i nieliniowy

nawet idealnie skolimowana, po przejściu przez przesłonę wykazuje słabą rozbieżność spowodowaną ugięciem. Tak więc, światło za przesłoną (rys. 14) wypełnia stożek o rozwartości θ_d. Rozbieżność dyfrakcyjna θ_d jest bardzo mała, czyli stożek świetlny praktycznie nie różni się od walca. Promienie padające od strony ośrodka optycznie gęstszego, przechodząc do optycznie rzadszego, mogą — przy dostatecznie dużych kątach padania — ulegać całkowitemu wewnętrznemu odbiciu (rys. 15). Jeśli rozbieżność dyfrakcyjna θ_d jest większa od kąta krytycznego θ_0, to promienie wychodzą z wiązki, natomiast promienie o rozbieżności mniejszej od kąta krytycznego wracają do środka wiązki. Porównując rozbieżność dyfrakcyjną θ_d z wartością kąta θ_0, można ocenić wpływ nieliniowej refrakcji i dyfrakcji na rozchodzenie się wiązki świetlnej w ośrodku. Jeśli $\theta_d > \theta_0$, to wiązka jest rozbieżna (rys. 14b), przy tym rozwartość jej stożka świetlnego jest mniejsza niż w ośrodku liniowym (rys. 14a). Jeśli $\theta_d = \theta_0$, to nieliniowa refrakcja kompensuje dyfrakcję; wiązka rozchodzi się jakby w optycznym, cylindrycznym falowodzie utworzonym przez siebie (rys. 14c). To zjawisko nazwano autokolimacją światła, a warunek $\theta_d = \theta_0$ wykorzystano do określenia kry-



Rys. 15. Powstawanie kanału autokolimacyjnego w nieliniowym ośrodku. W obszarze rozchodzenia się wiązki świetlnej współczynnik załamania wzrasta do wartości $n(E) = n+n_2E^2$ (kąt krytyczny θ_0 jest kątem dopełniającym do kąta całkowitego wewnętrznego odbicia)

warunek autokolimacji

tycznej mocy wiązki laserowej niezbędnej do jego wywołania. Okazuje się, że moc ta — w wypadku powszechnie używanych rozpuszczalników organicznych — nie jest zbyt wysoka i wynosi od kilkunastu do kilkudziesięciu kilowatów. Sytuacja przedstawiona na rys. 14d występuje, gdy $\theta_d < \theta_0$ oraz gdy moc wiązki przewyższa moc krytyczną. Promieniowanie załamuje się w kierunku osi wiązki, powodując najpierw samoogniskowanie światła, a po osiągnięciu dostatecznie małej średnicy wiązki rozpoczyna się autokolimacja. Trzeba podkreślić, że skolimowana w ten sposób wiązka ma średnicę ok. 50 μm i nie jest jednorodna. Składa się ona bowiem z wielu bardzo cienkich nici świetlnych. Energia świetlna zawarta w takiej nici może zostać po pewnym czasie zużyta na zjonizowanie molekuł ośrodka, w którym zjawisko autokolimacji zachodzi. Nić więc przestaje istnieć, a na jej miejsce może się pojawić nowa.

molekularny mechanizm autokolimacji

Z makroskopowego punktu widzenia samoogniskowanie i autokolimacja są możliwe dzięki wzrostowi współczynnika załamania światła ośrodka nieliniowego pod wpływem pola elektrycznego fali świetlnej. Zastanówmy się, jaki jest molekularny mechanizm powodujący wzrost tego współczynnika. Trzeba sobie zdać sprawę, że podstawowy mechanizm będący podstawą omawianych zjawisk musi być niezwykle szybki, jego czas relaksacji powinien być rzędu najwyżej 10^{-12} s. Wynika to z doświadczenia — stwierdzono bowiem, że pikosekundowe impulsy świetlne, a więc impulsy o czasie trwania ok. 10^{-12} s, wywołują również samoogniskowanie i autokolimację światła. Ponieważ omawiane zjawiska zachodzą zarówno w cieczach jak i ciałach stałych, podstawowy mechanizm winien się stosować do obydwu stanów skupienia.

Stwierdzono, że zasadniczą przyczyną samoogniskowania i autokolimacji światła jest wzrost polaryzowalności molekuł ośrodka nieliniowego w polu elektrycznym fali świetlnej. Oprócz tego podstawowe-

go mechanizmu pewną rolę odgrywają również inne procesy molekularne, a wśród nich reorientacja molekuł w polu fali świetlnej i ich hiperpolaryzowalność.

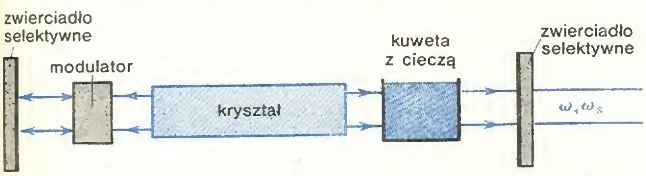
Wymuszone rozpraszanie światła

Od ponad stu lat wiadomo, że światło przechodząc przez ośrodki mętne, czyli niejednorodne optycznie, ulega bardzo silnemu rozpraszaniu. Z dokładnych obserwacji wynika, że również ośrodki przezroczyste, makroskopowo jednorodne, mogą rozpraszać światło. Rozpraszanie zachodzi wówczas np. na statystycznych fluktuacjach gęstości. Światło rozproszone przez układ atomów lub molekuł nie pochłaniających zachowuje niezmienioną długość fali, a zjawisko nazywa się rozpraszaniem Rayleigha. Oprócz takiego elastycznego rozpraszania występuje często nieelastyczne rozpraszanie światła, czyli zachodzące ze zmianą długości fali. W świetle rozproszonym, oprócz częstości podstawowej ω, obserwuje się wtedy częstości $\omega+n\Omega$ zw. częstościami antystokesowskimi i częstości $\omega-n\Omega$ zwane stokesowskimi, przy czym Ω jest równe częstości drgań oscylacyjnych lub rotacyjnych molekuł rozpraszających. Wiązka światła rozproszonego w układzie molekularnym niesie więc informacje o tym układzie. Takie nieelastyczne rozpraszanie światła nazywamy rozpraszaniem Ramana (→ Spektroskopia molekularna). Trzeba podkreślić, że zarówno rozpraszanie Ramana jak i Rayleigha są zjawiskami spontanicznymi, a więc rozproszone promieniowanie jest niespójne, a samo zjawisko nie ma progu energetycznego.

rozpraszanie Rayleigha

rozpraszanie Ramana

Sytuacja uległa radykalnej zmianie po zastosowaniu spójnego promieniowania laserowego w badaniach rozpraszania światła. Uproszczony układ doświadczalny umożliwiający takie badania przedstawia schematycznie rys. 16. W rezonatorze laserowym, utworzonym przez selektywne zwierciadła dielektryczne, oprócz kryształu aktywnego i komórki modulującej dobroć rezonatora jest umieszczone naczynie wypeł-



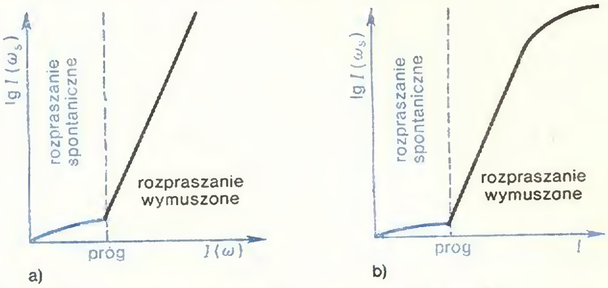
Rys. 16. Schemat układu służącego do badania rozpraszania światła laserowego

nione cieczą, w której zachodzi zjawisko Ramana. Natężenie pola elektrycznego fali świetlnej w rezonatorze takiego lasera osiąga wartości rzędu 10^8–10^{10} V/m, a więc zbliżone do natężeń pól wewnątrzatomowych. W tak silnych polach ujawnia się już wpływ ruchu elektronów na drgania całych molekuł.

Pojawienie się w rezonatorze laserowym promieniowania rozproszonego o częstościach ω_{as} i ω_s, będącego wynikiem spontanicznego rozpraszania Ramana, wpływa decydująco na przebieg zjawisk w rezonatorze. Promieniowanie to staje się bowiem czynnikiem wymuszającym dalszą emisję promieniowania rozproszonego. Fazy drgań poszczególnych molekuł rozpraszających światło stają się jednakowe, a promieniowanie rozproszone na tak drgających molekułach jest promieniowaniem spójnym. Oznacza to, że spontaniczne, niespójne początkowo, rozpraszanie staje się rozpraszaniem spójnym. Tak przebiegające zjawisko nazywa się wymuszonym rozpraszaniem Ramana. Warunkiem wymuszonego rozpraszania światła jest duża dobroć optyczna rezonatora, w którym to rozpraszanie zachodzi. Obecność zwierciadeł dielektrycznych w układzie doświadczalnym przedstawionym na rys. 16 zapewnia spełnienie tego warunku.

wymuszone rozpraszanie Ramana

Prawa rządzące rozpraszaniem światła wynikają z rozważań nad zależnością polaryzacji ośrodka od natężenia E pola elektrycznego fali świetlnej. Jeśli wziąć pod uwagę ośrodek złożony z molekuł polaryzowalnych, tzn. takich, w których rozkład ładunku elektrycznego zależy od natężenia zewnętrznego pola elektrycznego, to w polach o natężeniach rzędu 10^9–10^{10} V/m polaryzację opisują trzy wyrazy zależne od natężenia pola E. Dwa z nich są liniowymi funkcjami E i, jak można wykazać, opisują rozpraszanie Rayleigha i Ramana, trzeci natomiast jest członem nieliniowym, zależnym od trzeciej potęgi pola E. Człon ten opisuje zjawisko wymuszonego rozpraszania Ramana. Faza nieliniowej składowej polaryzacji jest określona jednoznacznie przez pole fali laserowej. Ponieważ spójne pole fali świetlnej emitowanej przez laser działa na wszystkie molekuły, to i fazy drgań ich momentów dipolowych, a co za tym idzie i polaryzacji, będą równe fazie promieniowania laserowego.



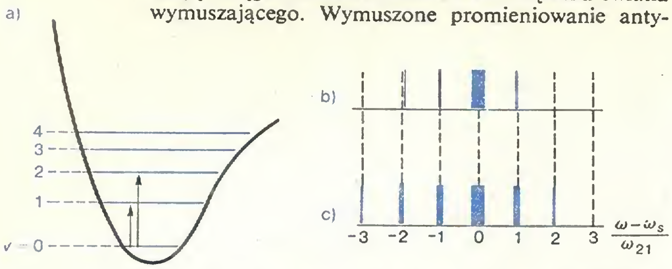
Rys. 17. Próg wymuszonego rozpraszania ramanowskiego: a) energetyczny, b) optyczny. Natężenie promieniowania stokesowskiego $I(\omega_s)$ rośnie wykładniczo wraz ze wzrostem natężenia promieniowania padającego $I(\omega)$. Wykładniczy charakter ma również zależność $I(\omega_s)$ od długości kuwety l (do pewnej długości maksymalnej)

Wymuszone rozpraszanie Ramana różni się pod wieloma względami od zwykłego, spontanicznego rozpraszania światła. Powyżej progu zjawiska natężenie wymuszonego promieniowania rozproszonego $I(\omega_s)$ rośnie bardzo szybko (wykładniczo) ze wzrostem natężenia światła padającego $I(\omega)$ (rys. 17a). Obecność progu energetycznego i odmienny przebieg zjawiska po jego przekroczeniu są charakterystyczne dla efektów wymuszonych. Należy podkreślić, że próg wymuszonego rozpraszania Ramana jest często zniekształcony innym progowym zjawiskiem nieliniowym, a mianowicie samoogniskowaniem światła. Istnieje również progowa wartość drogi optycznej promieniowania (rys. 17b) niezbędna do wystąpienia wymuszonego rozpraszania światła. Stopień spójności wymuszonego promieniowania ramanowskiego zbliżony jest do stopnia spójności promieniowania laserowego rozpraszanego na badanym układzie molekularnym. O spójności promieniowania rozproszonego świadczą nie tylko odpowiednie doświadczenia interferencyjne, ale również charakterystyczne dla emisji wymuszonej zwężenie linii widmowych. Szerokości linii stokesowskich w spontanicznym zjawisku Ramana wynoszą ok. 10 cm^{-1}, podczas gdy w zjawisku wymuszonym ulegają one zwężeniu do ok. 0,5 cm^{-1}; jeszcze węższe są linie antystokesowskie, których szerokość widmowa wynosi ok. 0,05 cm^{-1}.

różnice między wymuszonym i spontanicznym rozpraszaniem Ramana

W spontanicznym rozpraszaniu Ramana obserwuje się głównie przejścia, dla których oscylacyjna liczba kwantowa v zmienia się o jedność, tzn. $\Delta v = \pm 1$; mało prawdopodobne są natomiast przejścia z większą zmianą liczby kwantowej. Dlatego też natężenie spontanicznego promieniowania rozproszonego silnie maleje ze wzrostem Δv. Ze względu na anharmoniczność drgań molekularnych poziomy oscylacyjne nie są równoległe (rys. 18a), a pasma ramanowskie odpowiadające $\Delta v \neq 1$ mają częstości różniące się od dokładnych wiartości nadtonów (rys. 18b). W zjawisku wymuszonym natomiast natężenia linii stokesowskich,

dla których $\Delta v > 1$, mogą być bardzo duże i niekiedy mogą osiągać wartości zbliżone do natężenia światła wymuszającego. Wymuszone promieniowanie anty-

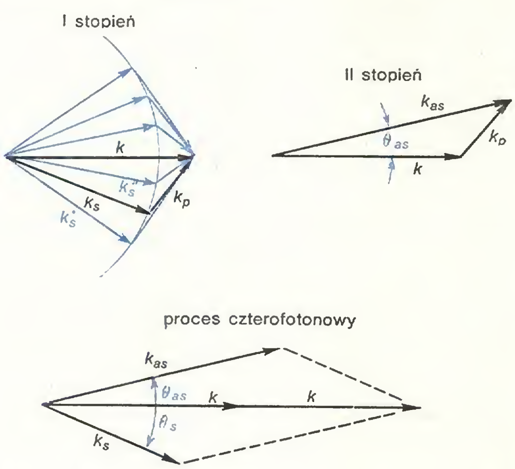


Rys. 18. Schemat poziomów energetycznych oscylatora anharmonicznego (a) oraz odpowiadające mu widma rozpraszania Ramana spontanicznego (b) i wymuszonego (c)

stokesowskie występuje również w najniższych temperaturach, a jego natężenie rośnie niekiedy w miarę oziębiania układu molekularnego. Sugeruje to, że wymuszone promieniowanie antystokesowskie związane jest z molekułami w podstawowym stanie energetycznym, a nie w stanie wzbudzonym. W przeciwieństwie do zjawiska spontanicznego, położenia pasm wymuszonego rozpraszania Ramana odpowiadają dokładnym wartościom częstości harmonicznych (rys. 18c).

rozkład kątowy promieniowania rozproszonego

Bardzo istotna różnica w obydwu zjawiskach Ramana ujawnia się w rozkładzie kątowym promieniowania rozproszonego. W zjawisku spontanicznym natężenie promieniowania rozproszonego jest funkcją ciągłą, słabo zależną od kąta obserwacji. W zjawisku wymuszonym natężenie składowej stokesowskiej jest również funkcją ciągłą kąta obserwacji, ale silnie zależną od geometrii układu doświadczalnego, a przeważająca część promieniowania jest emitowana w kierunku osi wiązki laserowej w stożku o niewielkiej rozwartości. Składowa antystokesowska zaś rozchodzi się jedynie w dokładnie określonym kierunku tworzącym kąt θ_{as} z osią wiązki laserowej. Pojawieniu się składowej antystokesowskiej towarzyszy spadek natężenia składowej stokesowskiej emitowanej pod kątem θ_s związanym z θ_{as} prostą relacją trygonometryczną wynikającą z rys. 19.



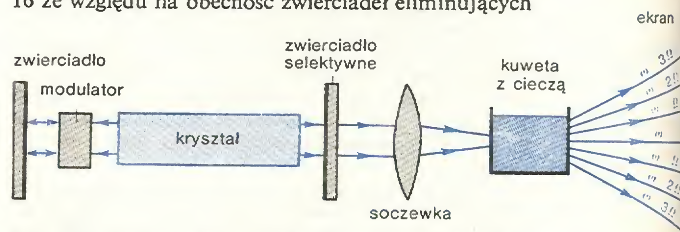
Rys. 19. Wymuszone rozpraszanie Ramana jako proces czterofotonowy. Górna część ilustruje dwustopniowy charakter rozpraszania

proces czterofotonowy

Zjawisko wymuszonego rozpraszania Ramana jest procesem czterofotonowym, w którym na miejsce dwóch fotonów światła laserowego pojawiają się dwa nowe fotony: stokesowski i antystokesowski. To, że charakterystyki kątowe promieniowania rozproszonego o częstościach ω_s i ω_{as} są istotnie różne, sugeruje odmienny mechanizm powstawania tych dwóch pasm. Przyjęto więc hipotezę, że rozpraszanie wymuszone jest procesem dwustopniowym. Pierwszy stopień to przemiana jednego z fotonów laserowych opisywanych wektorem falowym $\vec{k}$ w foton stokesowski opisywany wektorem $\vec{k_s}$. Częstość promieniowania laserowego jest większa od częstości stokesowskiej i dlatego nie całą energię tego promieniowania odnajdziemy w promieniowaniu rozproszonym. Różnica energii zostaje przekazana sieci jako energia fononu, któremu przypisujemy wektor falowy $\vec{k_p}$ (→ Dynamika sieci krystalicznej). Energia ta wydziela się w końcu w postaci ciepła. Pierwszy proces nie nakłada żadnych ograniczeń na kierunek emisji fotonu stokesowskiego, a więc prowadzi on do emisji dowolnie skierowanych fotonów stokesowskich i odpowiadających im fononów. Drugi stopień to wytwarzanie fotonu antystokesowskiego z drugiego fotonu laserowego oraz odpowiedniego fononu. Proces ten przebiega zgodnie z zasadą zachowania pędu, którą można zapisać w postaci równania wektorowego $\vec{k}+\vec{k_p} = \vec{k}_{as}$.

obserwacja składowej antystokesowskiej

Z powyższego wywodu widać, że emisja fotonu antystokesowskiego może zachodzić jedynie pod kątem θ_{as}, wynikającym z powyższej zależności wektorowej. Promieniowanie antystokesowskie, rozproszone pod kątem θ_{as} względem osi wiązki laserowej, nie może się pojawić w układzie przedstawionym na rys. 16 ze względu na obecność zwierciadeł eliminujących



Rys. 20. Schemat układu doświadczalnego umożliwiającego obserwację składowej antystokesowskiej wymuszonego rozpraszania Ramana

fotony nieosiowe. Aby więc zaobserwować składowe antystokesowskie, trzeba komórkę z cieczą umieścić poza rezonatorem laserowym w nierównoległej wiązce świetlnej. Prosty układ doświadczalny umożliwiający takie badania przedstawia rys. 20. (Promieniowanie jest rozpraszane wzdłuż pobocznic stożków, których osiowe przekroje zaznaczono na rysunku.) Kolejnym składowym antystokesowskim $\omega_{as} = \omega + n\Omega$ odpowiadają stożki o coraz to większej rozwartości. Tak więc na ekranie ustawionym prostopadle do osi układu doświadczalnego obserwuje się współosiowe barwne pierścienie odpowiadające poszczególnym wiązkom światła rozproszonego (il. 2, tabl. 1). Wynikiem rozpraszania światła w innych ośrodkach są podobne obrazy współśrodkowych barwnych pierścieni o przesunięciach częstości charakteryzujących rozpraszający ośrodek.

W wymuszonym rozpraszaniu Ramana występują zarówno charakterystyczne cechy zwyczajnego rozpraszania Ramana jak również i cechy promieniowania spójnego, emitowanego przez lasery. Emisja wymuszona w generatorach kwantowych możliwa jest dzięki inwersji obsadzeń odpowiednich poziomów energetycznych. W przypadku wymuszonego rozpraszania światła spójne wzmacnianie pola o częstości stokesowskiej ω_s zachodzi kosztem osłabienia pola pompującego. Układ umożliwiający takie wzmacnianie promieniowania nie wymaga więc wystąpienia inwersji obsadzeń poziomów energetycznych; jest to zasadnicza różnica między laserem a laserem ramanowskim. Laser ramanowski jest przykładem praktycznego zastosowania zjawiska wymuszonego rozpraszania światła. Jego znaczenie polega przede wszystkim na tym, że częstość promieniowania emitowanego przez urządzenie można zmienić łatwo i w szerokich granicach przez dobór właściwej cieczy rozpraszającej. Wymuszone rozpraszanie Ramana znalazło zastosowanie

laser ramanowski

w laboratoriach fizykochemicznych jako narzędzie badań oddziaływań wewnątrz- i międzymolekularnych, czyli jako metoda spektroskopowa znacznie czulsza od stosowanej szeroko klasycznej spektroskopii ramanowskiej.

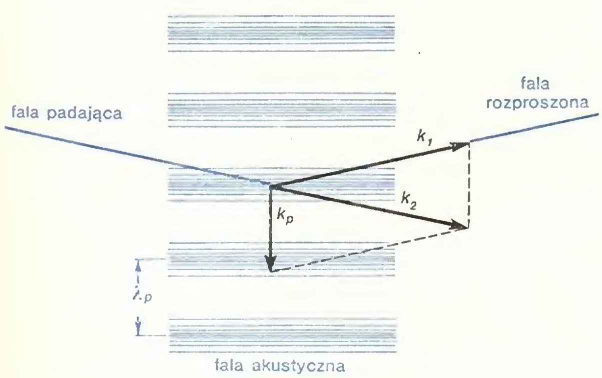
Rozpraszanie światła na fali akustycznej

Wymuszone rozpraszanie Ramana jest wynikiem oddziaływania fal elektromagnetycznych z poszczególnymi molekułami ośrodka. Innymi słowy, rozpraszanie Ramana odzwierciedla własności pojedynczych molekuł. Jeśli molekuły ośrodka rozpraszającego światło są uporządkowane i tworzą sieć krystaliczną, to padające promieniowanie świetlne oddziałuje nie tylko z poszczególnymi molekułami, ale może wzbudzić drgania tejże sieci, czyli falę akustyczną. Częstość tzw. gałęzi optycznej drgań sieci jest równa różnicy częstości światła wzbudzającego i rozpraszanego.

rozpraszanie Brillouina

Rozpraszanie światła na fali akustycznej (ultradźwiękowej) rozchodzącej się w ośrodku skondensowanym nazywa się rozpraszaniem Brillouina. Zjawisko staje się efektem wymuszonym, jeśli fala akustyczna, na której rozprasza się światło, wzbudzona jest przez to samo promieniowanie.

Zastanówmy się, jaki jest fizyczny mechanizm rozpraszania światła na fali akustycznej. Rozchodzenie się fali ultradźwiękowej o częstości ω_p w ośrodku można sobie wyobrazić jako rozprzestrzenianie się zagęszczeń i rozrzedzeń tego ośrodka z prędkością v_p. Odpowiada to periodycznym zmianom przenikalności elektrycznej i współczynnika załamania światła. Zmiany współczynnika załamania powodują częściowo odbicie padającej wiązki o częstości ω_2 (wektor falowy $\vec{k}_2$) w kierunku $\vec{k}_1$ (rys. 21). Tak więc fala akustyczna



Rys. 21. Odbicie światła o częstości ω_2 na periodycznych zagęszczeniach i rozrzedzeniach ośrodka spowodowanych propagacją fali akustycznej o częstości ω_p

rozchodząca się w ośrodku działa podobnie do siatki dyfrakcyjnej o stałej λ_p równej długości fali ultradźwiękowej. Istotna różnica między tymi dwoma elementami rozpraszającymi polega na tym, że „siatka akustyczna" przemieszcza się. Oznacza to, że promień odbity doznaje dopplerowskiego przesunięcia częstości prowadzącego w tym wypadku do obniżenia częstości fali odbitej ($\omega_1 < \omega_2$). Okazuje się, że różnica częstości promieniowania padającego i odbitego jest równa częstości fali akustycznej ($\omega_2 - \omega_1 = \omega_p$). Fala świetlna oddziałując z falą mechaniczną (akustyczną) przekazuje jej energię za pośrednictwem efektu elektrostrykcyjnego. Obecność w ośrodku fali o częstości ω_2 może doprowadzić więc do wzmocnienia fal ω_1 i ω_p. Jeśli to wzmocnienie wystarcza na skompensowanie strat i jeśli w układzie występuje dodatnie sprzężenie zwrotne, to wzbudzają się drgania elektromagnetyczne o częstości ω_1 i akustyczne o częstości ω_p. Taka generacja fal ω_1 i ω_p nazywa się wymuszonym rozpraszaniem Brillouina.

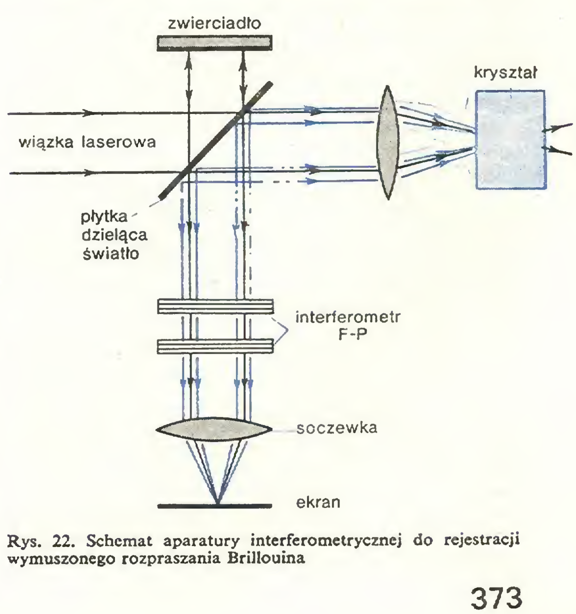
wymuszone rozpraszanie Brillouina

Rozchodząca się w ośrodku nieliniowym fala ultradźwiękowa ma bardzo dużą częstość, zazwyczaj w granicach 10^9–10^{10} s^{-1}. Zważywszy, że prędkość jej rozchodzenia się w ośrodku skondensowanym jest rzędu 10^3 m/s, oznacza to, że jej długość wynosi ok. 10^{-6} m. Jest to więc wielkość o trzy rzędy wielkości większa od typowych odległości międzyatomowych. Trzeba podkreślić, że częstość fali ultradźwiękowej jest jednak dużo mniejsza od częstości fal świetlnych ω_1 i ω_2 wynoszących w typowych doświadczeniach z laserem rubinowym ok. $5 \cdot 10^{-14}$ s^{-1}.

Rozpraszanie Brillouina można opisać kwantowo jako pojawienie się fotonu o częstości ω_1 i fononu o częstości ω_p zamiast fotonu o częstości ω_2. W opisie korpuskularnym równanie $\omega_2 - \omega_1 = \omega_p$ wyraża zasadę zachowania energii, a obowiązująca równocześnie zasada zachowania pędu wymaga, by $\vec{k}_2 = \vec{k}_1 + \vec{k}_p$. Ten warunek wektorowy mówi, że emisja tak fali akustycznej, jak i optycznej rozproszonej zachodzi w ściśle określonych kierunkach. Ponieważ prędkość propagacji fali akustycznej v_p jest znacznie mniejsza od prędkości rozchodzenia się światła, wektory $\vec{k}_2$ i $\vec{k}_1$ mają bardzo zbliżone długości, a częstość ω_2 bardzo nieznacznie różni się od ω_1. W doświadczeniu rozpraszanie Brillouina objawia się jako nadsubtelna struktura rozpraszania Rayleigha. Obserwowane doświadczalnie przesunięcia częstości dla wielu cieczy, jak np. anilina, benzen, toluen i woda, są rzędu 10^{-1} cm^{-1}, a dla kryształów są jeszcze mniejsze. Natomiast pasma w widmie ramanowskim doznają zmian częstości rzędu kilkuset cm^{-1}. Rozpraszanie Brillouina trzeba więc badać bardzo precyzyjnymi metodami — stosuje się np. spektrograf z interferometrem Fabry'ego–Pérota.

obserwacja wymuszonego rozpraszania Brillouina

Schemat aparatury umożliwiającej rejestrację wymuszonego rozpraszania Brillouina przedstawia rys. 22. Wiązka spójnego światła laserowego dużej mocy jest ogniskowana w badanej próbce. Promieniowanie rozproszone w kierunku wstecznym kieruje się do interferometru Fabry'ego–Pérota za pomocą zwierciadła półprzepuszczalnego ustawionego w wiązce pod kątem 45°. Obraz pierścieni interferencyjnych otrzymuje się na płycie fotograficznej ustawionej za długoogniskową soczewką. Okazuje się, że po przekroczeniu progowej wartości mocy wiązki laserowej na interferogramie pojawiają się dodatkowe pierścienie odpowiadające promieniowaniu rozproszonemu ze zmienioną częstością. Są one przesunięte w stosunku do pierścieni interferencyjnych odpowiadających wiązce



Rys. 22. Schemat aparatury interferometrycznej do rejestracji wymuszonego rozpraszania Brillouina

laserowej o dziesiąte części cm^{-1} w skali liczb falowych.

progowa gęstość mocy

Progowa gęstość mocy promieniowania laserowego niezbędna do wystąpienia wymuszonego rozpraszania Brillouina zależy przede wszystkim od własności sprężystych ośrodka rozpraszającego. W ciałach stałych typowa wartość progowej gęstości mocy światła wynosi ok. 10^{11} W/m². Taki poziom mocy zapewniają lasery pracujące z modulacją dobroci rezonatora, czyli wytwarzające impulsy „gigantyczne". W cieczach, dzięki występującemu w nich zjawisku samoogniskowania światła, gęstość mocy w kanale autokolimacyjnym znacznie przewyższa progową wartość mocy wymaganą, by wystąpiło wymuszone rozpraszanie Brillouina. Dlatego też doświadczalnie wyznaczony próg rozpraszania Brillouina pokrywa się z progiem samoogniskowania.

Naszkicowany mechanizm wymuszonego rozpraszania Brillouina prowadzi do prostego związku między względnym przesunięciem linii widmowych $(\omega_1-\omega_2)/\omega_1$, kątem rozproszenia i prędkością fali ultradźwiękowej generowanej w ośrodku rozpraszającym światło. Prędkość dźwięku, wyznaczona z pomiarów rozpraszania Brillouina, pozostaje w znakomitej zgodności z wartością wyznaczoną metodami bezpośrednimi. Jest to dodatkowym argumentem przemawiającym za słusznością modelu Brillouina.

Z analizy warunków wzmocnienia fali akustycznej w ośrodku rozpraszającym światło wynika, że w niektórych układach molekularnych, przy dostatecznie dużej mocy padającej wiązki świetlnej o częstości ω_2, natężenie światła przechodzącego może spaść do zera na końcu próbki. Oznacza to, że w wyniku rozpraszania Bruillouina ośrodek działa wówczas jak całkowicie odbijające zwierciadło, przy czym odbite promieniowanie ma częstość ω_1 przesuniętą o ω_p względem częstości ω_2 wiązki padającej. Zdolność odbijająca takiego zwierciadła wzrasta wraz ze wzrostem natężenia światła padającego od 0 do 1. Zjawisko to bywa niekiedy wykorzystywane w układach modulujących dobroć rezonatorów laserowych.

zwierciadło molekularne

Optyka nieliniowa jest młodą, dynamicznie się rozwijającą i bardzo już obszerną dziedziną optyki fizycznej. W tym krótkim omówieniu mogliśmy zwrócić uwagę tylko na niektóre zjawiska związane z wymianą energii między polami promieniowania o różnych częstościach. Dokładniejsze omówienie tych pasjonujących problemów można znaleźć w licznych monografiach poświęconych elektronice kwantowej i optyce nieliniowej.

S. A. Achmanow, R. W. Chochłow *Problemy nieliniejnoj optiki*, Moskwa 1964; N. Bloembergen *Nonlinear Optics*, Reading, Mass. 1977; S. Kielich *Podstawy optyki nieliniowej*, cz. 1, 2, Poznań 1972, 1973; *Elektronika kwantowa i optyka nieliniowa*, red. S. Kielich, F. Kaczmarek, A. Graja, Poznań 1975; J. L. Klimontowicz *Lasery i optyka nieliniowa*, Warszawa 1969; J. E. Midwinter, F. Zernike *Applied Nonlinear Optics*, New York 1973; A.H. Piekara *Nowe oblicze optyki*, Warszawa 1976; J. Stankowski, A. Graja *Wstęp do elektroniki kwantowej*, Warszawa 1972; A. Yariv *Quantum Electronics*, New York 1975.

Ultrakrótkie impulsy światła

Adam Kujawski

Od chwili zbudowania w 1960 r. pierwszego lasera nastąpił ogromny rozwój technik laserowych, które od tego czasu stale ulegają istotnym przeobrażeniom i ulepszeniom. Przykładem tego jest wynalezienie metody wytwarzania ultrakrótkich impulsów światła, zbudowanie laserów barwnikowych i innych o łatwo zmiennych długościach fali oraz zbudowanie laserów o dużej mocy. Omówimy tutaj technikę wytwarzania bardzo krótkich impulsów światła oraz sposoby pomiarów czasów ich trwania. Wspomnimy też o ważniejszych realizowanych obecnie zastosowaniach jak i o poszukiwaniu zupełnie nowych możliwości zastosowań.

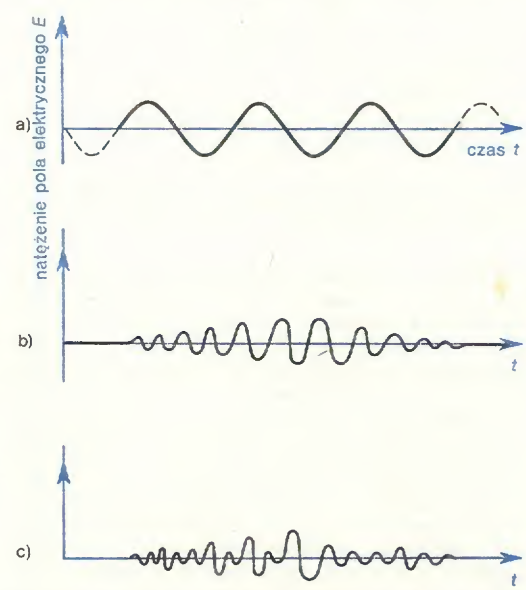
Czasy trwania impulsów, wytwarzanych przez pierwsze lasery rubinowe, były rzędu 10^{-6} s; obecnie dobrze opanowana jest technika wytwarzania impulsów o czasach trwania rzędu 10^{-12} s. Fakt skrócenia czasu o sześć rzędów wielkości świadczy najlepiej o postępie w osiąganiu możliwie najkrótszych impulsów. Podobnie, gęstość mocy ultrakrótkich impulsów obecnie wytwarzanych — mierzona zwykle w W/cm² — jest o wiele rzędów wielkości większa niż pierwszych impulsów laserowych.

gęstość mocy 10^{16} W/cm²

Opanowano już metody otrzymywania impulsów, które po zogniskowaniu mają gęstość mocy dochodzącą do 10^{16} W/cm². Warto sobie uświadomić, jaka jest długość impulsu pikosekundowego, to znaczy impulsu o czasie trwania $\tau = 10^{-12}$ s. Ponieważ prędkość rozchodzenia się impulsu jest równa prędkości światła, którą przyjmujemy $c = 3 \cdot 10^{10}$ cm/s, to długość impulsu $l = c\tau = 0{,}3$ mm. Tak więc impulsy pikosekundowe są „pociskami" świetlnymi o bardzo dużej gęstości mocy promieniowania elektromagnetycznego. Oddziaływanie światła z atomami i cząsteczkami ośrodka, na który pada tego rodzaju „pocisk" świetlny różni się pod wieloma względami istotnie od oddziaływania wiązki świetlnej padającej w sposób ciągły. Jest ono obecnie przedmiotem intensywnych badań w wielu laboratoriach optycznych.

Przypomnijmy, że częstości fal świetlnych, a więc częstości zakresu widzialnego widma elektromagnetycznego są rzędu 10^{15} s^{-1}. Jeśli fala elektromagnetyczna byłaby falą ściśle monochromatyczną opisaną funkcją sinus lub cosinus, to w ustalonym punkcie przestrzeni wykres zależności natężenia pola elektrycznego od czasu miałby postać taką, jak na rys. 1a. Taka fala trwałaby nieskończenie długo. W rzeczywistości fale, które wytwarzamy, mają skończony czas trwania i nie są ściśle monochromatyczne. Rys. 1b ilustruje zależność natężenia pola elektrycznego od czasu dla impulsu świetlnego o skończonym czasie trwania. W impulsie takim pole zmienia się regularnie z częstością z dobrym przybliżeniem równą

fala świetlna



Rys. 1. Zależność wartości natężenia pola elektrycznego od czasu: a) nieskończenie długo trwająca sinusoida, b) impuls o strukturze regularnej, c) impuls o strukturze nieregularnej

impulsy gigantyczne

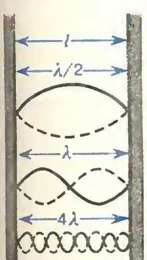
częstości fal świetlnych. Jeżeli częstość tę przyjmiemy za równą 10^{15} s^{-1}, a czas trwania impulsu za 10^{-12} s, to łatwo wywnioskujemy, że impuls zawiera 10^3 drgań. Na rys. 1b, rzecz jasna, zaznaczono znacznie mniejszą liczbę drgań. Należy podkreślić, że nie można dokładnie stwierdzić, czy obecnie wytwarzane impulsy pikosekundowe mają strukturę regularną, taką jak na rys. 1b, czy też drgania są nieuporządkowane i przypadkowe, tak jak schematycznie pokazuje to rys. 1c.

Aby zrozumieć zasadę wytwarzania ultrakrótkich impulsów, musimy zacząć od krótkiego wyjaśnienia zasady działania laserów dających impulsy gigantyczne. Nazwę tę nadano impulsom, których moc była o kilka rzędów wielkości większa od mocy promieniowania zwykłych laserów. Urządzenia takie powstały wkrótce po zbudowaniu pierwszego lasera i fakt ten stanowił istotny postęp w rozwoju elektroniki kwantowej. W laserze, który ma dawać impulsy gigantyczne, akcja laserowa nie może się rozwinąć, gdyż ośrodek czynny lasera jest pompowany (→ Lasery — podstawy działania) przy jednoczesnym popsuciu własności rezonatora optycznego. Jest to tak zwane przełączanie dobroci („dobroć" — wielkość określająca właściwości rezonansowe układu). Popsucie własności rezonatora można uzyskać np. przez umieszczenie komórki Kerra wewnątrz, między zwierciadłami. Najczęściej jest to naczynie z odpowiednią cieczą, którą może być np. bardzo czysty nitrobenzen — jego zdolność przepuszczania światła zależy od pola elektrycznego zewnętrznego, w którym ciecz się znajduje. Gdy komórka nie przepuszcza światła, to oczywiście układ dwu zwierciadeł nie może działać jako rezonator optyczny i ośrodek aktywny może być pompowany bez wywoływania akcji laserowej. Gdy energia zgromadzona w ośrodku czynnym jest dostatecznie duża, działanie rezonatora optycznego przez odpowiednie włączenie komórki zostaje przywrócone do normalnego stanu i następuje szybki rozwój akcji laserowej. Prowadzi to do skrócenia impulsu i powiększenia jego mocy. Metoda ta w różnych wersjach i ulepszeniach jest szeroko wykorzystywana do uzyskiwania impulsów nanosekundowych o mocach rzędu kilkudziesięciu megawatów.

Synchronizacja modów

Do istotnego skrócenia czasu trwania impulsów i powiększenia ich mocy przyczyniła się technika zwana synchronizacją modów (ang. mode-locking). Ze względu na ważność i oryginalność tej metody omówimy dokładniej jej podstawy fizyczne. Każde z drgań własnych pola elektromagnetycznego w rezonatorze optycznym odznacza się określoną częstością. Aby łatwiej zrozumieć, jakie drgania własne można wytwarzać w rezonatorze, rozważmy układ dwóch płaskich, równoległych i idealnie odbijających zwierciadeł, odległość między którymi wynosi l. Zakładamy, że wymiary zwierciadeł są nieskończenie duże. W obszarze między zwierciadłami może istnieć fala stojąca o takiej długości λ_m, że na odcinku o długości l mieścić się będzie całkowita liczba połowy długości fali $\lambda_m/2$, tzn. $m\lambda_m/2 = l$, gdzie m — liczba naturalna, tak jak schematycznie ilustruje to rys. 2. Ten właśnie związek określa drgania własne pola elektromagnetycznego w rezonatorze jednowymiarowym płaskim. Powstanie fali stojącej można oczywiście interpretować jako wynik interferencji dwu fal płaskich poruszających się w kierunkach przeciwnych i odbijających się od zwierciadeł rezonatora. Warto tutaj przypomnieć, że związek $m\lambda_m/2 = l$ jest ogólny i stosuje się również do wielu innych drgań własnych jednowymiarowych o zupełnie różnej naturze fizycznej, np. do drgań powietrza zamkniętego między dwiema ścianami. Ponieważ związek między długością fali λ i częstością ν dla fal elektromagnetycznych w próżni ma postać $c = \lambda\nu$, więc

częstości własne



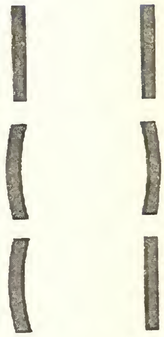
Rys. 2. Fale stojące elektromagnetyczne w rezonatorze utworzonym przez dwa nieskończone zwierciadła

równanie określające częstości drgań własnych przybiera postać

$$\nu_m = mc/2l. \qquad (1)$$

Gdy rezonator jest wypełniony ośrodkiem materialnym o współczynniku załamania n, wówczas c we wzorze (1) winno być zastąpione przez c/n. Rezonator płaski nieskończony, który tutaj omawiamy, jest bezstratny, to znaczy nie występują straty energii fal odbijających się od ścian rezonatora. Tego rodzaju idealizacja oznacza, że pola elektryczne i magnetyczne fali stojącej wytworzonej wewnątrz rezonatora istnieją tylko w obszarze między zwierciadłami i energia pola nie wydostaje się poza ten obszar. Rezonatory wykorzystywane w praktyce mają zwierciadła o wymiarach skończonych; przykłady różnych konfiguracji pokazuje rys. 3. Rezonatory takie nazywa się rezonatorami otwartymi. Drgania pola elektromagnetycznego, które można wytwarzać wewnątrz rezonatorów otwartych, są drganiami tłumionymi. Jest to wynikiem tego, że energia pola elektromagnetycznego opuszcza obszar centralny układu dwu zwierciadeł na skutek takich przyczyn, jak dyfrakcja na krawędzi zwierciadeł, nieidealnie odbijające powierzchnie itd. O tym, jak szybko są tłumione drgania w rezonatorze, decydują: kształt (układ zwierciadeł wklęsłych jest mniej stratny) oraz odległość między zwierciadłami. Nie omawiamy tutaj zupełnie problemu rozkładu linii sił pola elektrycznego i magnetycznego drgań własnych rezonatora optycznego. Ścisła teoria pozwala na ich wyznaczenie oraz wykazuje, że wzór (1), który uzasadniliśmy dla rezonatora nieskończonego, z bardzo dobrym przybliżeniem stosuje się do wielu rezonatorów otwartych używanych w praktyce. Z każdą częstością drgań wyznaczoną przez równanie (1) wiąże się pewien charakterystyczny rozkład przestrzenny linii sił pola elektromagnetycznego w rezonatorze otwartym. Rozkłady te noszą nazwę modów (rodzajów) promieniowania. Mody, które tutaj omawialiśmy (por. rys. 2) i dla których określiliśmy charakterystyczne częstości własne wzorem (1), nazywają się modami podłużnymi. Dotyczą one bowiem drgań i rozkładów pola wzdłuż osi rezonatora. Dla innych kierunków rozkłady pola i charakterystyczne częstości określone są w bardziej złożony sposób. Omawiamy tutaj jedynie mody podłużne, gdyż dokładne zrozumienie, czym one są, jest niezbędne do zrozumienia zasady wytwarzania impulsów pikosekundowych.

rezonator bezstratny



Rys. 3. Przykłady rezonatorów otwartych: a) zwierciadła płaskie, b) zwierciadła kuliste, c) układ zwierciadło kuliste - zwierciadło płaskie

mody promieniowania

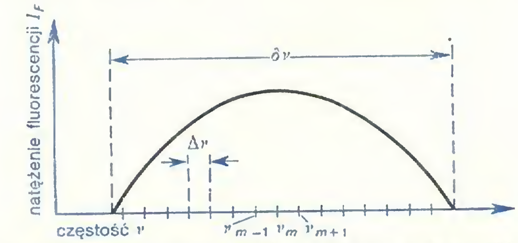
mody podłużne

Ze wzoru (1) wynika, że częstość modu sąsiedniego wynosi $\nu_{m+1} = (m+1)c/2l$ i różnica częstości $\Delta\nu = \nu_{m+1} - \nu_m$ dana jest przez

$$\Delta\nu = c/2l. \qquad (2)$$

Ponieważ $2l/c$ jest czasem przejścia sygnału świetlnego wewnątrz rezonatora od jednego zwierciadła do drugiego i z powrotem, różnica częstości modów sąsiednich równa jest odwrotności tego czasu. Dla rezonatora o długości $l = 1$ m, przyjmując $c = 3\cdot10^8$ m/s, otrzymujemy ze wzoru (2) $\Delta\nu = 150$ MHz. Ośrodek czynny wypełniający rezonator, pobudzany do świecenia przez odpowiednie „pompowanie" optyczne (→ Spektroskopia atomowa), wysyła światło, które odbija się od zwierciadeł rezonatora, co w wyniku interferencji prowadzi do powstania fali stojącej lub, mówiąc inaczej, drgań własnych o częstościach da-

wzbudzanie modów



Rys. 4. W zakresie szerokości widmowej $\delta\nu$ linii fluorescencyjnej mieści się skończona liczba częstości drgań własnych rezonatora

nych wzorem (1). Linia widmowa fluorescencji ośrodka czynnego ma pewną skończoną szerokość $\delta\nu$, tak że wzbudzane są tylko te mody rezonatora, których częstości mieszczą się w szerokości linii. Ilustruje to rys. 4. Dla rubinu szerokość $\delta\nu \approx 330$ GHz i stąd wynika, że w rezonatorze o długości $l = 10$ cm — po uwzględnieniu wartości prędkości światła w rubinie — można wzbudzić około 380 modów. Wspomnijmy od razu, że dla lasera ze szkła neodymowego i dla laserów barwnikowych szerokość linii fluorescencji $\delta\nu$ jest znacznie większa niż dla lasera rubinowego. Oznacza to, że liczba modów wzbudzanych w rezonatorze o takich samych wymiarach jest również odpowiednio większa. Ten fakt ma, jak zaraz zobaczymy, bardzo ważne znaczenie dla charakterystyki światła powstającego w laserze.

Mody podłużne wzbudzane przez świecenie ośrodka czynnego są w ogólności wzbudzane niezależnie od siebie. Oznacza to, że ich fazy drgań nie pozostają w żadnej zależności między sobą i w praktyce przyjmują wartości przypadkowe. Oznaczmy wartości natężenia pola elektrycznego m-tego modu w jakimś określonym punkcie przez E_m. Mamy wówczas $E_m = E_{0m} \sin(2\pi\nu_m t + \varphi_m)$, gdzie E_{0m} jest amplitudą drgań, a φ_m — fazą. Zakładając dla uproszczenia, że wektory pól elektrycznych wszystkich modów wzbudzanych w rozpatrywanym punkcie są tak samo skierowane, otrzymujemy dla wypadkowego natężenia E wyrażenie

nakładanie się modów

$$E = \sum_{m}^{N} E_{0m} \sin(2\pi\nu_m t + \varphi_m). \qquad (3)$$

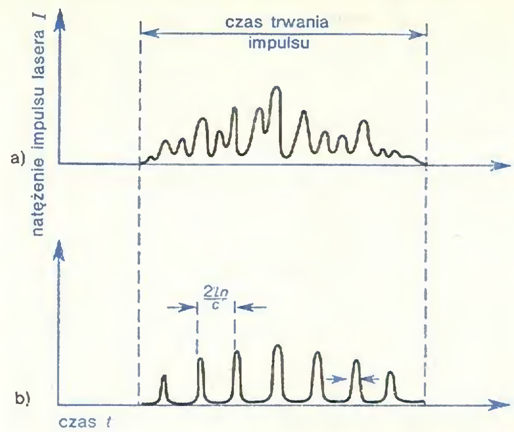
We wzorze (3) sumowanie rozciąga się po wszystkich wartościach m, które odpowiadają częstościom ν_m spełniającym równanie (1) i leżącym w zakresie częstości linii fluorescencyjnej. Nie możemy tutaj szczegółowo wyjaśnić sposobów obliczania E na podstawie wzoru (3) i zależności E od wyboru faz φ_m. Dwa najważniejsze wypadki są następujące: 1) gdy fazy φ_m są różne i mają wartości w ogólności wybrane przypadkowo (mówimy wówczas, że mody nakładają się niespójnie), to wypadkowe pole ma strukturę nieregularną, 2) gdy wszystkie fazy φ_m są sobie równe (spójne nałożenie się modów), wówczas powstałe pole tworzy regularny ciąg miniimpulsów. Zależność od czasu natężenia I, które jest wprost proporcjonalne do E^2, ilustrują odpowiednio rys. 5a i 5b. W pierwszym wypadku powstały impuls ma strukturę chaotyczną, w drugim tworzy się grzebień zawierający impulsy o znacznie krótszym czasie trwania. Dokładna analiza wykazuje, że czas trwania τ pojedynczego impulsu w grzebieniu dany jest przez $\tau \approx 1/\delta\nu$, gdzie $\delta\nu$ jest szerokością linii fluorescencyjnej. Tak więc, aby otrzymać krótkie impulsy, należy posługiwać się laserami, dla których $\delta\nu$ jest możliwie duże. Z tego punktu widzenia laser neodymowy jest lepszy niż rubinowy, gdyż jego linia fluorescencyjna jest szersza, a jeszcze lepsze możliwości dają lasery barwnikowe, dla których $\delta\nu \approx 10^{13}$ Hz. Energia niesiona przez każdy z miniimpulsów regularnego grzebienia, jak na rys. 5b, jest w ogólności większa niż energia pojedynczego miniimpulsu tworzącego się, gdy mody nakładają się niespójnie tak, jak na rys. 5a. W ten sposób nie tylko powstają pojedyncze miniimpulsy, lecz także w sposób kontrolowany uzyskują one dostatecznie dużą energię. Warunek równości faz modów podłużnych, który prowadzi do utworzenia ciągu miniimpulsów, nazywa się warunkiem synchronizacji modów.

warunek synchronizacji modów

realizacja synchronizacji modów

Jak praktycznie zrealizować synchronizację modów? Pierwszy sposób polega na umieszczeniu wewnątrz rezonatora modulatora, który moduluje natężenie światła z częstością równą $\Delta\nu$, określoną wzorem (2). Prowadzi to do ustalania się faz drgań modów w taki sposób, że ich wartości stają się prawie sobie równe i w rezultacie drgania własne nakładają się spójnie. Teoria tego rodzaju modulacji jest dość złożona i jedynie dla pełności opisu wspomnieliśmy o tej metodzie. Drugi sposób, szeroko stosowany w praktyce, polega na wstawieniu do komory rezo-

struktura impulsu

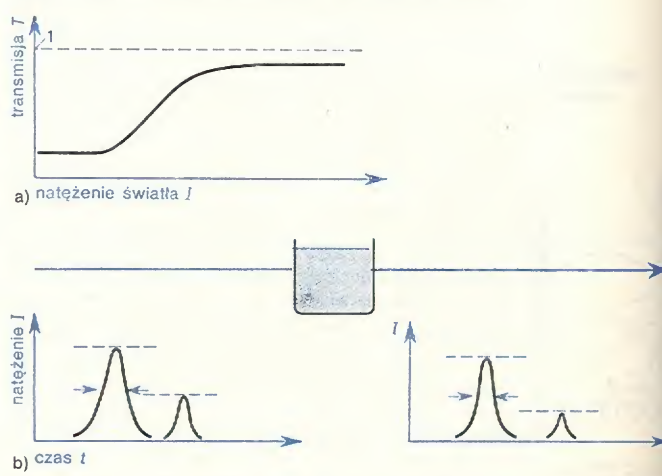


Rys. 5. Struktura impulsu zależy od wyboru wartości faz modów promieniowania: a) struktura chaotyczna, b) regularna

natora odpowiedniego materiału o nieliniowej charakterystyce pochłaniania — rys. 6a. Jest to zwykle barwnik organiczny, którego przepuszczalność (transmisja) zależy od natężenia światła padającego tak, jak ilustruje to rys. 6b. Zakładamy tutaj, że świecenie własne barwnika, to znaczy jego fluorescencja, jak i inne procesy molekularne są bardzo krótkotrwałe. Oznacza to, że ciecz po wzbudzeniu światłem powraca do stanu równowagi niemal natychmiast. Warunek ten jest dobrze spełniony w większości używanych układów wytwarzających impulsy pikosekundowe. Niemniej jednak, stanowi on jedno z ograniczeń w usiłowaniach osiągnięcia jeszcze krótszych, tzn. subpikosekundowych impulsów za pomocą techniki synchronizacji modów. Należy podkreślić, że element nieliniowy nie tylko zmniejsza natężenie, lecz jednocześnie skraca czas trwania impulsu.

działanie elementu nieliniowego

Jak przebiega akcja laserowa w obecności elementu nieliniowego wewnątrz rezonatora? Rozpatrzmy laser pracujący impulsowo. Na początku wzbudzania świecenia lasera impulsem lampy błyskowej wzbudzane mody podłużne mają nieduże natężenia i nakładają się przypadkowo tak, że wytwarzane pole elektromagnetyczne charakteryzuje zależność od czasu zilustrowania na rys. 5a. Promieniowanie wewnątrz rezonatora odbija się wielokrotnie od jego zwierciadeł i przechodząc przez naczynie z barwnikiem ulega

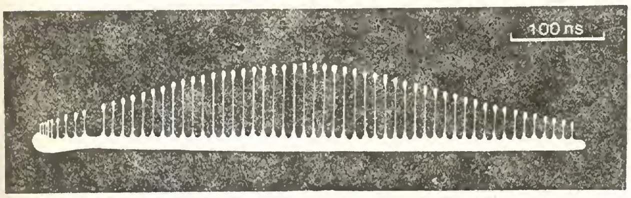


Rys. 6. Transmisja nieliniowa: a) zależność transmisji T (stosunek natężeń światła padającego i przechodzącego) od natężenia światła; b) kształt, jak również wysokość impulsu po przejściu przez ciecz o nieliniowej transmisji są zmienione

częściowemu pochłanianiu. Jednocześnie część promieniowania wychodzi z lasera przez jedno z półprzepuszczających zwierciadeł. W miarę wzrostu natężenia światła impulsu wzbudzającego wytwarzane chaotycznie, miniimpulsy mają większe natężenie i nieliniowy element układu zaczyna pełnić swoją właściwą rolę. Miniimpulsy o małych natężeniach są silniej pochłaniane niż inne o natężeniach większych, a te o dostatecznie dużych natężeniach przechodzą z małymi stratami z jednoczesnym zmniejszeniem czasu trwania (zob. rys. 6b). Tego rodzaju dyskryminacja prowadzi do tego, że tylko impuls najsilniejszy (czasem mogą być dwa o takim samym natężeniu) ulega wzmocnieniu. Promieniowanie wewnątrz lasera jest teraz właściwie pojedynczym impulsem odbijającym się między zwierciadłami i jednocześnie pompowanym. Częściowo opuszcza on rezonator przy odbiciu od zwierciadła półprzepuszczającego i dzieje się to oczywiście z częstością równą odwrotności czasu przejścia od jednego do drugiego zwierciadła i z powrotem. W rezultacie cały duży impuls wychodzący z lasera nie ma chaotycznej struktury, lecz składa się z ciągu miniimpulsów biegnących jeden za drugim w odstępach czasu $2ln/c$. W taki sam sposób działa element nieliniowy w przypadku lasera pracującego w sposób ciągły. Dokładniejsze rozważania wykazują, że działanie elementu nieliniowego opisane powyżej jest równoważne warunkowi równości faz modów nakładających się, który wcześniej omawialiśmy.

dyskryminacja miniimpulsów

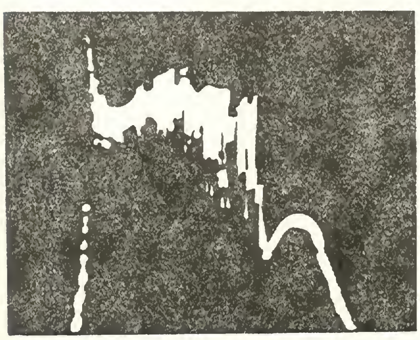
Rysunek 7 jest zdjęciem oscylogramu impulsu z lasera neodymowego, którego czas trwania wynosi ok. 500 ns. Dzięki synchronizacji modów składa się on



Rys. 7. Oscylogram impulsu wytworzonego w warunkach synchronizacji modów (jeden miniimpuls wycięto)

otrzymanie impulsu pikosekundowego

z ciągu miniimpulsów, każdy o czasie trwania kilkunastu pikosekund, biegnących jeden za drugim w odstępach 10 ns, co zgadza się z wartością obliczoną na podstawie wyrażenia $2ln/c$ przy $l = 150$ cm. Pojedynczy miniimpuls może być wycięty z grzebienia odpowiednio szybką migawką. Dalej można go przepuszczać przez ciecz o nieliniowej transmisji (zob. rys. 6), co powoduje dalsze jego skrócenie, ale i zmniejszenie natężenia. Natężenie zwiększa się z kolei wskutek przepuszczenia impulsu przez układ wzmacniający, którym jest napompowany ośrodek



Rys. 8. Oscylogram impulsu (z wieloma rozbłyskami) wytworzonego bez synchronizacji modów

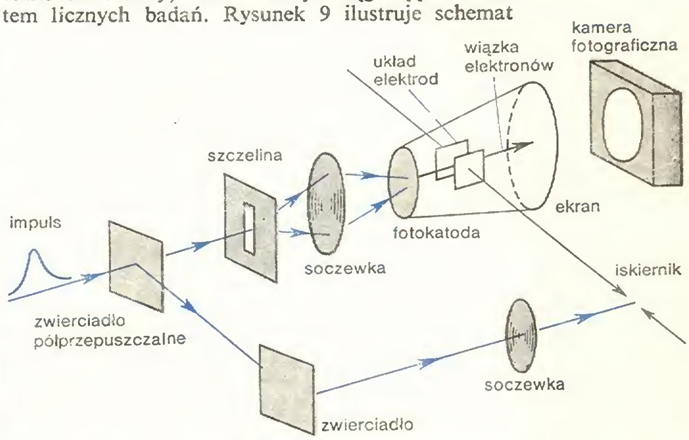
czynny, taki sam jak w laserze. Jeśliby w laserze neodymowym nie było komórki z barwnikiem nieliniowym, to impuls nie miałby regularnej struktury, takiej jak na rys. 7, lecz kształt charakterystyczny dla struktury chaotycznej pokazany na rys. 8.

Wyjaśniając zasady i sposób realizacji synchronizacji modów, nie poruszyliśmy wielu trudnych problemów eksperymentalnych. Mimo ich złożoności uzyskiwanie impulsów pikosekundowych z dobrą powtarzalnością stało się osiągnięciem wielu laboratoriów optycznych na świecie.

Pomiar czasu trwania impulsu

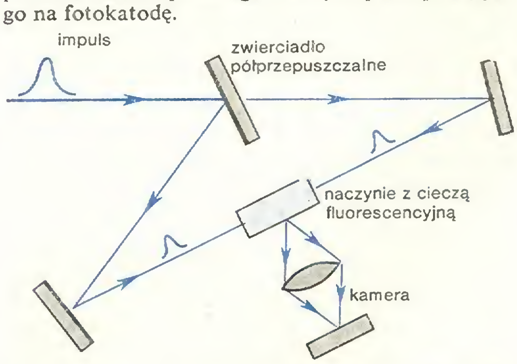
Problemy pomiarów czasów trwania impulsów pikosekundowych pod wieloma względami nie zostały jeszcze całkowicie rozwiązane. Posługiwanie się najszybszymi fotodetektorami i oscylografami pozwala na bezpośrednie pomiary czasów dłuższych niż 100 ps. W ostatnich latach skonstruowano specjalne kamery optyczno-elektroniczne (tzw. *streak cameras*) pozwalające na bezpośredni pomiar czasów rzędu 1 ps. Są to złożone układy, które w dalszym ciągu są przedmiotem licznych badań. Rysunek 9 ilustruje schemat

kamera optyczno-elektroniczna



Rys. 9. Schemat ilustrujący zasadę działania ultraszybkiej kamery optyczno-elektronicznej

działania tego rodzaju kamery. Impuls dzieli się na dwa, z których pierwszy po przejściu przez szczelinę i soczewkę pada na specjalną fotokatodę, a drugi — po zogniskowaniu — służy do włączenia na elektrodach napięcia, które po włączeniu rośnie w zadany sposób. Elektrony wyrzucone z fotokatody są odchylane proporcjonalnie do wielkości napięcia, tak że wybite w różnych czasach padają w różnych miejscach ekranu fosforyzującego. Fotografia ekranu pozwala określić przebieg czasowy impulsu padającego na fotokatodę.

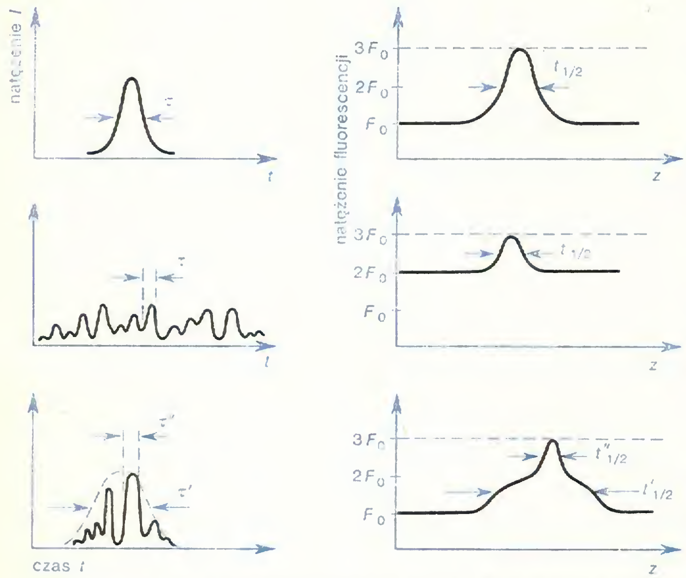


Rys. 10. Schemat ilustrujący zasadę pomiaru czasu trwania ultrakrótkich impulsów

Omówimy oryginalny, obecnie najbardziej powszechny sposób pomiaru, w którym wykorzystuje się nieliniowe zjawisko fluorescencji dwufotonowej. Na

wykorzystanie fluorescencji dwufotonowej.

rys. 10 pokazano schemat typowego układu doświadczalnego. Ultrakrótki impuls dzieli się na dwa o jednakowych natężeniach, które następnie interferują przechodząc przez naczynie zawierające odpowiednią ciecz fluorescencyjną. Najniższy stan wzbudzony molekuł cieczy odpowiada energii wzbudzenia równej energii dwu kwantów światła. Wzbudzenie cieczy do świecenia może więc nastąpić wówczas, gdy natężenie światła jest dostatecznie duże (przejścia wielofotonowe, → Optyka nieliniowa). W obszarze, gdzie się spotykają dwa impulsy, wypadkowe natężenie jest większe i ciecz świeci znacznie silniej. Świecącą ciecz fotografuje się kamerą o dużej zdolności rozdzielcze i na podstawie otrzymanego rozkładu natężenia określa się szerokość impulsu. Dokładne wyznaczenie szerokości nie jest rzeczą łatwą (rys. 11); ocena szero-



Rys. 11. Rozkład natężenia światła fluorescencji wzbudzanej światłem o różnym natężeniu I w funkcji położenia z

kości wiąże się istotnie z pomiarami kontrastu (tzn. stosunku różnicy maksymalnego i minimalnego natężenia świecenia fluorescencji do natężenia minimalnego). Najmniejsza wartość natężenia świecenia zależy od tła (szumu), na jaki nałożony jest impuls, którego szerokość chcemy mierzyć. We wszystkich wypadkach zachodzi $\tau = \gamma t_{1/2}$, tzn. czas trwania impulsu jest wprost proporcjonalny do połówkowej szerokości (mierzonej na połowie wysokości maksymalnego natężenia) linii fluorescencyjnej zarejestrowanej na kliszy fotograficznej. Współczynnik proporcjonalności zależy od kształtu impulsu (zwykle $\gamma = 1{,}5$) i nie może być wyznaczony przy użyciu omawianej techniki.

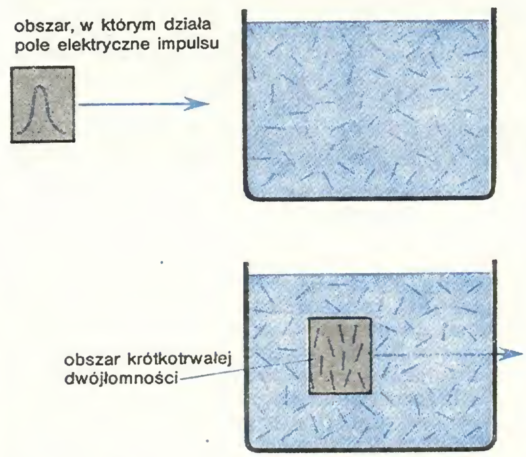
Zastosowanie ultrakrótkich impulsów

Tak jak zmienia się sposoby wytwarzania i detekcji, tak też stale rozszerza się zakres stosowania ultrakrótkich impulsów. Omówimy tutaj jedynie te zastosowania, które mają lub mogą mieć szczególne znaczenie dla badań fizycznych. Przede wszystkim zwrócimy uwagę Czytelnika na fakt, że w ośrodkach materialnych każdego rodzaju, a więc w gazach, cieczach i ciałach stałych zachodzą procesy, których czas trwania lub częstotliwości ich występowania są rzędu kilku do kilkudziesięciu pikosekund. Atomy i cząsteczki pozostają bowiem w ciągłym ruchu, którego charakter i rodzaj zależy od ich budowy oraz warunków fizycznych, w jakich się znajdują. W przypadku gazów i cieczy zderzenia atomów i cząsteczek mogą prowadzić do takich odkształceń ich powłok elektronowych, że te atomy i cząsteczki pozostają we wzbudzonych stanach o wyższej energii i mogą z kolei przekazywać innym energię wzbudzenia. W ciałach stałych ruch drgający atomów lub cząsteczek tworzących sieć krystaliczną może być bardzo złożony i niezwykle szybki. Śledzenie i badanie tego rodzaju zjawisk wymaga odpowiedniego zegara, którego skala pozwoli mierzyć odstępy czasu rzędu pikosekund. Ultrakrótkie impulsy o czasach trwania rzędu nanosekund pozwoliły zbadać przebieg wielu zjawisk o takich właśnie czasach trwania. Teraz jednak w centrum zainteresowań są przebiegi zjawisk o pikosekundowych skalach czasu. Przykładem jest świecenie fluorescencyjne o bardzo krótkim czasie trwania, np. materiał pobudzony do świecenia może świecić przez okres 10^{-8} s i krócej. Dotychczasowe, istniejące od wielu lat odpowiednio szybkie migawki i układy elektroniczne pozwalały na badanie świecenia o czasach trwania nie krótszych niż 10^{-9} s.

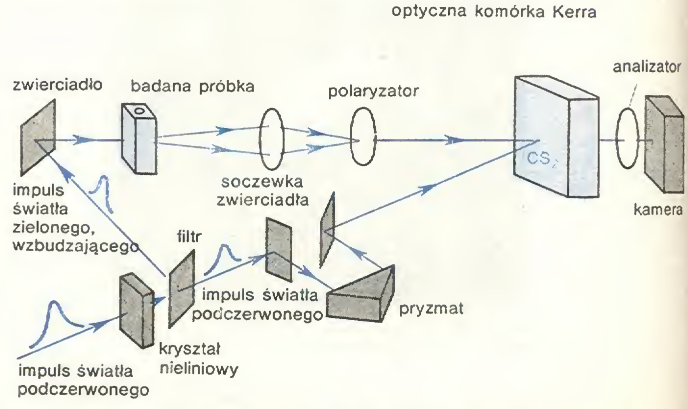
zjawiska pikosekundowe

dwójłomność wymuszona przez falę świetlną

Pole elektryczne impulsu laserowego rozchodzącego się w cieczy może tworzyć w cieczy krótkotrwałą dwójłomność. Mówiąc inaczej, ciecz nie mająca własności kryształu anizotropowego staje się dwójłomna w obszarze, przez który przechodzi fala świetlna, i może polaryzować światło podobnie jak kryształ. Na rys. 12 schematycznie pokazano, jak molekuły cieczy reorientują się w polu fali świetlnej, co zwykle prowadzi do wzrostu współczynnika załamania w obszarze fali. Ponieważ czas powrotu do stanu równowagi (czas relaksacji) dla wielu cieczy jest rzędu kilku pikosekund, okres istnienia obszaru dwójłomnego, wy-



Rys. 12. Pole elektryczne impulsu orientuje molekuły tak, że wytwarza się krótkotrwała dwójłomność



Rys. 13. Schemat wykorzystania ultrakrótkiej migawki do badania świecenia fluorescencji

tworzonego impulsem pikosekundowym, mierzy się czasem kilku pikosekund. Naczynie z cieczą, w której dwójłomność jest wymuszona przez silnie pole elektryczne impulsu świetlnego, nazywa się optyczną komórką Kerra. Została ona wykorzystana przy konstrukcji ultraszybkiej migawki, której czas przepuszczania światła wynosi około 8 ps. Tak szybkie zadziałanie nie byłoby możliwe przy użyciu zwykłej komórki Kerra.

optyczna komórka Kerra

ultraszybka migawka

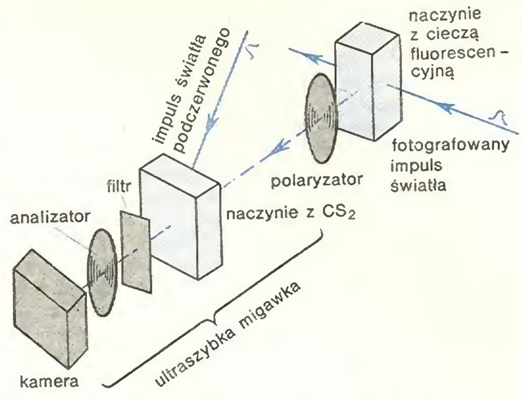
Ultraszybkie migawki wykorzystuje się np. do badań krótkich czasów świecenia fluorescencyjnego. W tym doświadczeniu (rys. 13) ultrakrótki impuls światła podczerwonego o długości fali 10,6 μm, pochodzący z lasera neodymowego, pada na kryształ nieliniowy (np. KDP), który daje drugą harmoniczną, tzn. po przejściu impulsu przez kryształ pojawia się dodatkowo impuls światła zielonego o długości fali 5,3 μm. Odpowiednie zwierciadło z filtrem (opuszczone na rysunku) przepuszcza impuls światła podczerwonego, który padając na naczynie z dwusiarczkiem węgla (optyczna komórka Kerra) wymusza dwójłomność cieczy na czas 8 ps. Impuls światła zielonego wzbudza światło fluorescencyjne, które pada na układ — polaryzator, optyczna komórka Kerra, analizator. Polaryzator i analizator są ustawione prostopadle, tak że gdy ciecz nie jest dwójłomna, światło nie przechodzi do fotodetektora. Gdy impuls promieniowania podczerwonego wymusi na krótki czas dwójłomność, kierunek liniowej polaryzacji świecenia fluorescencyjnego ulega zmianie i światło przechodzi przez analizator do fotodetektora. Zmieniając drogę optyczną impulsu podczerwonego (przez przesuwanie pryzmatu) można otwierać migawkę na 8 ps wcześniej lub później, co pozwala badać różne odcinki czasów świecenia fluorescencji trwających kilkadziesiąt pikosekund i określać krzywą spadku natężenia. W ten sposób zbadano przebieg świecenia różnych związków organicznych i dzięki temu określono podstawowe właściwości ruchów oscylacyjno-rotacyjnych cząsteczek.

Tego rodzaju badania mają duże znaczenie dla wielu zjawisk fotochemicznych i biofizycznych, w których światło inicjuje bardzo szybkie procesy. Z podobnych względów stosuje się także wzbudzenia pikosekundowe do badania drgań atomów sieci krystalicznej ciała stałego. We wszystkich tego rodzaju eksperymentach nowe możliwości przyniosły lasery barwnikowe, które działając w odpowiednim układzie z synchronizowanymi modami mogą być źródłem impulsów pikosekundowych o różnych długościach fali. Przy tym mogą one być przestrajane na różne długości fal w sposób ciągły. Taka technika jest znacznie efektywniejsza (a nawet czasami niezastąpiona) od otrzymywania fal o różnych długościach za pośrednictwem zjawisk nieliniowych, jak to już omawialiśmy na przykładzie wytwarzania światła o barwie zielonej. W rezultacie wzbudzanie impulsami pikosekundowymi (również laserem pracującym w sposób ciągły) może być bardzo selektywne. Oznacza to, że w układach atomowych i cząsteczkowych można badać przejścia między jakimiś dwoma dowolnie wybranymi poziomami energetycznymi naświetlając światłem, którego fotony mają energię dokładnie równą różnicy energii tych poziomów. Stwarza to zupełnie nowe możliwości i wytycza kierunek rozwoju współczesnej spektroskopii optycznej.

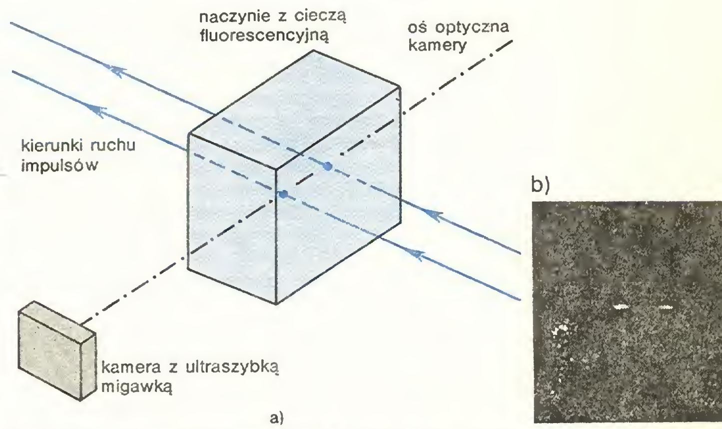
wykorzystanie laserów barwnikowych

Ultraszybką migawkę, wyżej omówioną, zastosowano również w oryginalnym doświadczeniu, w którym sfotografowano impuls światła. Rysunek 14 ukazuje schemat doświadczenia. Rysunek 15 ilustruje wersję eksperymentu, w której w naczyniu z cieczą fluorescencyjną poruszają się prostopadle do linii celowania obiektywu dwa impulsy pikosekundowe, jeden obok drugiego. Gdyby migawka została otwarta na okres dłuższy niż 8 ps, zdjęcie pokazałoby ciągłą linię świecenia cieczy. Otwarcie na krótki czas pozwala jednak uchwycić impuls w locie. Aby stało się to w odpowiednim momencie, zarówno impulsy fotografowane, jak i impuls otwierający migawkę pochodzący z tego samego lasera, są przez odpowiedni układ optyczny przesuwane w czasie względem siebie. W pierwszej chwili oglądania zdjęcia zadziwia fakt,

fotografia impulsu światła



Rys. 14. Schemat układu do fotografowania ultrakrótkiego impulsu



Rys. 15. Dwa ultrakrótkie impulsy tworzą obiekt relatywistyczny; a) oba impulsy znajdują się na osi optycznej układu; b) otrzymane zdjęcie

że klisza rejestruje dwa impulsy, mimo że wzdłuż linii celowania „widoczny" powinien być tylko jeden, a drugi — zasłonięty. Związane to jest z tym, że prędkość światła jest skończona i do obiektywu oraz na kliszę fotograficzną dochodzi jednocześnie światło rozproszone w cieczy przez impuls pierwszy i drugi. Mówiąc inaczej, krótkotrwałe otwarcie migawki wpuszcza fotony pochodzące od impulsu pierwszego (w tym momencie nie jest już on na osi optycznej obiektywu!) oraz fotony pochodzące od impulsu drugiego wysłane wcześniej, gdy nie był on jeszcze na osi optycznej. Rejestrowane na kliszy fotony pochodzące od pierwszego i drugiego impulsu docierają do niej jednocześnie i określają różne położenia źródeł, z których zostały wysłane. Jest to prosty i znakomicie ilustrujący przykład, że obserwacja (fotografia) obiektów poruszających się z prędkościami relatywistycznymi, tzn. z prędkościami bliskimi lub równymi prędkości światła, daje obrazy, których kształt można wyjaśnić tylko z uwzględnieniem skończonej prędkości światła. W omawianym przykładzie obiektem takim jest układ dwóch impulsów. Poruszony tutaj problem został opracowany i wyjaśniony z teoretycznego punktu widzenia dużo wcześniej.

Innym ważnym zjawiskiem, obok dwójłomności cieczy, które może być wywołane silnym polem elektrycznym impulsu, jest jego samoogniskowanie. Zmiana współczynnika załamania w obszarze impulsu powoduje, że impuls rozchodzi się w ośrodku niejednorodnym, co zmienia jego kształt tak, jak uczyniłaby to soczewka skupiająca. Jest to w ogólności złożony

samoogniskowanie impulsu

problem, dotyczący również poprzecznego kształtu impulsu, którego tutaj zupełnie nie poruszamy.

Powiemy teraz krótko o możliwości zastosowania ultrakrótkich impulsów do celów mikrosyntezy jądrowej (→ Energia termojądrowa). Badania te podjęte w wielu krajach — również i w Polsce — mają na celu wykorzystanie dużej gęstości energii impulsu laserowego do wytworzenia stanu materii o bardzo wysokiej gęstości i temperaturze, w której możliwa jest reakcja syntezy jąder lekkich. Fakt, że pojawiło się tego rodzaju nowe podejście do problemu reakcji termojądrowych i że jest ono przedmiotem badań jako jedno z możliwych praktycznych rozwiązań, wynika z postępu w wytwarzaniu i pomiarach ultrakrótkich impulsów laserowych. Zasadnicza idea jednego z takich rozwiązań polega na tym, że kropla mieszaniny deuteru i trytu zostaje z wielu stron równomiernie oświetlona wysokoenergetycznymi impulsami laserowymi. Odrzucenie zewnętrznej części kulki (szybkie wyparowanie) prowadzi do ściśnięcia reszty i wytworzenia dużej gęstości, co powinno istotnie poprawić wydajność reakcji termojądrowej. Z punktu widzenia techniki ultrakrótkich impulsów nie jest najważniejszą sprawą ich bardzo krótki czas trwania, lecz odpowiedni profil, tzn. specjalny dobór krzywej wzrostu i zaniku natężenia w czasie. Również niezbędna jest duża energia impulsu. Pierwsze impulsy nanosekundowe niosły energię rzędu 0,1 J, obecnie takie impulsy mogą nieść energię 10000 J, a prowadzi się prace nad dalszym jej powiększeniem. W gorącej plazmie, w której zachodzą reakcje termojądrowe, ma się do czynienia z wieloma bardzo złożonymi procesami. Informacje o nich można otrzymać śledząc szybko ich przebieg. Tu również znajdują zastosowanie techniki, w których wykorzystuje się impulsy pikosekundowe.

mikrosynteza jądrowa

Kończąc omawianie zastosowania ultrakrótkich impulsów, wspomnijmy jeszcze o możliwości wykorzystania ich do pompowania układów atomowych, które byłyby źródłem promieniowania spójnego w zakresie dalekiego nadfioletu i rentgenowskim. Problem ten jest przedmiotem bardzo intensywnych badań w wielu ośrodkach naukowych. Innym przykładem poszukiwania zastosowań jest możliwość wykorzystania impulsu pikosekundowego w przełączniku optoelektrycznym. Urządzenie to może włączyć układ elektroniczny na czas kilku pikosekund, jak też służyć tylko do ultraszybkiego włączania lub wyłączania.

pompowanie w zakresie nadfioletu i rentgenowskim

Liczne i różnorodne zastosowania laserów bardzo często wykorzystują impulsową pracę lasera, a więc krótkie impulsy światła. Ograniczyliśmy się do tych aktualnych problemów i osiągnięć dotyczących impulsów światła, których dalszy rozwój z pewnością odegra ważną rolę w fizyce molekularnej, fizyce plazmy i fizyce ciała stałego. Szczególną uwagę poświęciliśmy impulsom pikosekundowym. Wiadomo już, że techniką synchronizacji modów uzyskano impulsy o czasach trwania krótszych niż 1 ps. Ten zakres nie jest jednak w pełni opanowany. Z uwagą więc śledzi się prace teoretyczne i doświadczalne przewidujące uzyskiwanie impulsów o czasach krótszych niż 1 ps, z wykorzystaniem wymuszonych zjawisk rozpraszania światła. Technika ta jest dopiero w początkowym stadium badań. Bez wątpienia opanowanie i wykorzystanie światła w formie krótkich i ultrakrótkich impulsów należy do jednej z najciekawszych i aktualnie rozwijanych technik laserowych i ma duże znaczenie dla różnych badań naukowych.

R. R. Alfano, S. L. Shapiro *Ultrafast Phenomena in Liquids and Solids*, Scientific Amer., 228, No. 6, 42 (1973); M.A. Duguay *Light Photographed in flight*, Am. Scientist 59, 551 (1971); A. Piekara *Nowe oblicze optyki*, Warszawa 1976; S. Kaliski *Laserowa kompresja i synteza termojądrowa plazmy*, Delta nr 3, 1975; H. Klejman *Lasery*, Warszawa 1979.

Holografia

Romuald Pawluczyk

Oko ludzkie oraz wszystkie przyrządy służące do zapisu obrazów świetlnych mogą rejestrować tylko przedmioty składające się z takich ośrodków materialnych, które zaburzają padające na nie fale świetlne (lub same je emitują) w sposób wyróżniający je z otoczenia. Zaburzone lub wysyłane przez przedmiot promieniowanie rozchodzi się w określonych kierunkach, ma określony skład widmowy, a składowe fale mają określoną polaryzację, amplitudę i fazę. Ogólnie mówiąc — stan fizyczny światła pochodzącego od poszczególnych fragmentów przedmiotu jest różny, wskutek czego w przestrzeni powstają różnice w rozkładzie natężenia i składu widmowego (barwy) światła.

przedmiot i jego obraz

Jeżeli ten rozkład zostanie odwzorowany za pomocą układu optycznego, to otrzymany obraz może być utrwalony na powierzchni odpowiedniego materiału światłoczułego (fotografia) lub odebrany przez odpowiedni układ odbiorczy (nerw oczny w oku, układ elektroniczny w kamerze telewizyjnej itp.). W każdym wypadku odbiornik uzyskuje jednak tylko tę informację o przedmiocie, która się przejawia w rozkładzie natężenia (odbiorniki typu czarno-białego lub jasnociemnego) oraz barwy (odbiorniki do odbioru barwnego), natomiast zostaje zatracona informacja związana z zaburzeniem fazy oraz stanu polaryzacji fali świetlnej. Utrata informacji o fazie ma szczególnie istotne znaczenie, gdyż z tym wiąże się utrata informacji o sposobie zaburzenia fazy fali świetlnej przez przedmiot, a co za tym idzie — zatraceniu ulega trójwymiarowość ze wszystkimi jej atrybutami: głębią, perspektywą i paralaksą. Zapisany w ten sposób obraz nie jest pod względem optycznym równoważny z przedmiotem, gdyż fale świetlne ukształtowane przez obraz na ogół całkowicie różnią się pod względem rozkładu fazy od fal świetlnych ukształtowanych przez przedmiot (wynika to ściśle z dyfrakcyjnej teorii odwzorowania optycznego).

Aby więc odtworzony obraz był optycznie równoważny z rejestrowanym przedmiotem, oprócz rozkładu amplitudy (natężenia) i barwy powinien być zarejestrowany, a następnie odtworzony rozkład fazy ukształtowanych przez przedmiot fal świetlnych. Warunek ten spełniają metody holograficzne (grecki wyraz holographéō znaczy pisać w całości, nie skracając).

holographéō

Fizyczne podstawy holografii

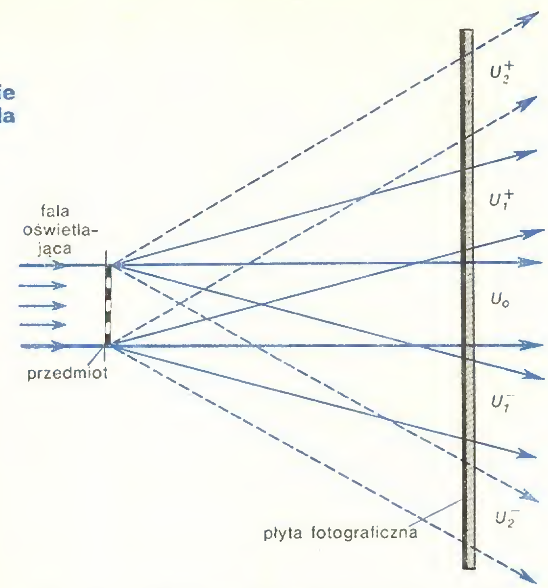
Dyfrakcyjna teoria odwzorowania optycznego

Powstanie holografii jest konsekwencją rozwoju dyfrakcyjnej teorii odwzorowania optycznego, stworzonej w zeszłym stuleciu przez E. Abbego w celu wyjaśnienia zasad działania mikroskopu. Teoria ta okazała się bardziej uniwersalna i tłumaczy również działanie wszystkich innych układów, w których następuje odwzorowanie przedmiotu za pomocą pól falowych (np. fal dźwiękowych, radiowych, rentgenowskich czy fal przedstawiających rozchodzenie się cząsteczek materii — elektronów, protonów itp.).

Rozpatrzmy z punktu widzenia tej teorii najprostszy przypadek odwzorowania optycznego — odwzorowanie płaskiego przedmiotu zaburzającego

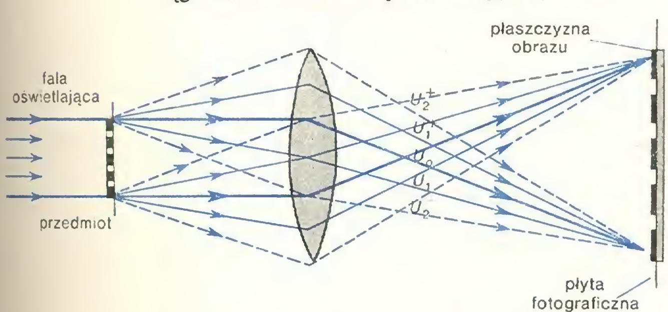
tylko amplitudę. Światło oświetlające przedmiot (rys. 1) zachowuje się tak, jak gdyby przedmiot składał się z dużej liczby nałożonych na siebie siatek dyfrakcyjnych amplitudowych. Dla każdego przedmiotu liczba składowych siatek, kierunek prążków w danej siatce, stała siatki oraz maksymalny i minimalny stopień zaburzenia amplitudy padającej fali (głębokość modulacji) są w jednoznaczny sposób ustalone.

ugięcie światła



Rys. 1. Ugięcie światła na przedmiocie. Fala oświetlająca oddziałuje z przedmiotem w taki sposób, jak gdyby składał się on z siatek dyfrakcyjnych o różnych częstościach, z których każda ugina część światła w dwóch kierunkach (U_1^+ i U_1^-; U_2^+ i U_2^-)

Na każdej siatce część padającej fali świetlnej ulega ugięciu w dwóch kierunkach, położonych symetrycznie względem padającej fali (uwzględniamy tylko ugięcie I rzędu). Kierunki i kąty ugięcia ściśle zależą od orientacji prążków i stałej siatki, a amplituda fal ugiętych — od głębokości modulacji. Tak więc każda siatka składowa przedmiotu jest reprezentowana przez parę fal o określonej amplitudzie, ugiętych w określonych kierunkach. W miarę oddalania się od przedmiotu rozkład natężenia światła wytwarzany przez te fale coraz bardziej się różni od rozkładu natężenia na powierzchni przedmiotu. W płaszczyźnie obserwacji (leżącej w pewnej odległości od przedmiotu) rozkład natężenia może być tak różny od rozkładu natężenia na powierzchni przedmiotu, że na jego podstawie nie można rozpoznać przedmiotu (można się o tym przekonać, wstawiając bezpośrednio w to miejsce płytę fotograficzną, która po naświetleniu nie da obrazu). Niemniej jednak, pole świetlne w dalszym ciągu zawiera całkowitą informację o przedmiocie,



Rys. 2. Otrzymywanie obrazu w układach optycznych. Po przejściu przez optyczny układ skupiający (soczewkę) ugięte na przedmiocie fale świetlne ponownie nakładają się na siebie w płaszczyźnie obrazowej i interferując odtwarzają obrazy siatek składowych przedmiotu czyli dają jego obraz

która jest zakodowana w postaci par fal o określonych amplitudach, padających na płaszczyznę obserwacji z określonych kierunków pod określonymi kątami. Informacja o kierunkach i kątach padania poszczególnych fal zostanie odwzorowana na tej płaszczyźnie jako odpowiedni rozkład fazy (znaczy to, że utrata informacji o rozkładzie fazy w świetle ugiętym jest równoważna z utratą informacji o okresach i kierunkach siatek tworzących przedmiot).

Aby ze światła zaburzonego przez przedmiot można było odzyskać informację o tym przedmiocie (np. obraz przedmiotu), rozkład amplitudy i fazy należy przetransformować na rozkład natężenia światła taki, jak w płaszczyźnie przedmiotu. Zadanie to spełniają różnorodne układy optyczne (soczewki aparatu fotograficznego, soczewka w oku, a w optyce elektronowej — układy wytwarzające odpowiednią konfigurację pól elektrycznych i magnetycznych — soczewki magnetyczne; rys. 2). Układy takie odtwarzają wiernie tylko obrazy przedmiotów płaskich, natomiast deformują przestrzennie obrazy przedmiotów trójwymiarowych. Dostępne fotodetektory (np. płyty fotograficzne, siatkówka oka itp.) nie mogą wiernie zarejestrować nawet tych zdeformowanych obrazów. Wiernie rejestrują one tylko te fragmenty obrazu, które leżą dokładnie na powierzchni fotodetektorów, pozostałe zaś fragmenty są rejestrowane z tym większą utratą informacji, im dalej się znajdują od powierzchni fotodetektora.

dwustopniowość odwzorowania optycznego

Z powyższych rozważań wynika, że obrazy powstają w dwóch etapach: najpierw światło dokonuje analizy przedmiotu i koduje informację o jego własnościach optycznych w postaci amplitudy i fazy zaburzonego przez przedmiot pola świetlnego, tzw. pola dyfrakcyjnego (matematycznie zjawisko to opisuje się za pomocą tzw. transformacji Fouriera), a następnie układ optyczny transformuje to pole w rozkład natężenia (dokonuje syntezy — matematycznie odpowiada to odwrotnej transformacji Fouriera), który może już być zarejestrowany przez dostępne fotodetektory (→ Optyka fourierowska).

Gdyby istniały sposoby zapisu oraz odtwarzania rozkładu amplitudy i fazy pola dyfrakcyjnego, czyli gdyby można było oddzielić proces syntezy obrazu od procesu analizy, to przy zachowaniu pewnych warunków stosowane w obu tych procesach promieniowanie mogłoby się różnić swymi własnościami. Po raz pierwszy ideę tę wysunął w 1920 r. polski fizyk Mieczysław Wolfke, który pragnął ją wykorzystać w mikroskopii rentgenowskiej i zaproponował stosowanie promieniowania rentgenowskiego w procesie analizy, a światła widzialnego w procesie syntezy. Idea ta była później jeszcze kilkakrotnie podejmowana, jednakże nie znalazła urzeczywistnienia, ponieważ nie było metody zapisu rozkładu fazy i amplitudy pola dyfrakcyjnego. Problem udało się rozwiązać dopiero twórcy holografii, D. Gaborowi, w 1948 r.; do zapisu rozkładu fazy i amplitudy zaproponował on zastosowanie metody interferencyjnej.

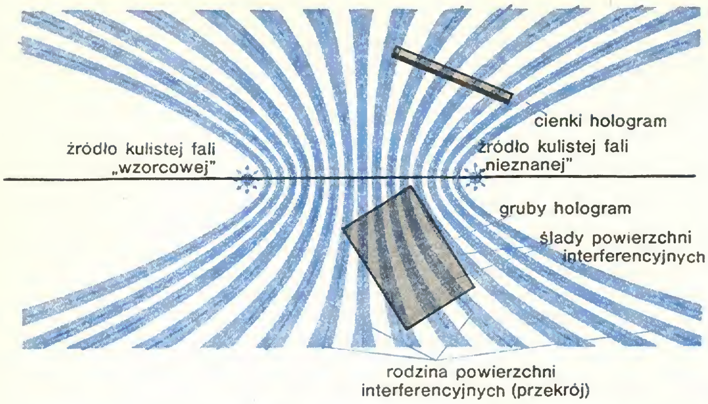
metoda interferencyjna

Istota metody interferencyjnej polega na tym, że fala o nieznanym rozkładzie fazowym jest nakładana na spójną z nią falę wzorcową (rys. 3 → Spójność światła). W wyniku interferencji obu fal tworzy się przestrzenny układ maksimów i minimów natężenia światła (pole interferencyjne). Na płaszczyźnie przecinającej pole interferencyjne te minima i maksima tworzą określoną strukturę. W wypadku światła monochromatycznego w cienkiej płycie fotograficznej jest ona rejestrowana jako układ naprzemiennych jasnych i ciemnych prążków interferencyjnych, a w grubym ośrodku światłoczułym — jako układ jasnych i ciemnych powierzchni. Dwa sąsiednie prążki (powierzchnie) o tym samym natężeniu są zbiorami punktów, w których różnica faz interferujących promieni różni się o 2π (co w swobodnej przestrzeni odpowiada zmianom różnicy dróg optycznych interferujących promieni o jedną długość

fala odniesienia

fali użytego światła). Jeżeli kształt powierzchni falowej jednej z tych fal (fali odniesienia) jest znany, to

na podstawie obrazu prążków (powierzchni) można wyznaczyć kształt powierzchni falowej drugiej fali — fali badanej.



Rys. 3. Przekrój powierzchni interferencyjnych tworzących się w wyniku interferencji światła spójnego emitowanego przez dwa źródła punktowe

D. Gabor udowodnił, że jeżeli do pola interferencyjnego wprowadzi się płaską płytę fotograficzną i zarejestruje na niej prążki interferencyjne, a taki obraz interferencyjny (hologram) oświetli się taką samą falą jak fala odniesienia, to w wyniku ugięcia na tym obrazie powstanie fala będąca jak gdyby dalszym ciągiem fali badanej. Odkrycie tego zjawiska dało początek holografii.

hologram

Obrazy holograficzne

fala przedmiotowa

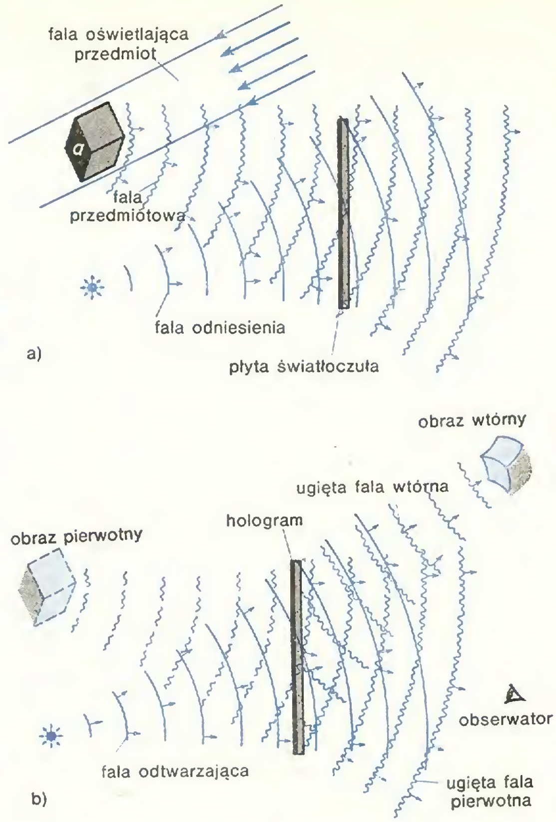
Dalsze badania wykazały, że do wiernego odtworzenia fali badanej (w holografii nazywanej falą przedmiotową, gdyż zazwyczaj jest ona w jakiś sposób ukształtowana przez przedmiot) fala wzorcowa, zwana w holografii falą odniesienia, nie musi mieć powierzchni falowej o znanym kształcie, jak również obojętny jest kąt pomiędzy kierunkami padania interferujących fal (rys. 4a). Obie fale (przedmiotowa i odniesienia) interferują między sobą, tworząc przestrzenny układ maksimów i minimów pola interferencyjnego, zarejestrowanych w hologramie w postaci prążków (w cienkim ośrodku światłoczułym) lub powierzchni (w grubym).

fala odtwarzająca

Aby z hologramu uzyskać wierne odtworzenie zarejestrowanej fali przedmiotowej, fala odtwarzająca musi być wierną „kopią" fali odniesienia, a więc musi padać na hologram z tej samej strony, z tego samego kierunku i pod tym samym kątem co fala odniesienia w czasie zapisu (rys. 4b). Światło fali odtwarzającej ulega częściowemu ugięciu na siatce tworzącej hologram. (Jeżeli grubość materiału światłoczułego jest odpowiednio mała, porównywalna z odstępem między prążkami, to jej wpływ na ugięcie światła można pominąć). Ugięcie następuje w dwóch różnych kierunkach i za hologramem oprócz wiązki nieugiętej powstają dwie fale ugięte. Jedna z nich jest jak gdyby dalszym ciągiem fali przedmiotowej i obserwator ma wrażenie, że ogląda rzeczywisty przedmiot przez okienko o wymiarach hologramu. Rozcinając hologram na mniejsze części, nie ogranicza się wielkości obrazu zwanego pierwotnym, lecz tylko zmniejsza się możliwość zmian punktu obserwacji (podobnie jak przy zmniejszaniu wymiarów okienka). Zmniejszanie wymiarów hologramu dopóty nie wpływa na jakość obrazu, dopóki są one większe od wymiarów źrenicy oka. Przy dalszym zmniejszaniu jakość obrazu zaczyna ulegać pogorszeniu, następuje rozmycie konturów, a przy wymiarach porównywalnych z długością fali światła użytego w procesie analizy — obraz całkowicie znika.

obraz pierwotny

Druga fala ugięta też zawiera informację o przedmiocie, jednakże w bardzo złożonej postaci i, na ogół biorąc, informacji tej nie można wykorzystać. Sytuacja



Rys. 4. Zapisywanie hologramu (a) i odtwarzanie w świetle monochromatycznym (b). Na hologramie zapisuje się obraz interferencyjny utworzony w wyniku interferencji fali przedmiotowej z falą odniesienia. Do odtwarzania stosuje się falę identyczną z falą odniesienia. Jedna z fal ugiętych na hologramie jest „przedłużeniem" fali przedmiotowej, dzięki czemu odtwarza obraz przedmiotu. Wtórna fala ugięta na ogół nie daje wiernego obrazu (nieoświetlona ściana *a* sześcianu nie jest rejestrowana na hologramie)

się zmienia, gdy fale odniesienia i odtwarzająca są falami płaskimi lub sferycznymi. Jeżeli fala odniesienia i fala odtwarzająca są falami płaskimi, padającymi prostopadle na powierzchnię hologramu, to druga fala ugięta tworzy również wierny obraz przedmiotu, zwany wtórnym. Fala ta (w odróżnieniu od fali pierwszej, która się zachowuje tak, jak gdyby się rozchodziła od obrazu) zbiega się do obrazu, a jej kierunek ugięcia jest przeciwny do kierunku ugięcia fali pierwotnej, przy czym w tym wypadku oba te obrazy są zwierciadlanym odbiciem jeden drugiego względem płaszczyzny hologramu. Obserwator oglądający wtórny obraz widzi go zawieszonym w przestrzeni przed hologramem, przy czym ten obraz ma odwróconą głębię, jest więc obrazem pseudoskopowym. Źródłem światła tworzącym oba obrazy jest hologram, do oka obserwatora mogą więc docierać tylko promienie rozchodzące się w nieskończonym ostrosłupie, którego jeden z przekrojów stanowi hologram, a wierzchołek — oko. Z tego względu obserwator będzie widział tylko te fragmenty odtworzonych obrazów, które się mieszczą w objętości tak wyznaczonego ostrosłupa. Do obejrzenia innych fragmentów obrazu zarejestrowanego na hologramie należy zmienić punkt obserwacji.

obraz wtórny

Jeżeli po wykonaniu hologramu nie został on powiększony lub pomniejszony, a fala odtwarzająca jest identyczna z falą odniesienia, to pierwotny obraz jest wierną kopią przedmiotu, natomiast obraz wtórny

na ogół jest zniekształcony. Po przeskalowaniu hologramu (powiększeniu lub pomniejszeniu), po zmianach kształtu powierzchni falowej lub też długości fali odtwarzającej zniekształceniu ulegają oba obrazy. Odtworzone obrazy mogą być powiększone lub pomniejszone, również położenie odtworzonych obrazów w przestrzeni może być inne niż przedmiotu.

powiększenie lub pomniejszenie obrazu

Technicznie najprostszym sposobem uzyskania powiększonych (pomniejszonych) obrazów jest zmiana krzywizny powierzchni falowej fali odtwarzającej w stosunku do fali odniesienia lub zmiana odległości źródła fali odtwarzającej od hologramu. Jeżeli po wykonaniu hologram został powiększony (dla hologramu pomniejszonego m jest ułamkiem), długość fali odtwarzającej jest λ_c, a długość fali odniesienia i przedmiotowej — λ_o (wiązki odniesienia i odtwarzania są falami kulistymi), to słuszne są następujące relacje między współrzędnymi obrazów (wskaźniki l) i przedmiotu (wskaźniki p) (rys. 5):

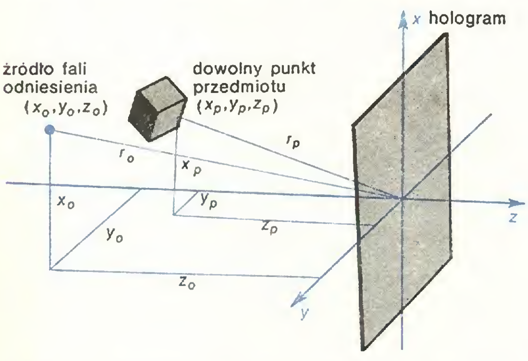
$$z_l = \frac{z_p}{\dfrac{z_p}{z_c} - \dfrac{k\lambda_c}{m^2\lambda_o}\left(\dfrac{z_p}{z_o} - 1\right)},$$

$$x_l = \left(\frac{k\lambda_c x_p}{m\lambda_o z_p} - \frac{k\lambda_c x_o}{m\lambda_o z_o} + \frac{x_c}{z_c}\right) z_l.$$

Dla współrzędnej y wzór jest taki sam jak dla współrzędnej x, należy tylko w miejsce x wstawić y.
Gdy $x \ll z$ i $y \ll z$ powiększenie obrazów M_l wynosi:

$$M_l = \frac{m}{1 - \dfrac{z_p}{z_o} + \dfrac{m^2\lambda_o}{k\lambda_c}\,\dfrac{z_p}{z_o}};$$

x_o, y_o, z_o — współrzędne źródła fali odniesienia, x_c, y_c, z_c — współrzędne źródła fali odtwarzającej. Wskaźnik l i współczynnik k przybierają wartość 1 dla obrazu pierwotnego oraz −1 dla obrazu wtórnego.

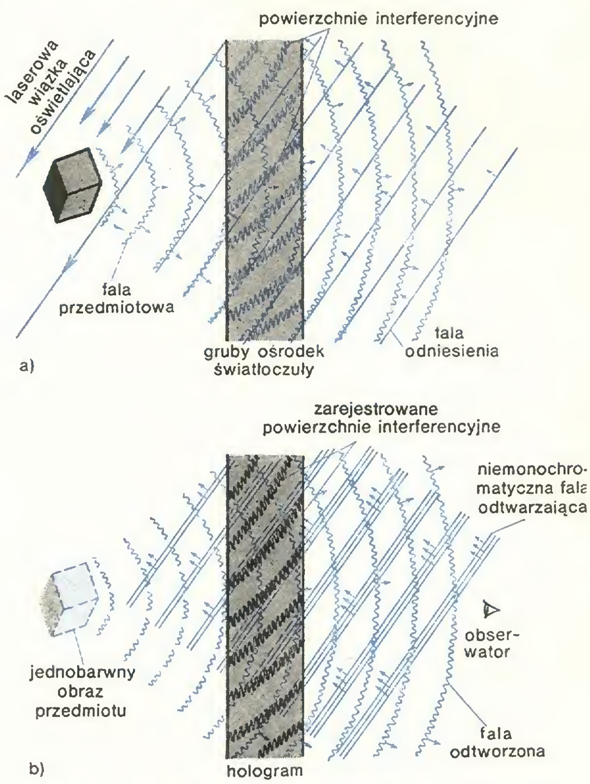


Rys. 5. Wybór układu współrzędnych do opisu współzależności między położeniem poszczególnych elementów układu holograficznego; przy odtwarzaniu położenie środka układu współrzędnych oraz orientacja układu nie ulegają zmianom

Obrazy holograficzne otrzymywane w świetle niemonochromatycznym

Z przedstawionych wyżej wzorów wynika, że położenie odtworzonych obrazów oraz ich powiększenie zależą od długości fali światła użytego do odtwarzania. Bardziej szczegółowe rozważania prowadzą do wniosku, że obrazy odpowiadające różnym długościom fal będą poprzesuwane względem siebie, wskutek czego przy odtwarzaniu hologramów światłem niemonochromatycznym nastąpi tęczowe rozmycie odtworzonych obrazów. Rozpoznanie takich obrazów jest trudne, a przeważnie wręcz niemożliwe. Jeżeli natomiast grubość stosowanego materiału światłoczułego będzie dużo większa od odstępu między prążkami rejestrowanymi na płaskim hologramie, to przestrzenny układ maksimów i minimów pola interferencyjnego zostanie zapisany w postaci przestrzennych naprzemianległych jasnych i ciemnych powierzchni interferencyjnych. Wskutek właściwości takiej przestrzennej struktury interferencyjnej hologram jest w stanie odtworzyć obraz tylko w świetle o takiej długości fali, jaka była użyta przy zapisie, i to pod warunkiem, że fala odtwarzająca rozchodzi się z tego samego kierunku co fala odniesienia. Przy

grubość materiału światłoczułego



Rys. 6. Wykonywanie i odtwarzanie hologramów rekonstruowanych w świetle niemonochromatycznym: a) hologram wykonuje się przy użyciu wysokospójnego światła laserowego, b) odtworzenie następuje w świetle niemonochromatycznym

oświetleniu falą niemonochromatyczną taki hologram sam wybiera z wiązki świetlnej światło o właściwej barwie, dzięki temu można uzyskać dobrej jakości obrazy w białym świetle punktowej żarówki lub nawet w świetle słonecznym.

obrazy w świetle białym

Barwa i ostrość odtworzonego obrazu tym bliższe są oryginalnym, im grubszy ośrodek jest stosowany do zapisu hologramu. Sposób wykonywania takich hologramów został zaproponowany po raz pierwszy w 1962 r. przez J. Denisiuka.

W celu zapisania możliwie dużej liczby powierzchni falowych w metodzie tej fala przedmiotowa i fala odniesienia padają na ośrodek światłoczuły z przeciwnych stron (rys. 6a). Przy odtwarzaniu obrazów (rys. 6b) na zarejestrowanych powierzchniach interferencyjnych następuje wsteczne ugięcie światła, a fale ugięte na poszczególnych powierzchniach interferują między sobą, tworząc jednobarwny obraz.

Otrzymywanie hologramów

Na hologramach rejestrowanych w cienkich ośrodkach światłoczułych odległość między sąsiednimi prążkami zależy od kąta zawartego między kierunkami padania interferujących fal. W trójwymiarowych hologramach nieprzezroczystych przedmiotów

ta odległość waha się od ułamka do pojedynczych mikrometrów, a w hologramach przedmiotów przezroczystych może być nieco większa (kilka mikrometrów). Aby taki obraz interferencyjny można było zarejestrować, materiały światłoczułe muszą mieć wysoką zdolność rozdzielczą — minimum 2000 linii na mm (zwykłe materiały fotograficzne mają ok. 100 linii na mm). Po naświetleniu i wywołaniu materiału światłoczułego obraz interferencyjny zapisany w postaci złożonej siatki prążków interferencyjnych można zobaczyć w dużym powiększeniu (il. 143, tabl. 37).

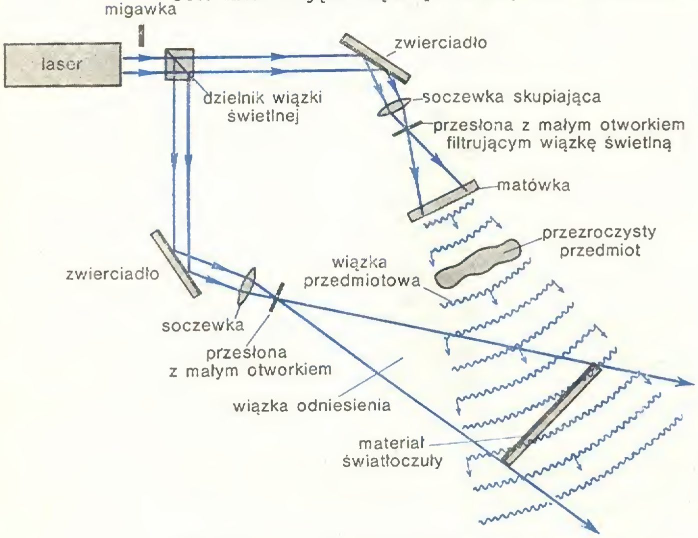
holografia w świetle laserowym

Przy wykonywaniu hologramów obraz interferencyjny musi być stabilny przez cały czas naświetlania. Kontrast obrazu zależy od spójności stosowanego promieniowania (→ Spójność światła) oraz od stanu polaryzacji interferujących fal. Uzyskanie dostatecznie spójnego światła o odpowiedniej intensywności z termicznych źródeł praktycznie jest niemożliwe, dlatego to obecny rozwój holografii nastąpił dopiero po wynalezieniu laserów (→ Lasery — podstawy działania); za prawdziwą datę narodzin holografii można uznać lata 1962–1964, kiedy to E. N. Leith oraz J. Upatnieks wspólnie opublikowali serię prac dotyczących holografii w świetle laserowym.

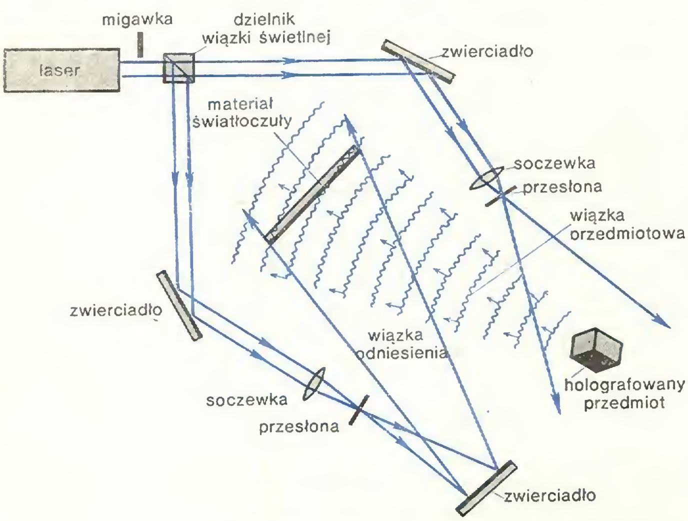
Nawet z zastosowaniem laserów stabilny obraz interferencyjny można otrzymać tylko wtedy, gdy obie interferujące wiązki pochodzą z jednego źródła, a różnica dróg optycznych interferujących promieni (mierzona od miejsca, gdzie wiązka została podzielona, do miejsca ich interferencji na powierzchni materiału światłoczułego) nie przekracza długości spójności. Warunki te spełniają układy, których schematy pokazano na rys. 7 i 8.

lasery stosowane w holografii

Ze względu na dobre parametry użytkowe i stosunkowo niską cenę w laboratoriach holograficznych najczęściej są używane lasery helowo-neonowe (He-Ne). Ponadto stosuje się lasery argonowe, a także impulsowe lasery rubinowe lub z kryształami YAG w układzie zwiększającym spójność wytwarzanego przez nie promieniowania (w wypadku lasera YAG promieniowanie podczerwone jest zmieniane metodą podwajania częstości na promieniowanie widzialne; → Optyka nieliniowa). Dla uzyskania dobrej jakości hologramów układ holograficzny należy zabezpieczyć przed drganiami oraz przed wpływem czynników zewnętrznych (zmiany temperatury, ciśnienia i wilgotności powietrza). Toteż badania holograficzne często prowadzi się w przystosowanych do tego celu pomieszczeniach. Elementy układu holograficznego umieszcza się na masywnych stołach w sposób zapewniający dostateczną sztywność całego układu (il. 140, tabl. 36). Jeśli przedmiot holografowany jest ze swej natury niestabilny (np. obiekty żywe), stosuje się lasery inpulsowe o odpowiednio krótkim impulsie i dużej energii.



Rys. 7. Schemat układu do holografowania przedmiotów przezroczystych



Rys. 8. Schemat układu do holografowania przedmiotów nieprzezroczystych

Zastosowanie holografii

Rejestrując pełną informację amplitudowo-fazową pola świetlnego, holografia znacznie wierniej odzwierciedla rzeczywistość niż dotychczasowe metody zapisu obrazów. Wskutek tego możliwości jej zastosowań są również szersze niż metod tradycyjnych, co już obecnie można zauważyć w nauce i technice. Odtwarzanie trójwymiarowości jest tak ważną cechą holografii, że niewątpliwie otworzy jej drogę do zastosowania również w życiu codziennym.

Interferometria holograficzna

Holograficzna metoda utrwalania fazy fali świetlnej nie tylko spowodowała rozwój dotychczasowych metod badań interferencyjnych, ale i przyczyniła się do powstania nowych.

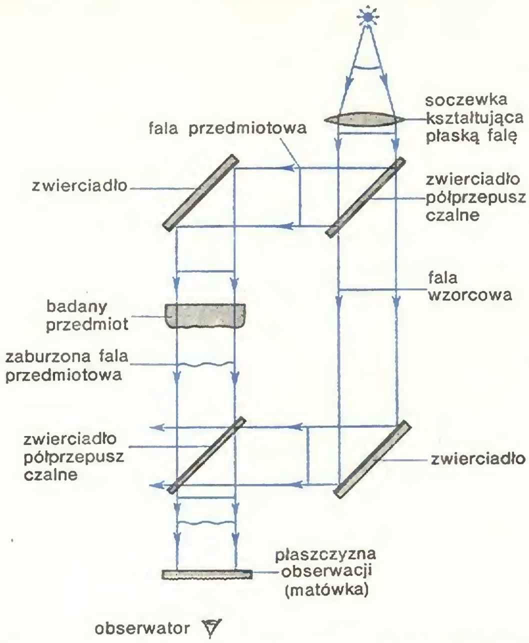
interferometria tradycyjna

Zjawisko interferencji znajdowało praktyczne zastosowanie już przed wynalezieniem holografii, służyło do precyzyjnych pomiarów odległości, kształtu powierzchni płaskich lub sferycznych, jednorodności materiałów optycznych i wielu innych. Istota tradycyjnych pomiarów interferencyjnych polega na tym, że w odpowiednim układzie optycznym światło emitowane przez spójne źródło dzieli się na dwie fale o dokładnie znanym, zazwyczaj jednakowym kształcie powierzchni falowej. Jedna z tych fal, zwana falą przedmiotową, oddziałuje z badanym obiektem, np. niejednorodną płytą szklaną (której jakość należy sprawdzić), wskutek czego ulega zaburzeniom. Fala ta nakłada się następnie z powrotem na drugą (niezaburzoną) falę — falę wzorcową (rys. 9), obie interferują, a w płaszczyźnie obserwacji tworzą obraz interferencyjny, na podstawie którego można wyciągnąć wnioski o badanym obiekcie.

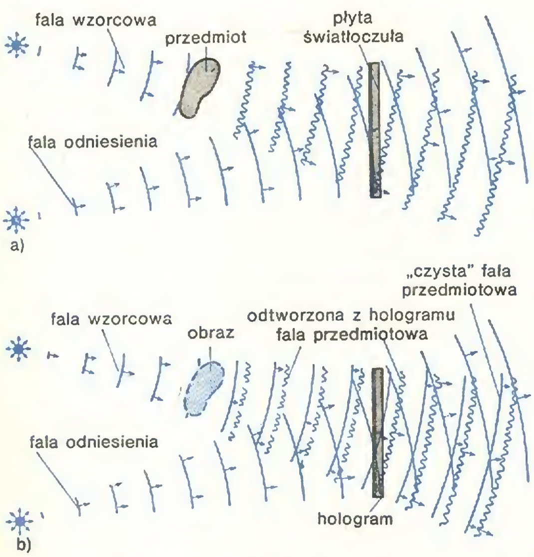
Ponieważ w zwykłych interferometrach każdą z fal kształtują inne elementy układu optycznego, to — aby uzyskać łatwe w interpretacji obrazy interferencyjne — stosowane elementy optyczne muszą być wykonane z bardzo wysoką precyzją, co znacznie podnosi koszty budowy interferometrów.

Przed wynalezieniem holografii można było prowadzić badania interferometryczne tylko przedmiotów umieszczonych w układzie optycznym, co było

szczególnie niedogodne przy zjawiskach zmiennych i krótkotrwałych. Holograficzne utrwalenie takich zjawisk pozwala na prowadzenie badań interfero-



Rys. 9. Schemat działania dwuwiązkowego holograficznego interferometru do badania jednorodności przezroczystych płyt



Rys. 10. Holograficzne badania interferencyjne: a) wykonanie hologramu obiektu badanego, b) uzyskanie interferogramu holograficznego — odtworzona z hologramu fala badana nakłada się na niezaburzoną przez obiekt falę przedmiotową

metrycznych w warunkach i w czasie dogodnym dla eksperymentatora, przy czym do badań można wykorzystać dowolny układ holograficzny, umożliwiający holografowanie badanego obiektu. Aby przeprowadzić badanie interferometryczne, należy po wykonaniu hologramu (rys. 10a) usunąć z układu badany obiekt i ustawić gotowy hologram dokładnie w miejscu, w którym został wykonany. Po oświetleniu hologramu falą wzorcową (wolną od przedmiotu falą przedmiotową) i falą odniesienia za hologramem w kierunku rozchodzenia się fali wzorcowej będą się jednocześnie rozchodziły dwie fale świetlne: fala wzorcowa oraz odtworzona z hologramu fala zawierająca informację

fala wzorcowa i fala badana

o badanym obiekcie — fala badana (rys. 10b). Interferując między sobą, wytworzą one obraz interferencyjny jak w normalnym interferometrze. Ponieważ obie interferujące fale kształtuje ten sam układ optyczny, to z obrazu interferencyjnego można wnioskować o własnościach przedmiotu (obraz ten nie zależy od kształtu powierzchni falowej fali wzorcowej). Dzięki temu w interferometrach holograficznych można znacznie obniżyć wymagania co do jakości używanych elementów optycznych, a jedynym istotnym warunkiem jest dokładne ustawienie hologramu w miejscu jego wykonania.

obiekty o własnościach stałych

Przy badaniu obiektów o własnościach optycznych stałych lub zmieniających się powoli wygodniej jest zapisać na hologramie falę wzorcową, a przedmiot wprowadzać do układu na okres badań. Taki układ jest ścisłym analogiem zwykłego interferometru dwuwiązkowego.

Ponieważ w omówionych holograficznych układach interferencyjnych kształt fali wzorcowej nie odgrywa istotnej roli, zatem na drodze tej fali może się znajdować dowolny element optyczny z matówką włącznie — ważne jest jedynie to, aby się znajdował w tym samym miejscu zarówno w czasie zapisu hologramu, jak i w czasie badań interferencyjnych. Dzięki temu możliwe jest prowadzenie badań różnych procesów zachodzących w obszarach ograniczonych przezroczystymi nieregularnymi ściankami lub ściankami o optycznie niskiej jakości (np. procesów cieplnych w żarówkach czy w wysokociśnieniowych palnikach rtęciowych, procesów spalania paliwa w specjalnie przygotowanych komorach).

obiekty o własnościach zmiennych

Badanie obiektów o zmieniających się własnościach polega na porównywaniu zapisanego na hologramie obrazu obiektu w początkowym stanie fizycznym z jego obrazem w stanie zmienionym przez przyłożenie określonego układu sił, zmianę ciśnienia wewnątrz lub na zewnątrz badanego obiektu, zmianę jego temperatury, wprowadzenie odkształceń relaksacyjnych, zmianę wilgotności itp.

interferometria w czasie rzeczywistym

Istnieją dwie metody takich badań. W pierwszej, tzw. interferometrii w czasie rzeczywistym, hologram badanego obiektu, znajdującego się w określonym stanie fizycznym, umieszcza się w miejscu, w którym został wykonany, i przedmiot pozostawia się w miejscu, w którym był holografowany. Między falą odtworzoną z hologramu i falą ukształtowaną przez przedmiot występuje interferencja i na podstawie uzyskanego obrazu interferencyjnego można uzyskać informację o zmianach własności optycznych wewnątrz obiektu (w badaniach w świetle przechodzącym) lub zmianach jego powierzchni (zob. il. 146, tabl. 39).

interferometria dwuekspozycyjna

Dwa stany tego samego obiektu w dwóch różnych chwilach można także porównać dokonując dwukrotnego naświetlania płyty światłoczułej w tym samym układzie (interferometria dwuekspozycyjna). Jeżeli się taki hologram oświetli falą odtwarzającą, zostaną odtworzone dwie fale świetlne, odpowiadające dwóm różnym stanom obiektu. Fale te utworzą obraz interferencyjny umożliwiający wykrycie powstałych w obiekcie zmian. Gdy mamy do czynienia z szybkozmiennymi zjawiskami, do uzyskania dwuekspozycyjnego interferogramu stosuje się lasery impulsowe dużej mocy, umożliwiające uzyskanie dwóch silnych błysków laserowych w krótkim odstępie czasu.

badanie materiałów

Obie metody służą do porównania dwóch stanów tego samego obiektu, przy czym za ich pomocą można wykryć zmiany zachodzące nawet w wiązkach ukształtowanych przez silnie rozpraszający obiekt, czego nie można byłoby w ogóle dokonać metodami zwykłej interferometrii. Metody te, jako metody nie niszczące i umożliwiające wykrycie minimalnych zmian położenia lub kształtu obiektów rozpraszających znalazły zastosowanie w różnorodnych badaniach własności mechanicznych materiałów i elementów. Metoda dwuekspozycyjna, szczególnie z zastosowaniem laserów dwuimpulsowych, jest użyteczna również w warunkach przemysłowych. Metody interferometrii ho-

lograficznej są szeroko stosowane w przemyśle rakietowym, lotniczym, samochodowym, energetyce itd. Jako przykład można podać zastosowanie holografii do wykrywania wad w oponach. W tym celu oponę, układ holograficzny i hologram umieszcza się wewnątrz komory podciśnieniowej (il. 141, tabl. 37). Na hologramie są rejestrowane dwa stany powierzchni opony przy dwóch różnych ciśnieniach otaczającego powietrza. Miejsca wadliwe przejawiają się w postaci zaburzenia struktury powstających prążków interferencyjnych. Ponadto interferometria dwuekspozycyjna z zastosowaniem laserów dwuimpulsowych (il. 142, tabl. 37) służy do badań deformacji i drgań elementów maszyn i urządzeń w czasie ich pracy (badania drgań poszycia samolotów i śmigłowców, karoserii i silników samochodowych, części obrabiarek itd.).

badanie drgań

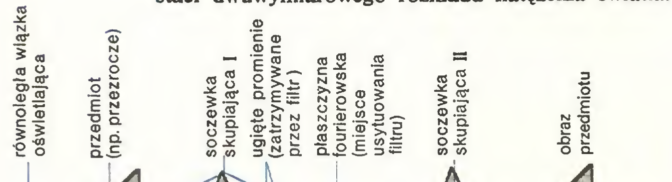
Do badania stacjonarnych drgań o niewielkiej amplitudzie stosuje się metodę badań z uśrednianiem obrazu w czasie rejestracji, zwaną metodą uśredniania czasowego. Polega ona na tym, że naświetlanie hologramu trwa dłużej niż okres drgań. Wówczas na hologramie rejestrują się kolejne obrazy przedmiotu w każdej fazie ruchu. Przy odtwarzaniu interferujące fale tworzą obraz uśredniony, na którym węzły drgań przejawiają się w postaci jasnych obszarów otoczonych prążkami o kontraście zanikającym w miarę wzrostu aplitudy drgań (il. 144, tabl. 38). Do badania drgań o większej amplitudzie (powyżej kilku mikrometrów) stosuje się impulsową metodę dwuekspozycyjną lub holograficzną metodę stroboskopową. Ta druga polega na tym, że przedmiot i hologram są oświetlane za pomocą krótkotrwałych błysków świetlnych padających w chwilach, gdy się obiekt znajduje w dwóch ściśle określonych fazach drgań.

zastosowanie w mikroskopii

Interferometria holograficzna znajduje również zastosowanie w mikroskopii, służy także do wyznaczania kształtu przedmiotów oraz do zwiększania czułości metod interferencyjnych (il. 145, tabl. 38).

Optyczne przetwarzanie i przechowywanie informacji

Różnica pomiędzy układami elektronicznymi i optycznymi sprowadza się do tego, że w układach elektronicznych informacja jest przekazywana i przetwarzana w postaci sekwencji impulsów, natomiast w układach optycznych jest ona przedstawiana w postaci dwuwymiarowego rozkładu natężenia światła.



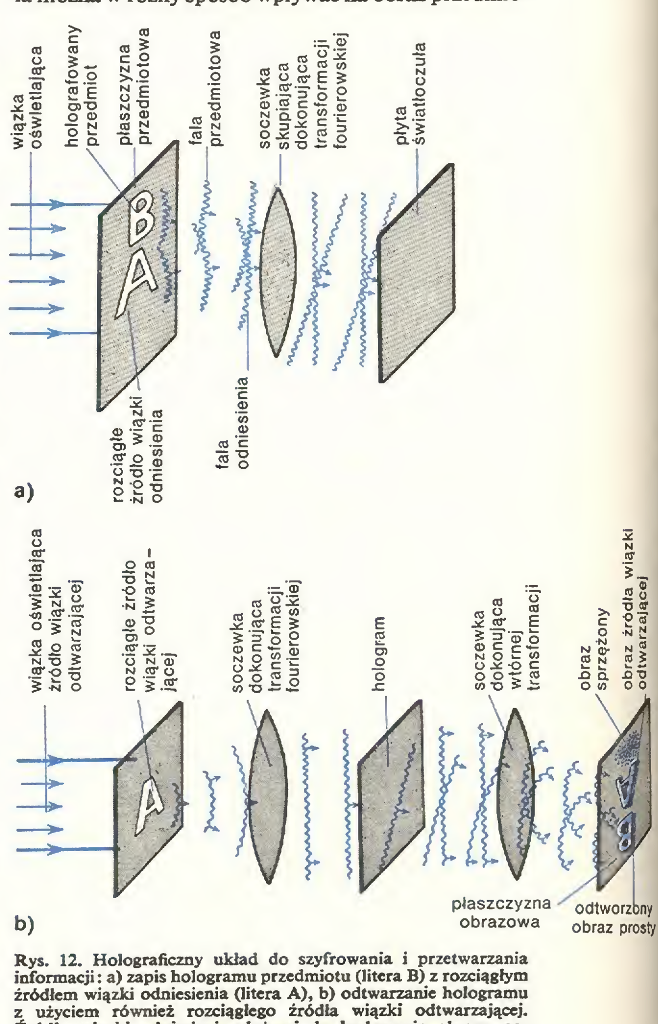
Rys. 11. Schemat układu optycznego do analizy fourierowskiej i filtracji częstości przestrzennych. Soczewka *I* dokonuje transformacji fourierowskiej przezrocza i w płaszczyźnie fourierowskiej tworzy jego transformatę, którą soczewka *II* przetwarza na obraz obiektu. Wstawiając w płaszczyźnie fourierowskiej odpowiednie filtry amplitudowo-fazowe można modyfikować obraz przedmiotu tworzony przez soczewkę *II*

Dzięki tej właśnie różnicy układ optyczny może jednocześnie przekazywać i przetwarzać dużą ilość informacji, co gwarantuje dużą szybkość działania.

Specyfika pracy układów optycznych czyni je szczególnie przydatnymi do realizacji działań typu mnożenia funkcji przez siebie oraz niektórych przekształceń całkowych. Jako przykład można podać, że zwykła soczewka skupiająca wykonuje w sposób natychmiastowy dość złożoną operację z dziedziny wyższej matematyki, a mianowicie przekształcenie Fouriera dowolnej funkcji dwóch zmiennych, reprezentowanej jako przepuszczalność przezrocza (przezrocze osłabia amplitudę padającej fali świetlnej w sposób proporcjonalny do wartości analizowanej funkcji w danym punkcie). Jeżeli się takie przezrocze ustawi w przedniej płaszczyźnie ogniskowej soczewki skupiającej, to rozkład amplitudy i fazy światła w tylnej płaszczyźnie ogniskowej tej soczewki będzie przedstawiał transformatę Fouriera analizowanej funkcji. Ustawiając za tą płaszczyzną drugą soczewkę tak, aby jej przednia płaszczyzna ogniskowa pokrywała się z tylną płaszczyzną ogniskową pierwszej soczewki, w tylnej płaszczyźnie ogniskowej drugiej soczewki uzyska się obraz przezrocza (rys. 11).

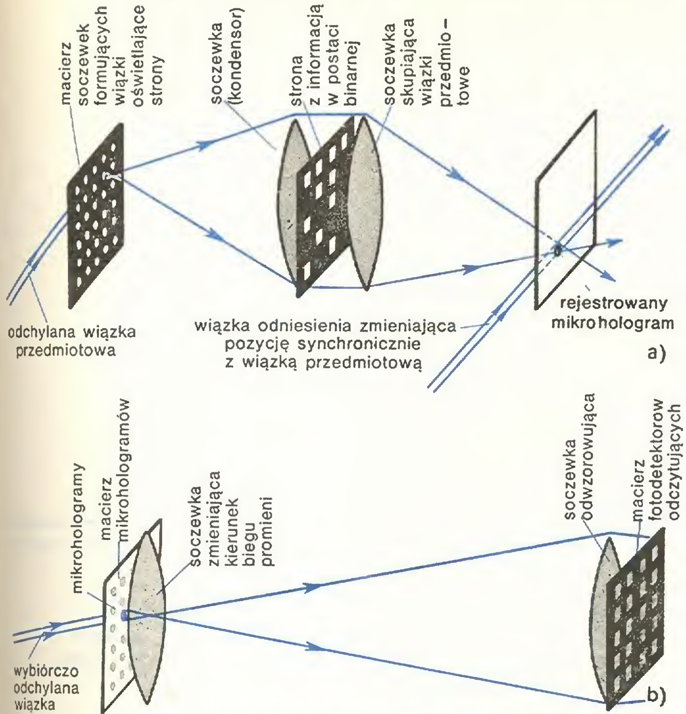
przekształcenie Fouriera w układzie optycznym

Przez wstawienie między soczewkami w ich wspólnej płaszczyźnie ogniskowej elementu optycznego zaburzającego amplitudę i fazę przechodzącego światła można w różny sposób wpływać na obraz przedmio-



Rys. 12. Holograficzny układ do szyfrowania i przetwarzania informacji: a) zapis hologramu przedmiotu (litera B) z rozciągłym źródłem wiązki odniesienia (litera A), b) odtwarzanie hologramu z użyciem również rozciągłego źródła wiązki odtwarzającej. Źródło wiązki odniesienia służącej do kodowania obrazu powinno mieć możliwie duże wymiary i możliwie złożoną strukturę amplitudowo-fazową; natomiast do rozpoznawania obrazów używa się jako przedmiotu — punktowego źródła światła, a wówczas w miejscu prostego obrazu tworzy się obraz będący funkcją korelacji między rozciągłymi źródłami wiązki odniesienia i wiązki odtwarzającej

tu (np. polepszyć jego ostrość, usunąć fragment obrazu; rys. 11). Elementy służące do tego celu nazywają się filtrami częstości przestrzennych. Wykonanie



Rys. 13. Pamięć holograficzna: a) zapisywanie mikrohologramu, b) odtwarzanie i odczytywanie informacji

filtry częstości przestrzennych

amplitudowo-fazowego filtru zwykłymi metodami jest bardzo pracochłonne. W niektórych wypadkach takie filtry można łatwo wykonać metodami holograficznymi. Taką metodę zastosowano np. do wyostrzania zdjęcia wirusa fd, uzyskanego za pomocą mikroskopu elektronowego; dzięki temu po raz pierwszy wyraźnie uwidoczniono strukturę heliksu tego wirusa, podobną do heliksu cząsteczki białka DNA (il. 147, tabl. 39).

Jeżeli do zapisu i odtwarzania hologramu zastosujemy układy optyczne, w których źródło fali odniesienia i źródło fali przedmiotowej mają złożoną rozciągłą strukturę, to obraz holograficzny otrzymany przy odtwarzaniu falą świetlną różniącą się od fali odniesienia będzie miał również złożoną strukturę. Obraz ten będzie czytelny, jeżeli fala odtwarzająca będzie identyczna z falą odniesienia. Jeżeli fala odniesienia będzie miała odpowiednio złożoną strukturę (np. źródłem fali odniesienia będzie oświetlona matówka o zupełnie przypadkowej strukturze), to odtworzenie obrazu będzie możliwe tylko w wypadku zastosowania przy odtwarzaniu tej samej matówki. Z teoretycznego punktu widzenia jest to idealny sposób szyfrowania informacji (rys. 12).

szyfrowanie informacji

Ten sam układ może być wykorzystany do optycznego porównywania dwóch obrazów mających określone rozkłady amplitudy i fazy. Jako przedmiot stosuje się wówczas punktowe źródło światła, a porównywane przedmioty odgrywają rolę odpowiednio rozciągłego źródła fali odniesienia i fali odtwarzającej. Przy otwarzaniu punktowy obraz otrzymuje się tylko wtedy, gdy oba przedmioty mają identyczną strukturę amplitudowo-fazową. W przeciwnym razie odtworzony obraz ma kształt rozmytej plamki, a stopień jej rozmycia może być miarą różnic w porównywanych przedmiotach. Najbardziej znane przykłady zastosowania holografii do tego celu — to identyfikacja odcisków palców czy analiza grafologiczna rękopisów. Metoda ta może też służyć do automatycznego poszukiwania określonych danych z dużego zbioru (np. wyszukiwanie prac o określonej tematyce, zawierających określone słowa-hasła, jak holografia, lasery itp.).

pamięć holograficzna

Obecnie produkowane materiały fotograficzne o wysokiej zdolności rozdzielczej umożliwiają gromadzenie dużej ilości informacji na małych powierzchniach (duża gęstość zapisu informacji, przewyższająca wszelkie inne znane sposoby zapisu — magnetyczne, elektroniczne itp.).

W odróżnieniu od innych metod, w których zapis informacji ma lokalny charakter, na hologramie informacja o poszczególnych fragmentach obrazu może być zapisana na całej jego powierzchni; wskutek tego lokalne uszkodzenia hologramu nie powodują straty odpowiedniego fragmentu obrazu, lecz jedynie spadek jego jasności, proporcjonalny do wielkości uszkodzonej części, oraz zmniejszenie kontrastu.

gęstość zapisu informacji

Teoretyczna analiza wykazuje, że holograficznymi metodami na płaskich hologramach można rejestrować informację z gęstością ok. $8 \cdot 10^9$ bitów/cm², a z użyciem hologramów objętościowych z gęstością 10^{12}–10^{13} bitów/cm³. W praktyce na cienkich emulsjach fotograficznych uzyskuje się gęstość zapisu 10^4–10^5 bitów/ mm² — przy zapisie informacji w układzie binarnym (w postaci ciemnych lub jasnych kwadratów) lub 2000 liter alfabetu — przy holograficznej rejestracji tekstu drukowanego

Już obecnie zbudowano kilka urządzeń do magazynowania informacji w postaci mikrohologramów z zarejestrowanymi stronami tekstu, obrazami graficznymi lub stronami zawierającymi informację w postaci binarnej (rys. 13). Umożliwiają one przedstawienie zarejestrowanej informacji w postaci wizualnej, a niektóre z nich mogą być również podłączone do komputerów jako zewnętrzna pamięć stała.

Holografia w technice audiowizualnej

holografia w dydaktyce

W przyszłości holografia znajdzie niewątpliwie zastosowanie w dziedzinie dydaktyki; trójwymiarowe obrazy holograficzne zastąpią z czasem inne pomoce naukowe we wszystkich wypadkach, w których ważne będzie zachowanie trójwymiarowości, a demonstracja naturalnych obiektów nie będzie możliwa, jak również niemożliwe będzie zastąpienie ich modelami (np. rzadkie lub chronione okazy fauny i flory, jedyne w swoim rodzaju okazy geologiczne lub wykopaliskowe, niepowtarzalne dzieła sztuki, przedmioty historyczne itp.).

Zarejestrowane na cienkim materiale światłoczułym obrazy holograficzne muszą być odtwarzane w świetle monochromatycznym, toteż odtworzone obrazy mają barwę światła odtwarzającego i nie odzwierciedlają barwy przedmiotu. Tego typu hologramy mogą więc znaleźć zastosowanie wówczas, gdy barwa obrazu nie odgrywa istotnej roli, natomiast ważne jest zachowanie trójwymiarowości.

hologramy barwne

Dalsze rozpowszechnienie holografii do celów demonstracyjnych będzie zależało od możliwości odtworzenia naturalnej barwy przedmiotu. Sposoby uzyskiwania hologramów barwnych przedmiotów niesamoświecących są już znane i sprowadzają się do wykonywania hologramu z użyciem trójbarwnego promienia laserowego, natomiast nie można na razie uzyskać w sposób bezpośredni hologramów przedmiotów samoświecących. W obu wypadkach zagadnienie to jest bardzo złożone i wciąż pozostaje do rozwiązania wiele problemów technicznych i technologicznych.

Po rozwiązaniu wszystkich tych trudności, a zwłaszcza po opanowaniu techniki holografowania żywych obiektów oraz technologii odtwarzania hologramów w świetle nielaserowym, można spodziewać się powstania atelier holograficznych, w których można będzie zamawiać portrety holograficzne.

Ponieważ metodami holograficznymi uzyskuje się z płaskiego interferencyjnego obrazu obraz trójwymiarowy, zdawałoby się rzeczą naturalną wykorzy-

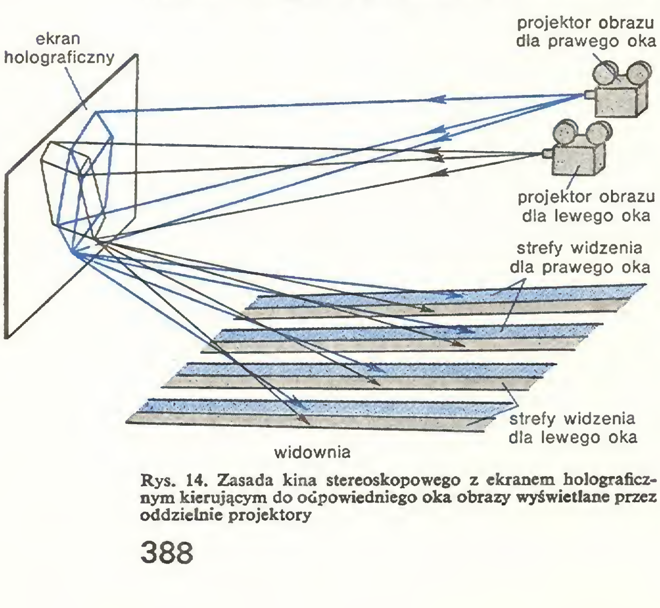
kino holograficzne

stanie holografii w kinie i telewizji. W tym jednak wypadku dochodzi dodatkowy problem — utrwalenia ruchu. Ze względów technicznych wydaje się mało prawdopodobne, aby kiedykolwiek zrealizowano kino lub telewizję trójwymiarową w postaci szybko wymienianych hologramów wielkości ekranu. W kinie holograficznym badania zmierzają obecnie w dwóch kierunkach. Jeden ze sposobów rozwiązania tego zagadnienia polega na tym, że za pomocą lasera impulsowego o odpowiednio wysokiej mocy i odpowiedniej częstości powtarzania błysków na błonie filmowej jest rejestrowana sekwencja hologramów pomniejszonych trójwymiarowych obrazów przedmiotu, które przy odtwarzaniu są ponownie powiększane i za pomocą specjalnych ekranów (wykonanych również metodami holograficznymi) są odwzorowywane w przestrzeni przed widownią.

Innym rozwiązaniem są kina stereoskopowe. Są w nich wyświetlane dwa obrazy — jeden odpowiada lewemu, drugi prawemu oku — kierowane niezależnie do każdego oka za pomocą ekranów holograficznych. Ekrany te, wykonane w postaci grubego hologramu, mają tę właściwość, że rzutowane na nie obrazy z projektora ustawionego w określonym punkcie przestrzeni są odbijane na widownię w taki sposób, że można je obserwować tylko wówczas, gdy oko widza trafi w jedną z dużej liczby stref, na które podzielona jest widownia. Naprzemianległe strefy różnią się tym, że do jednej trafia obraz wyświetlany z jednego projektora, do drugiej — z drugiego. Odległość między strefami jest dobrana tak, że zawsze jedno oko znajduje się w jednej strefie, drugie — w drugiej (rys. 14). Jeżeli więc projektory jednocześnie wyświetlają dwa obrazy odpowiadające punktom widzenia poszczególnych oczu, to u widza powstaje wrażenie głębi. Ekrany te muszą być wykonane z bardzo wysoką dokładnością, co przy dużych ich wymiarach jest poważną przeszkodą; jednak z czasem na pewno można ją będzie przezwyciężyć.

telewizja holograficzna

Bezpośrednie zastosowanie holografii do celów telewizyjnych nie jest w chwili obecnej możliwe ze względu na ograniczoną przepustowość kanału telewizyjnego; nie może on w standardowym czasie 0,04 s przekazać ilości informacji niezbędnej do uzyskania hologramu. 1 mm^2 dobrego hologramu zawiera mniej więcej tyle informacji, co pełny obraz telewizyjny. Aby tę informację można było odczytać, układ analizujący kamery powinien mieć zdolność rozdzielczą 10–20 razy wyższą od obecnie stosowanej, przy czym analizowana powierzchnia powinna być równa powierzchni ekranu telewizyjnego. Aby zaś cały ten obraz można było przekazać przez układ telewizyjny w odpowiednim czasie, jego przepustowość powinna być ok. 10^5 razy większa niż kanałów stosowanych obecnie. Następnym problemem jest wytworzenie hologramu w odbiorniku telewizyjnym z taką samą zdolnością rozdzielczą jak na wejściu, a także odtworzenie obrazu z przekazanego hologramu. Na razie zagadnienia te nie zostały w zadowalający sposób rozwiązane. Wydaje się, że jedynym wyjściem byłoby znalezienie sposobu redukcji przekazywanej informacji, tym bardziej, że jak wykazują badania — do wywołania pełnego wrażenia głębi wystarczy ilość informacji najwyżej 10–15 razy większa niż w zwykłym obrazie telewizyjnym.



Rys. 14. Zasada kina stereoskopowego z ekranem holograficznym kierującym do odpowiedniego oka obrazy wyświetlane przez oddzielnie projektory

Inne zastosowania

Znając matematyczny opis kształtu przedmiotu, jego położenie w przestrzeni oraz położenie źródła oświetlającego przedmiot i źródła fali odniesienia, można obliczyć rozkład natężenia światła w płaszczyźnie hologramu, a tym samym i jego zaczernienia. Taki sam rozkład zaczernień płyty holograficznej można uzyskać naświetlając ją odpowiednio punkt po punkcie. W ten sposób można uzyskać hologram obiektu realnie nie istniejącego, co wykorzystuje się do modelowania kształtu nowych wyrobów z pominięciem procesu wykonywania modelu plastycznego. Zaletą holograficznego modelu jest możliwość przekształceń obrazu przez zmianę geometrii układu odtwarzającego. Ta właściwość znajduje praktyczne zastosowanie do modelowania w polu widzenia pilota trójwymiarowego obrazu lądowiska przy ślepym lądowaniu samolotu.

modelowanie holograficzne

pomiar kształtu

Metody holograficzne można także stosować do pomiaru kształtu lub rozkładów przestrzennych. Stosując lasery impulsowe, tą metodą można badać zachowanie się różnych obiektów w ruchu.

Kiedy oświetlimy hologram falą sprzężoną do fali odniesienia, to w miejscu, w którym się znajdował przedmiot, powstanie jego wierny obraz. Dokładność odwzorowania i zdolność rozdzielcza w tym obrazie będą tym większe, im dokładniej fala odtwarzająca będzie oddawała kształt fali odniesienia oraz im większa będzie powierzchnia hologramu. Odtworzony w ten sposób obraz można obserwować bezpośrednio na matówce lub fotografować bez stosowania żadnych elementów optycznych (zazwyczaj obniżających jakość obrazu), wstawiając płytę fotograficzną w miejscu tworzenia się obrazu. Następnie na podstawie tego obrazu można dokładnie wyznaczyć kształt obiektu.

Holograficznie można także z wysoką zdolnością rozdzielczą rzutować obrazy. Zarejestrowany na hologramie obraz o złożonej drobnej strukturze może być rzutowany na dowolną powierzchnię przez oświetlenie hologramu falą odtwarzającą we wspomniany wyżej sposób. Ponieważ zdolność rozdzielcza układu optycznego zależy od jego apertury, czyli od stosunku wymiarów poprzecznych hologramu do jego odległości od obiektu, to można zapewnić odwzorowanie z dużą zdolnością rozdzielczą. Z tego względu metoda ta może być przydatna w mikroelektronice — do nanoszenia odpowiednich obrazów w trakcie produkcji układów scalonych. Dodatkowe zalety holografii polegają na tym, że rzutowanie odbywa się w sposób bezdotykowy i że małe uszkodzenia powierzchni hologramu nie mają istotnego wpływu na jakość rzutowanego obrazu.

zastosowanie w mikroelektronice

Przytoczone tu przykłady zastosowań holografii obejmują zaledwie niewielką ich część. Na wymienienie zasługują jej zastosowania w technologii elementów optycznych (do wykonywania siatek dyfrakcyjnych, elementów ogniskujących, elementów optyki zintegrowanej, odchylaczy wiązek świetlnych itd.), w mikroskopii rentgenowskiej i elektronowej (metodami holograficznymi udało się po raz pierwszy uzyskać obraz powłoki L atomów neonu i argonu), w radiolokacji (wykorzystanie holografii umożliwia uzyskanie na lecącym samolocie obrazów radiolokacyjnych o zdolności rozdzielczej porównywanej

z uzyskiwaną w zdjęciach lotniczych), w geofizyce (holografia służy do obróbki sejsmogramów), w tomografii rentgenowskiej (do uzyskania trójwymiarowych obrazów) i w wielu innych dziedzinach.

Przedstawiony, niezupełny zresztą, przegląd zastosowań holografii dobitnie ukazuje, jak stosunkowo prosta idea wywodząca się z podstawowych teorii fizycznych może doprowadzić do ujawnienia wielkich możliwości technicznych, inspiruje do szukania możliwości praktycznego jej zastosowania. Jednocześnie na przykładzie holografii widać, że wielu takich możliwości jeszcze nie wykorzystano, że kryją je w sobie teorie fizyczne na pozór całkowicie poznane i że do ujawnienia ich trzeba nieco odmiennego niż powszechnie przyjęte — spojrzenia.

A. Piekara *Nowe oblicze optyki*, Warszawa 1976; H. Royer, P. Śmigielski, J.Ch. Viénot *Holografia optyczna*, Warszawa 1975, *Holografia, podstawy fizyczne, teoria i zastosowanie*, pr. zbiorowa, Warszawa 1979.

Optyka fourierowska

Andrzej G. Kalestyński

Wykorzystanie światła jako nośnika informacji, a układów optycznych jako urządzeń służących do jej przetwarzania charakteryzowało optykę od samych jej początków. Postęp w optyce, który nastąpił w ostatnich latach, jest związany z jednej strony z pojawieniem się źródeł światła spójnego — laserów, z drugiej — z nową metodą opisu zjawisk zachodzących w optycznym torze przesyłania informacji, polegającą na zastosowaniu języka zaczerpniętego z telekomunikacji. Należy jednak pamiętać, że owo wykorzystanie pojęć i opisu typowego dla kanałów łączności nieoptycznej stało się możliwe dzięki poprzednim pracom i odkryciom optyków: E. Abbego, J. Rayleigha, A. Michelsona, M. Wolfkego (profesora Politechniki Warszawskiej), F. Zernikego, A. Mareshala, N. Francona, D. Gabora i in.

Dopóki wierność odwzorowań optycznych oceniano okiem, nie było większej potrzeby uściślania kryteriów jakości obrazów optycznych. Dopiero rozwój telewizji, radiolokacji oraz radioastronomii skłonił do spojrzenia na proces tworzenia obrazów z punktu widzenia teorii informacji, która się początkowo rozwijała na gruncie badania kanałów łączności nieoptycznej. Pojawienie się i doskonalenie laserów, wydajnych źródeł światła spójnego, tych swoistych anten świetlnych, otworzyło wielkie możliwości przed optycznym przenoszeniem i przetwarzaniem danych i ich zbiorów (czyli wiadomości) przy użyciu światła jako nośnika informacji; podkreślić trzeba, że w procesie tym ważne jest nie tylko otrzymanie wiernego obrazu (ewentualnie powiększonego lub pomniejszonego), lecz również jego odpowiednie przekształcenie. W optyce wiadomość to po prostu obraz. Przenoszenie i przetwarzanie wiadomości w kanałach optycznych można opisać za pomocą pojęcia transformacji optycznej. Przejście od rozkładu pola świetlnego przedmiotu do rozkładu pola świetlnego w obrazie dyfrakcyjnym, wytworzonym przez ten przedmiot, jest przykładem transformacji optycznej pola świetlnego, zachodzącej dzięki zjawiskom ugięcia i interferencji światła. Transformacjom optycznym ulega światło przechodzące przez rozmaite ośrodki naturalne lub elementy optyczne. Transformacje Fouriera i Fresnela, odpowiadające dwóm typom zjawisk dyfrakcyjnych — dyfrakcji Fraunhofera i dyfrakcji Fresnela, są podstawą opisu większości procesów wytwarzania obrazów w przyrządach optycznych zarówno tradycyjnych, jak i całkiem nowych, np. w komputerach optycznych czy układach holograficznych. Optyka fourierowska opisuje procesy przetwarzania sygnałów optycznych z punktu widzenia transformacji Fouriera.

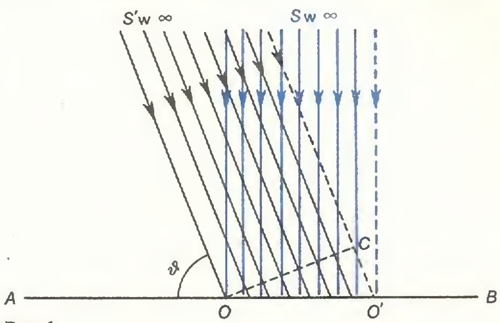
światło jako nośnik informacji

optyczne przetwarzanie informacji

Transformacja Fouriera w optyce

Istotę transformacji Fouriera można wytłumaczyć na następującym przykładzie. Wyobraźmy sobie, że punktowe źródło światła monochromatycznego S (rys. 1) znajduje się tak daleko („w nieskończoności"), że falę świetlną, która dociera do odcinka AB prostej prostopadłej do kierunku rozchodzenia się światła, można uważać za falę płaską. Obserwator dysponujący metodą pomiaru natężenia światła (wielkości związanej z jego energią) i fazy (wielkości związanej z drogą, którą przebywa fala świetlna) stwierdzi, że obie te wielkości są jednakowe wzdłuż całego odcinka AB. Załóżmy następnie, że źródło światła leży w punk-

transformacja Fouriera



Rys. 1.

cie S, tak że odcinek AB leży pod kątem ϑ do kierunku rozchodzenia się światła. Obserwator i tym razem stwierdzi, iż natężenie światła na prostej AB jest jednakowe. Okaże się jednak, iż fazy wyznaczone np. w punktach O i O' będą różne z powodu różnicy dróg optycznych CO'. Wzdłuż odcinka AB faza zmienia się periodycznie z okresem zależnym od kąta ϑ.

Rozważaną sytuację można odwrócić i potraktować odcinek AB jako źródło światła o jednakowym natężeniu i jednakowych fazach wzdłuż całej długości odcinka AB. Można się spodziewać, że wtedy światło o jednakowej fazie będzie się rozchodziło w kierunku S. Natomiast jeśli faza będzie się zmieniała wzdłuż AB w określony wyżej sposób, to światło o jednakowej fazie będzie się rozchodziło w kierunku S', a nie S. Zatem istnieje związek między przestrzennym rozkładem pola wzdłuż odcinka (współrzędna x) a rozkładem kątowym (współrzędna ϑ). Jeden rozkład pola świetlnego może być przekształcony w drugi rozkład pola świetlnego za pomocą przekształcenia Fouriera $\mathcal{F}$:

$$f(x) \overset{\mathcal{F}}{\rightleftarrows} F(\vartheta), \text{ czyli } \begin{cases} F(\vartheta) = \mathcal{F}\{f(x)\}, \\ f(x) = \mathcal{F}\{F(\vartheta)\}, \end{cases}$$

co oznacza, że funkcja $F(\vartheta)$ jest transformatą Fouriera $f(x)$ i odwrotnie — funkcja $f(x)$ jest transformatą funkcji $F(\vartheta)$, a $\mathcal{F}\{\}$ jest operatorem przekształcenia Fouriera.

transformata Fouriera

Podobne rozumowanie można przeprowadzić w przypadku dwóch zmiennych, rozpatrując zamiast odcinka AB płaszczyznę prostopadłą do kierunku S. Dochodzi się wówczas do przekształceń:

$$f(x,y) \overset{\mathcal{F}}{\rightleftarrows} \Psi(\vartheta, \varphi).$$

Wyobraźmy sobie, że znamy pewne pole świetlne na płaszczyźnie x_0, y_0, czyli znamy rozkład natężenia pola i rozkład jego fazy we wszystkich punktach tej

płaszczyzny. Pytamy, jakie będzie pole świetlne w płaszczyźnie x, y odległej od z od płaszczyzny x_0, y_0? Otóż J. W. Rayleigh zaważył, że aby otrzymać odpowiedź na to pytanie, najlepiej rozłożyć wyjściowe pole świetlne na nieskończony zbiór fal płaskich rozbiegających się pod różnymi kątami ϑ i φ (→ Holografia). Oznacza to dokonanie transformacji Fouriera i przejście w opisie od współrzędnych x_0, y_0 do współrzędnych kątowych ϑ, φ. Otrzymuje się w ten sposób tzw. widmo kątowe $\hat{P}(\vartheta, \varphi)$ pola świetlnego danego w płaszczyźnie x_0, y_0, czyli rozkład amplitud i faz płaskich fal składowych w zależności od kierunków ich rozchodzenia się określonych kątami ϑ, φ. Rozkład pola świetlnego (natężenia i fazy) w płaszczyźnie x, y odległej o z od płaszczyzny wejściowej otrzymuje się przez całkowanie widma kątowego wejściowego pola świetlnego po współrzędnych kątowych. Rozkład pola świetlnego na zbiór elementarnych fal płaskich przypomina rozkład drgań struny o unieruchomionych końcach na drgania składowe o częstościach równych wielokrotnościom częstości podstawowej (częstości harmoniczne). Funkcję opisującą kształt struny w danej chwili można przedstawić w postaci sumy drgań sinusoidalnych, zwanej szeregiem Fouriera. W optyce zwykle nie ma wyróżnionego dyskretnego zbioru fal składowych, dlatego posługujemy się całkami, a nie szeregami Fouriera.

widmo kątowe pola świetlnego

Zamiast współrzędnych kątowych ϑ, φ można wprowadzić zmienne $\omega_x = 2\pi \cos\vartheta/\lambda, \omega_y = 2\pi \cos\varphi/\lambda$, gdzie λ — długość fali świetlnej. Zmienne ω_x, ω_y nazywają się częstościami (częstotliwościami) przestrzennymi i odgrywają rolę analogiczną do częstości czasowej ω_t, występującej m.in. w radiotechnice. Widmo kątowe wyrażone w zmiennych ω_x, ω_y, czyli $\hat{P}(\omega_x, \omega_y)$ nazywa się widmem częstościowym przestrzennym pola świetlnego. Do opisu barwy fali świetlnej wystarcza częstość czasowa, natomiast do pełnego opisu rozchodzenia się w przestrzeni trójwymiarowej sygnałów optycznych trzeba się jeszcze posłużyć częstościami przestrzennymi. Z tego już widać, że układy optyczne są sposobniejsze do przetwarzania informacji niż układy elektroniczne. Niewielka bowiem zmiana barwy w optyce odpowiada modulacji częstości (czasowych) w elektronice, ale modulacja częstości przestrzennych w optyce nie ma żadnego odpowiednika w telekomunikacji nieoptycznej. Tor optyczny jest układem optycznym składającym się z filtrów częstości przestrzennych i modulatorów, którymi są np.: przestrzeń swobodna, soczewki sferyczne, cylindryczne czy stożkowe, siatki dyfrakcyjne różnych typów oraz najrozmaitsze przesłony umieszczane w różnych płaszczyznach optycznych. W torze (układzie) optycznym można zawsze rozróżnić wejście i wyjście, tzn. miejsce, gdzie się wprowadza dane (sygnały optyczne), oraz miejsce, gdzie się je wyprowadza (w postaci np. obrazów optycznych). W wypadku płaskich sygnałów optycznych mówi się o płaszczyźnie wejściowej i wyjściowej.

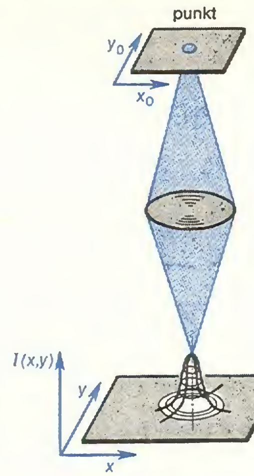
częstości przestrzenne

tor optyczny

Zdolność rozdzielcza układu optycznego, który dokonał przetworzenia wejściowego sygnału optycznego w obraz przedmiotu, wiąże się ściśle z tzw. funkcją rozmycia. Wyjaśnimy to pojęcie na przykładzie soczewki wytwarzającej obraz świecącego punktu. Obraz punktowego przedmiotu nigdy w praktyce nie jest punktem. Dyfrakcja na skończonej źrenicy soczewki, aberracje, a także i rozproszenia powodują, iż obraz punktu jest rozmyty. Nawet jeśli się posłużymy np. soczewką sferyczną o dobrze skorygowanych aberracjach (bezaberracyjną) i całkowicie usuniemy rozproszenie w układzie optycznym, obraz przedmiotu punktowego będzie rozmyty. Ogólnie biorąc, pole świetlne obrazu przedmiotu punktowego w płaszczyźnie obrazowej soczewki można opisać za pomocą zespolonej funkcji rozmycia $\hat{R}(x,y) = R(x,y)e^{i\varphi(x,y)}$. Kwadrat modułu tej funkcji $|\hat{R}(x,y)|^2$ jest równy rozkładowi natężenia światła obrazu i można go wyznaczyć doświadczalnie (rys. 2).

funkcja rozmycia

Funkcja $\varphi(x, y)$ określa przesunięcie fazy między przedmiotem i obrazem. Jeśli to jest układ bezaberracyjny, funkcja rozmycia wszystkich punktów x_0,



Rys. 2. Rozmycie obrazu punktowego źródła światła. Otworek jest źrenicą fotodetektora służącego do pomiaru rozkładu natężenia rozmycia

y_0 przedmiotu jest symetryczna, a funkcja φ równa jest zeru. W rzeczywistych układach optycznych, gdy występują np. aberracje poprzeczne, $\varphi \neq 0$.

Obraz przedmiotu złożonego, składającego się z wielu punktów, można opisać sumując funkcje rozmycia wszystkich punktów. Zdolność rozdzielcza układu optycznego jest tym większa, im ostrzejsze maksimum ma funkcja rozmycia tego układu dla przedmiotu punktowego. Funkcję rozmycia często nazywa się odpowiedzią impulsową układu optycznego.

zdolność rozdzielcza

W procesie tworzenia obrazu ważne jest także, w jakim stopniu kontrast z przedmiotu może być przeniesiony do obrazu. Dwie linie szare przedmiotu łatwo się zlewają w obrazie w jedną, podczas gdy linię czarną sąsiadującą z białą ten sam układ optyczny może rozdzielić. Miarą kontrastu jest współczynnik $\gamma = (I_{max}-I_{min})/(I_{max}+I_{min})$, gdzie I_{max} i I_{min} — natężenie światła w najjaśniejszych i najciemniejszych miejscach przedmiotu. Współczynnik przenoszenia kontrastu przez układ optyczny można określić jako stosunek kontrastu obrazu γ_{obrazu} do kontrastu przedmiotu $\gamma_{przedmiotu}$:

$$H = \gamma_{obrazu}/\gamma_{przedmiotu}.$$

Okazuje się, że współczynnik H zależy od częstości przestrzennych ω_x, ω_y; nosi on nazwę funkcji przenoszenia kontrastu lub funkcji przenoszenia modulacji $H(\omega_x, \omega_y)$.

Przedmiot oświetlony, czyli sygnał optyczny na wejściu układu optycznego charakteryzuje się także rozkładem fazy. Zatem oprócz przenoszenia kontrastu trzeba także brać pod uwagę funkcję przenoszenia fazy, $\Psi = \Psi(\omega_x, \omega_y)$. Tak więc funkcja $\hat{H}(\omega_x, \omega_y) = H(\omega_x, \omega_y)e^{i\Psi(\omega_x, \omega_y)}$ jest pełną optyczną funkcją przenoszenia danego układu optycznego. Działanie układu optycznego można przeto opisać albo za pomocą funkcji przenoszenia $\hat{H}(\omega_x, \omega_y)$, albo funkcji rozmycia $\hat{R}(x, y)$, przy czym jedna z tych funkcji jest transformatą fourierowską drugiej.

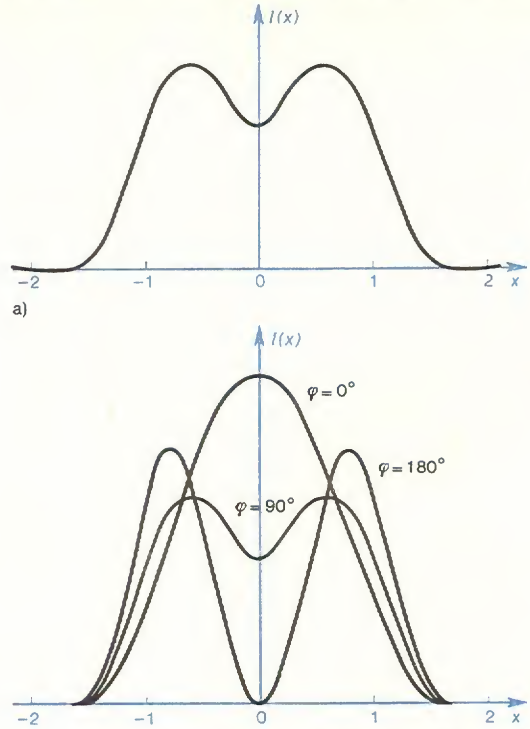
funkcja przenoszenia układu optycznego

Funkcję przenoszenia układu optycznego można wyznaczyć z porównania rozkładu natężenia światła przedmiotu i obrazu (w wypadku oświetlenia niespójnego) lub rozkładu amplitud i fazy pola świetlnego przedmiotu i obrazu (w wypadku oświetlenia spójnego).

Znajomość funkcji przenoszenia układu optycznego spójnego lub niespójnego umożliwia wyznaczenie sygnału (obrazu) na wyjściu układu przy danym sygnale na wejściu.

Układy optyczne różnią się nieco swym działaniem — zależnie od tego, czy użyto źródła światła spójnego czy też niespójnego (koherentnego czy nie-

koherentnego). Chodzi tu nie tylko o spójność czasową, lecz także spójność przestrzenną, czyli występowanie stałej różnicy faz w całym przekroju wiązki



Rys. 3. Rozkład natężenia w obrazie dwóch źródeł punktowych, leżących w minimalnej odległości rozróżnialnej (wg kryterium Rayleigha): źródła niespójne (a) i źródła spójne o różnicy faz φ (b)

oświetlającej przedmiot na wejściu układu optycznego (→ Spójność światła). Dla uproszczenia mówi się o koherentnych i niekoherentnych układach optycznych. Różnica między obu typami układów ujawnia się wyraźnie w ich zdolności rozdzielczej. Rysunek 3a przedstawia rozkład natężenia w obrazie dwóch punktów oświetlanych światłem niespójnym, znajdujących się w odległości minimalnej wg kryterium Rayleigha (zgodnie z tym kryterium odwrotność zdolności rozdzielczej, czyli minimalna odległość między dwoma rozróżnialnymi punktami $\Delta\varepsilon = 1{,}22 d\lambda/l$, gdzie d — odległość obrazu od płaszczyzny soczewki wyjściowej, l — średnica źrenicy wyjściowej układu optycznego, zwykle jest to średnica najmniejszej soczewki użytej w układzie, λ — długość fali). Na rys. 3b pokazano rozkład natężenia obrazu owych dwóch punktów, z których każdy oświetlony jest światłem spójnym. Rozkład ten zależy od różnicy fazy światła w tych punktach. Jeśli różnica faz $\varphi = 90°$, to rozkład jest taki jak przy oświetleniu niespójnym; jeśli $\varphi = 0°$, zdolność rozdzielcza jest mniejsza niż w przypadku oświetlenia niespójnego, jeśli zaś $\varphi = 180°$, zdolność rozdzielcza jest większa. Ale sama zdolność rozdzielcza nie decyduje o tym, który typ oświetlenia jest lepszy. Zalety oświetlenia spójnego ujawniają się przy zapisywaniu pełnej informacji o sygnale świetlnym (amplitudy i fazy), przy filtracji częstości przestrzennych itd.

zdolność rozdzielcza układów koherentnych i niekoherentnych

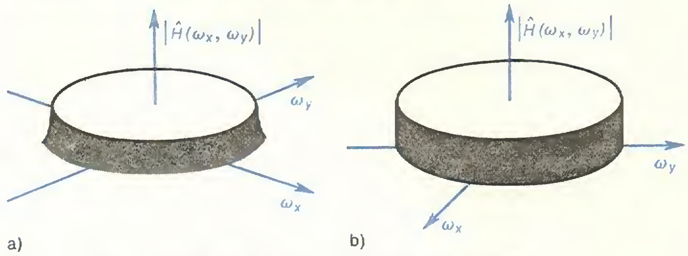
Omówimy dalej w ujęciu optyki fourierowskiej dwa podstawowe elementy optyczne: przestrzeń swobodną i soczewkę skupiającą.

Przestrzeń swobodna

Elementem optycznym jest nawet sama przestrzeń zawarta między płaszczyznami wejściową x_0, y_0 i wyjściową x, y odległymi o z. Choć się to wydaje dziwne, jednak przestrzeń swobodna jest także filtrem częstości przestrzennych i jej funkcja przenoszenia nie jest tożsamościowo równa jedności, jak można by oczekiwać. W rzeczywistości moduł funkcji przenoszenia jest następującą funkcją ω_x, ω_y (rys. 4a):

$$|\hat{H}(\omega_x, \omega_y)| = \begin{cases} 1 \text{ przy } (\omega_x^2+\omega_y^2) \leqslant k^2 = (2\pi/\lambda)^2, \\ \exp\left(-z\sqrt{\omega_x^2+\omega_y^2-k^2}\right) \\ \qquad\qquad \text{przy } \omega_x^2+\omega_y^2 > k^2. \end{cases}$$

Przestrzeń swobodna dla $z \gg \lambda$ jest filtrem częstości przestrzennych; nie przepuszcza ona częstości określonych warunkiem $\omega_x^2+\omega_y^2 > k^2$, odpowiadających tzw. falom niejednorodnym. W miarę wzrostu z (grubości warstwy) moduł $|\hat{H}(\omega_x, \omega_y)|$ coraz bardziej przypomina walec (rys. 4b). Filtrujące właściwości



Rys. 4. Moduł funkcji przenoszenia przestrzeni swobodnej w funkcji częstości przestrzennych; a) odległość z porównywalna z długością fali światła λ, b) $z \gg \lambda$

przestrzeni swobodnej, polegające na usuwaniu fal niejednorodnych, wynikają z ogólnych praw rozchodzenia się fal świetlnych w przestrzeni pozbawionej ośrodków materialnych. Przy dostatecznie dużej wartości z, sięgającej do bliskiej strefy dyfrakcji — strefy Fresnela, funkcję przenoszenia przestrzeni swobodnej można przedstawić wzorem:

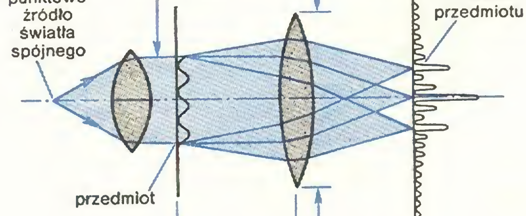
funkcja przenoszenia przestrzeni swobodnej

$$\hat{H}(\omega_x, \omega_y) = \exp[(-iz/2k)(\omega_x^2+\omega_y^2)].$$

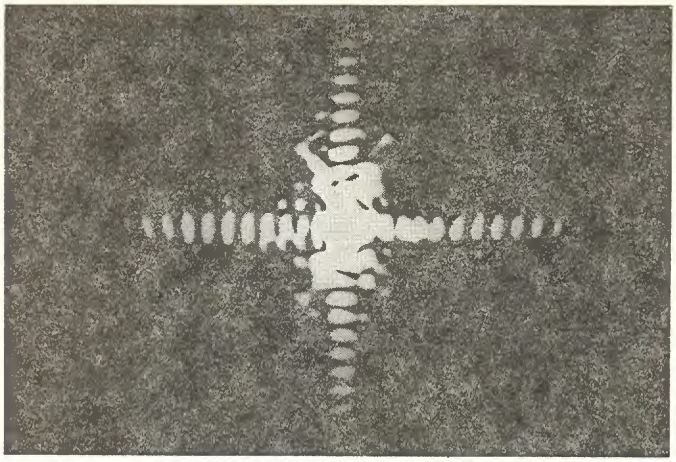
Soczewka skupiająca

Aby lepiej zrozumieć właściwości soczewki jako swoistego modulatora realizującego przekształcenie Fouriera, rozpatrzmy doświadczenie, w którym wiązka światła spójnego oświetla przeźrocze leżące przed soczewką w odległości równej np. jej odległości ogniskowej (rys. 5). W płaszczyźnie ogniskowej soczewki powstanie pewien charakterystyczny rozciągły obraz (światło skupiłoby się w jednym punkcie, gdyby samą soczewkę oświetlić bezpośrednio falą płaską i gdyby soczewka ta miała wymiary poprzeczne nieskończone). Ów rozciągły obraz jest obrazem Fouriera przedmiotu umieszczonego na wejściu układu (jest on także zwany widmem Wienera optycznego sygnału przedmiotu wejściowego). Soczewka dokonuje zatem przekształcenia Fouriera. Gdy światło trafia na swej drodze na przeszkodę (przeźrocze), ulega ugięciu, nie rozchodzi się dalej w tym samym kierunku, lecz rozbiega się na boki i interferuje, tworzy obrazy inter-

widmo Wienera



Rys. 5. Schemat działania soczewki sferycznej jako operatora przekształcenia Fouriera





Rys. 6. Obraz Fouriera pola świetlnego przedmiotowego (tzw. widmo Wienera) powstałego przez oświetlenie światłem spójnym przezrocza przedstawiającego czarny kwadrat z białym trójkątem w środku (obok)

ferencyjne, które na pozór mało mają wspólnego z przedmiotem. Obrazy są tym bardziej kontrastowe, wyraźniejsze, im światło ma większą zdolność interferowania, czyli im bardziej jest spójne czasowo i przestrzennie. Fale świetlne ugięte na przeźroczu skupiają się po przejściu przez źrenicę soczewki i nadal interferują, tworząc w płaszczyźnie ogniskowej soczewki ów charakterystyczny obraz Fouriera przedmiotu. Przyjrzyjmy się temu obrazowi (rys. 6). Widzimy, że najwięcej światła rozchodzi się w kierunkach bliskich osi układu optycznego, a najmniej — w dużym oddaleniu od osi. Punkty o różnej jasności w obrazie Fouriera przedstawiają skupione fale płaskie, na które można rozłożyć pole świetlne sygnału wejściowego, czyli w tym wypadku — pole świetlne ugięte na przeźroczu. Odpowiadają one kierunkom rozchodzenia się tych fal płaskich, czyli odpowiednim częstościom przestrzennym. Przebieg zjawiska możemy ująć krótko w następujący sposób: w wyniku ugięcia światła na przeźroczu powstaje pole świetlne, którego energia dzięki działaniu soczewki skupia się głównie w zakresie niskich częstości przestrzennych, a tylko niewielka jej część — w zakresie wysokich częstości przestrzennych; oznacza to, że w pewnych kierunkach przedmiot ugina więcej światła, w innych mniej. Obserwowanie lub fotografowanie ogniska soczewki jest utrwaleniem tego rozkładu natężenia światła w funkcji częstości przestrzennych. Soczewka odgrywa rolę realizatora przekształcenia Fouriera. Funkcja przenoszenia soczewki skupiającej ma postać:

funkcja rzenoszenia soczewki

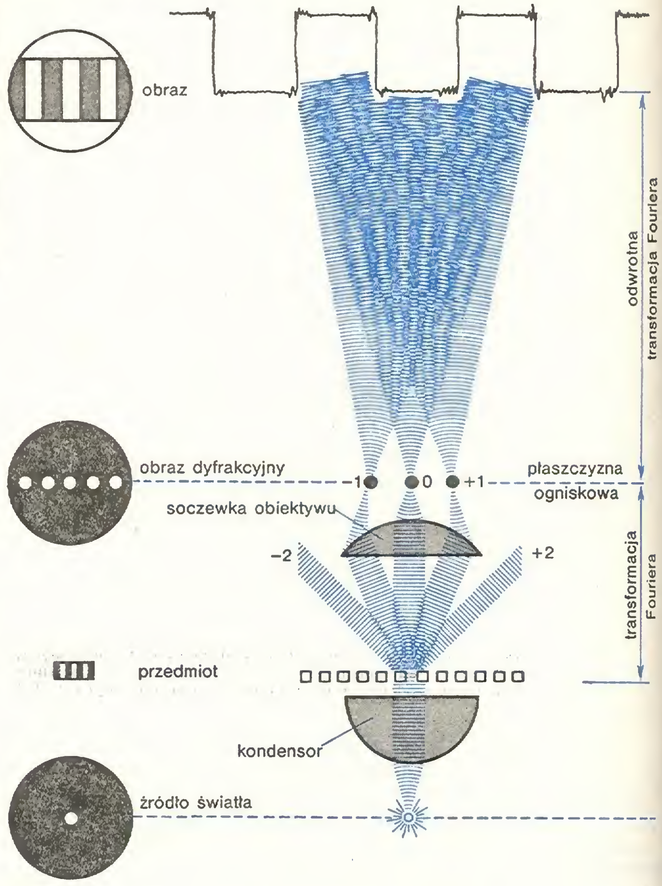
$$\hat{H}(\omega_x, \omega_y) = A \exp[i(f/2k)(\omega_x^2 + \omega_y^2)],$$

gdzie f — ogniskowa soczewki, A = const.

Soczewka o wymiarach poprzecznych nieskończonych działa podobnie jak gruba warstwa przestrzeni swobodnej, taka że grubość warstwy z sięga do dalekiej strefy dyfrakcji, zwanej strefą dyfrakcji Fraunhofera. W obu wypadkach sygnał optyczny wejściowy jest na wyjściu przekształcony w swoją transformatę Fouriera (z dokładnością do stałych czynników fazowych).

teoria Abbego mikroskopu

Doniosłe znaczenie transformacji Fouriera w procesie tworzenia obrazu zostało po raz pierwszy stwierdzone przez E. Abbego już w ubiegłym stuleciu. Zauważył on, że proces powstawania obrazu w mikroskopie jest dwustopniowy. Najpierw soczewka obiektywowa wytwarza pierwotny obraz Fouriera w swojej płaszczyźnie ogniskowej; w drugim etapie — pierwotny obraz Fouriera przedmiotu zostaje przekształcony w obraz optyczny (odwrotne przekształcenie Fouriera), który oglądamy przez okular (rys. 7).



Rys. 7. Dwustopniowy schemat powstawania obrazu w mikroskopie w świetle spójnym podany przez E. Abbego (wg J. R. Meyer-Arendt *Wstęp do optyki*, Warszawa 1977)

Filtrowanie częstości przestrzennych

Płaszczyzny w układzie optycznym, w których się pojawia obraz Fouriera pola świetlnego przedmiotu, umieszczonego na wejściu układu, nazywa się płaszczyznami widmowymi częstości przestrzennych. Uzyskany w płaszczyźnie widmowej obraz Fouriera sygnału optycznego można częściowo zasłaniać; taka operacja nosi nazwę filtracji częstości przestrzennych. Problemem filtracji częstości przestrzennych zajmowali się fizycy już od dziesięcioleci. Jako przykład można przytoczyć osiągnięcia N. Mareshala i współpracowników, którzy za pomocą filtracji częstości przestrzennych polepszali obrazy przekazywane przez fototelegrafię kopiową (il. 149, 150, tabl. 40). Przesłanianie częściowe obrazu Fouriera wpływa na cały obraz przedmiotu, każdy bowiem punkt obrazu Fouriera odpowiada całemu polu świetlnemu przedmiotu. Jak wpływa? Zależy to od sposobu filtrowania.

Powróćmy jeszcze do rys. 7, ilustrującego etapy powstawania obrazu w mikroskopie. Obraz przedmiotu składającego się z pasków na przemian czarnych i przezroczystych jest nieostry, rozmyty. Dlaczego? Rzeczywista soczewka ma skończone wymiary poprzeczne i odgrywa dwojaką rolę: modulatora (realizatora przekształcenia Fouriera) i ekranu czar-

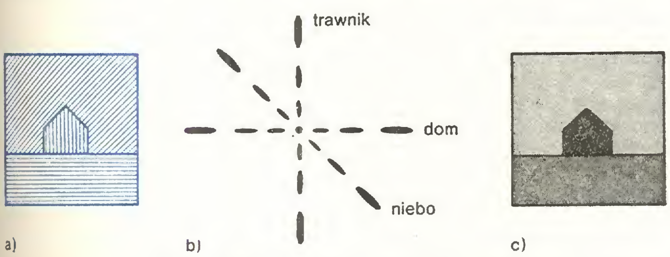
nego z otworem w środku, czyli jest także dolnoprzepustowym, dość wąskopasmowym filtrem częstości przestrzennych. Okazuje się, że natężenie światła odpowiadające niższym częstościom przestrzennym w obrazie Fouriera przedmiotu świecącego lub oświetlonego wiąże się z zarysami jego rysunku czy rzeźby, a natężenie odpowiadające najwyższym częstościom przestrzennym — z ostrością krawędzi rysunku. Obcięcie wyższych częstości prowadzi w obrazie do rozmycia krawędzi, do zmniejszenia zdolności rozdzielczej.

filtr częstości przestrzennych

Przesłony umieszczane celowo w płaszczyźnie widmowej danego układu optycznego nazywamy filtrami częstości przestrzennych lub ich maskami. Mogą one być rozmaite — zależnie od celu, jakiemu służą. Doskonałym przykładem jest doświadczenie Abbego–Portera, które miało duże znaczenie dla omawianej dziedziny fizyki. Przedmiot w postaci siatki z kwadratowymi oczkami oświetlano światłem spójnym (il. 148, tabl. 40) W płaszczyźnie ogniskowej układu optycznego wstawiano maskę filtrującą obraz Fouriera — szczelinę poziomą lub pionową, dzięki czemu otrzymywano obrazy składające się z pionowych lub poziomych pasków czarnych i białych.

Modulacja θ

Jedną z metod filtrowania częstości przestrzennych jest tzw. metoda modulacji θ, dzięki której uzyskuje się zamierzoną zmianę odcienia szarości fragmentów obrazu. Zasadę tej modulacji wyjaśnimy na prostym przykładzie (rys. 8). Weźmy pod uwagę widok zawierający różne elementy, np. niebo, dom, trawnik (na fotografii czarno-białej elementy te są przedstawione różnymi odcieniami szarości). Każdy element

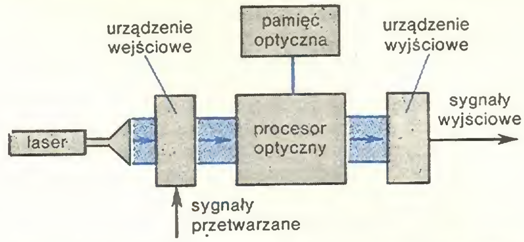


Rys. 8. Modulacja theta: a) widok wymodelowany z siatek dyfrakcyjnych zorientowanych pod różnymi kątami θ, b) obraz Fouriera tego modelu, c) obraz wyjściowy o różnych odcieniach szarości otrzymanych dzięki odpowiedniej filtracji częstości przestrzennych

widoku wykonuje się z kawałków plastykowej siatki dyfrakcyjnej o prążkach zorientowanych pod różnymi kątami θ (stąd nazwa metody). Przez odpowiednie przesłanianie różnych fragmentów obrazu Fouriera tak wymodelowanego widoku zmienia się odcień szarości elementów. Jeśli np. w końcowym obrazie dom ma być czarny, to trzeba przesłonić odpowiednie maksima widma częstości przestrzennych domu, jeśli zaś niebo ma być jasne — trzeba przepuścić cały jego obraz Fouriera. Ponieważ maksimum jasności obrazu Fouriera, odpowiadające częstościom przestrzennym bliskim 0, czyli ugięciu rzędu zerowego, nie zależy od kąta θ, trzeba je w tej metodzie filtrowania zasłaniać całkowicie. Modulację θ można stosować nie tylko w przytoczonej tu wersji czarno-białej, lecz także w barwnej.

Komputery optyczne

Łatwość dokonywania przekształcenia Fouriera za pomocą układów optycznych wykorzystuje się w koherentnych układach optycznych do przeprowadzania operacji matematycznych, czyli w komputerach optycznych. Schemat budowy komputera optycznego przedstawia rys. 9. Właściwą maszyną matematyczną jest układ optyczny dokonujący żądanej operacji ma-



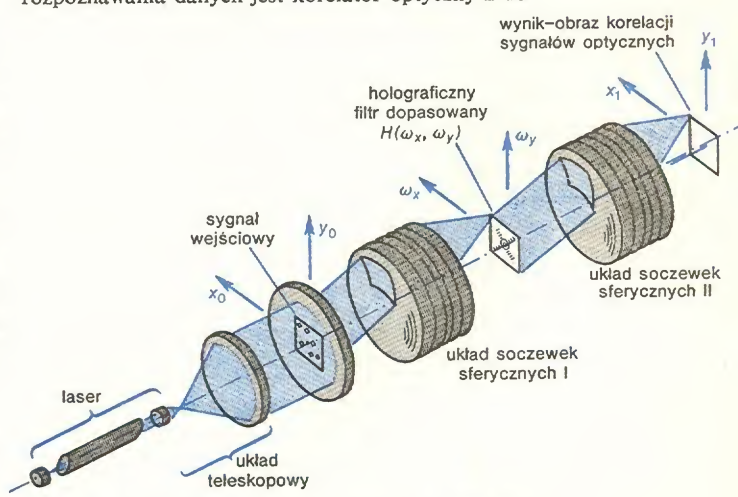
Rys. 9. Schemat blokowy komputera optycznego. Laser zaopatrzony jest w układ teleskopowy poszerzający wiązkę. Urządzenie wejściowe zmienia dane wejściowe w wejściowe sygnały optyczne, urządzenie zaś wyjściowe zmienia wyjściowe sygnały optyczne w dane pożądane dla odbiorcy

tematycznej, zw. procesorem optycznym. Pamięć operacyjną stanowi zbiór optycznych filtrów częstości przestrzennych, umieszcza się je w płaszczyźnie widmowej. Urządzenie wyjściowe służy do odbierania wyniku obliczeń w postaci obrazów optycznych, także do przetwarzania wyjściowych sygnałów optycznych na inne (np. elektryczne) — zgodnie z zapotrzebowaniem odbiorcy. Rodzaj użytego filtru częstości przestrzennych decyduje o rodzaju operacji matematycznej przeprowadzanej przez dany komputer optyczny. Jeśli liczba kolejnych operacji matematycznych wykonywanych przez procesor optyczny jest znaczna, trzeba w odpowiednim etapie obliczeń zmieniać filtry wstawiane w płaszczyzny widmowe układu. Filtry przechowywane są w bloku pamięci optycznej.

pamięć optyczna

Komputery optyczne to wyspecjalizowane maszyny optyczne przystosowane do wykonywania określonych operacji matematycznych. Jako przykład omówimy komputer optyczny służący do rozpoznawania danych. Jego zadanie polega na tym, aby ze zbioru sygnałów podawanych na wejściu komputera wybrać poszukiwany, czyli inaczej mówiąc — ze zbioru funkcji wybrać określoną (trzeba np. zidentyfikować linie papilarne palca osobnika podejrzanego o zbrodnię na podstawie policyjnej kartoteki linii papilarnych kryminalistów, liczącej zwykle wiele tysięcy pozycji). Wprawdzie istnieją elektroniczne maszyny cyfrowe przystosowane do rozpoznawania danych, lecz takie zadania są dla nich pracochłonne i długotrwałe, podczas gdy komputer optyczny wykonuje je łatwo i szybko.

Typowym i już klasycznym układem optycznym do rozpoznawania danych jest korelator optyczny z do-



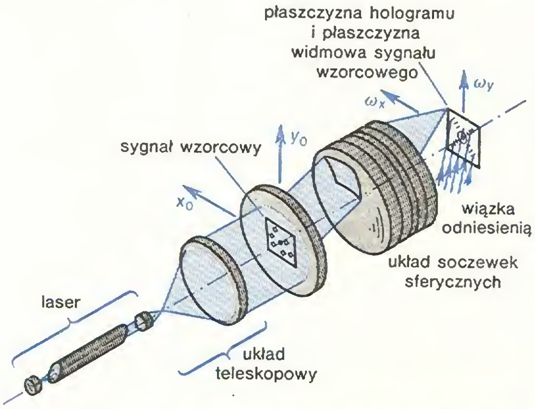
Rys. 10. Schemat układu optycznego korelatora Van der Lugta: laser z układem teleskopowym poszerzającym wiązkę i układy soczewek dokonujące przekształcenia Fouriera (I) i odwrotnego przekształcenia Fouriera (II)

korelator optyczny Van der Lugta

pasowanym filtrem holograficznym, zaproponowany przez Van der Lugta w 1963 r. (rys. 10). Na wejściu korelatora podaje się za pomocą przetwornika (może być nim zwykłe przezrocze) sygnał optyczny, który należy porównać z jakimś wzorcowym sygnałem optycznym, zapisanym w pamięci komputera w postaci dopasowanego filtru holograficznego częstości przestrzennych. Sygnał wejściowy poddawany jest najpierw transformacji Fouriera w układzie soczewek sferycznych, następnie pole świetlne, będące transformatą Fouriera sygnału wejściowego, odtwarza hologram wzorcowy (holograficzny filtr dopasowany), umieszczony dokładnie w płaszczyźnie widmowej pierwszego stopnia korelatora.

otrzymywanie filtru dopasowanego

Holograficzny filtr dopasowany Van der Lugta jest rodzajem hologramu Fouriera (→ Holografia), na którym zarejestrowano obraz interferencyjny widma Fouriera sygnału zadanego i wiązki odniesienia (widmo Fouriera jest rozkładem amplitudy i fazy sygnału optycznego w funkcji częstości przestrzennych, a obserwowane widmo Wienera jest tylko rozkładem natężenia w funkcji tych częstości). Zwykle do otrzymywania takich hologramów wykorzystuje się pierwszy stopień korelatora optycznego (rys. 11).



Rys. 11. Schemat powstawania dopasowanego holograficznego filtru częstości przestrzennych sygnałów wzorcowych

Drugi stopień korelatora dokonuje odwrotnego przekształcenia Fouriera pola świetlnego wiązki ugiętej na hologramie. Powstają trzy wiązki odtworzone — wiązka zerowa, rozchodząca się w zasadzie w kierunku osi układu, nie niosąca żadnej pożytecznej informacji, oraz dwie wiązki ugięte. Jedną z nich można nazwać korelacyjną, niesie ona informację o zgodności (niezgodności) rozpoznawanych obiektów. Drugą wiązkę ugiętą można nazwać splotową (odpowiada matematycznej operacji zwanej splotem), w tym wypadku nie odgrywa ona roli. Ugięta wiązka korelacyjna niesie informację o autokorelacji, gdy jest zgodność sygnału rozpoznawanego z sygnałem wzorcowym, lub o korelacji wzajemnej, gdy takiej zgodności nie ma. Rozkład jasności w ugiętej wiązce informacyjnej odpowiadający autokorelacji cechuje duża gęstość energii w środku obrazu, w wypadku obrazów korelacji wzajemnej rozkład jasności jest szeroko rozmyty. Wielkość natężenia w środku obrazu korelacyjnego (mierzonego za pomocą fotodetektora o małej średnicy) świadczy o identyczności lub odmienności sygnałów porównywanych i wzorcowych.

szybkość komputerów optycznych

Optyczne komputery zapewniają niezwykłą szybkość przeprowadzania operacji matematycznych — niezależnie od stopnia ich skomplikowania. Ogranicza ją szybkość wprowadzania i wyprowadzania danych, wymiana masek filtrujących częstości przestrzenne itp. Komputery optyczne dokonują przede wszystkim przekształceń całkowych, choć można ich także użyć do różniczkowania. Wadą ich jest jeszcze mała dokładność obliczeń (rzędu 1%) oraz ścisła specjalizacja; każdy komputer optyczny może wykonywać tylko określony rodzaj obliczeń. Nie są to maszyny matematyczne uniwersalne. Jaka jest ich przyszłość? Prawdopodobnie wkrótce nastąpi era maszyn hybrydowych — swoistych połączeń matematycznych maszyn optycznych i elektronicznych cyfrowych. Część zadań, szczególnie pracochłonnych, będą szybko wykonywały maszyny optyczne, a pozostałe dokładne obliczenia — elektroniczne maszyny cyfrowe. Takie hybrydowe maszyny już działają. Natomiast w dalszej przyszłości, po pełniejszym poznaniu możliwości zastosowania światła spójnego oraz dzięki zminiaturyzowaniu koherentnych układów optycznych (co należy do optyki zintegrowanej), nastąpi zapewne era powszechnego przesyłania i przetwarzania danych torami optycznymi. Cała telekomunikacja i maszynowa technika liczenia będą wyzyskiwały światło jako główny nośnik informacji. Elektronikę uzupełni czy nawet zastąpi nowa dziedzina — fotonika.

fotonika

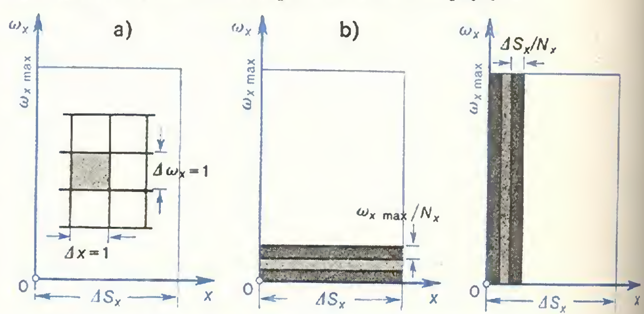
Struktura informacyjna sygnału optycznego

Z punktu widzenia optyki fourierowskiej można rozważać także strukturę informacyjną sygnału świetlnego. Krótkie omówienie tego zagadnienia rozpoczniemy od przypomnienia zasady nieokreśloności sygnałów radiowych: $\Delta\omega_t \cdot \Delta t \geqslant 1$, gdzie ω_t — częstość czasowa, t — czas trwania sygnału. Zasadę tę można rozszerzyć na informacyjne sygnały optyczne:

$$\Delta\omega_x \cdot \Delta x \geqslant 1 \quad \text{i} \quad \Delta\omega_y \cdot \Delta y \geqslant 1.$$

komórka informacyjna pola świetlnego

Zgodnie z zasadą nieokreśloności sygnałów optycznych w wypadku monochromatycznego pola świetlnego informacyjną czterowymiarową przestrzeń o współrzędnych ω_x, x, ω_y, y można podzielić na czterowymiarowe komórki o skończonej objętości $\Delta\omega_x \cdot \Delta x \cdot \Delta\omega_y \cdot \Delta y = 1$. Kształt takiej komórki informacyjnej można wybrać dowolnie, zachowując jedynie jej objętość. Rysunki 12a, b, c ukazują różne możliwości wyboru komórek w przestrzeni dwuwymiarowej ω_x, x. Każdej takiej komórce odpowiada niezależny pomiar amplitudy i fazy pola świetlnego sygnału. Zatem struktura informacyjna sygnału optycznego jest dyskretna. Analiza problemu pomiaru pola świetlnego z punktu widzenia teorii informacji prowadzi do wniosku, że całkowity pomiar optyczny, obejmujący pomiar rozkładu amplitudy i fazy przedmiotu świecącego, jest możliwy jedynie przy oświetleniu spójnym, natomiast przy oświetleniu przestrzennie niespójnym



Rys. 12. Diagram informacyjny: a) komórki elementarne Gabora, b) komórki elementarne Fouriera, c) komórki elementarne Shannona

informację o rozkładzie fazy traci się bezpowrotnie. Liczba niezależnych parametrów (wyników pomiarów) potrzebnych do opisania pola świetlnego w źrenicy wyjściowej układu optycznego równa się N przy oświetleniu spójnym i N^2 przy oświetleniu niespójnym; N jest liczbą Shannona, czyli liczbą informacyjnych przestrzennych stopni swobody:

$$N = N_x N_y \approx 2\Delta S_x \cdot \omega_{x\,\max} \cdot \Delta S_y \cdot \omega_{y\,\max},$$

gdzie ΔS_x i ΔS_y — wymiary liniowe obrazu sygnału w kierunku osi x i y, a $\omega_{x\,\max}$, $\omega_{y\,\max}$ — odpowiednie

szerokości pasma częstości przestrzennych przepuszczane przez rzeczywisty układ optyczny (czynnik 2 odpowiada dwóm niezależnym stanom polaryzacji światła). W wypadku niemonochromatycznego pola świetlnego liczbę tę należy uzupełnić dodatkowym czynnikiem: $N_t \approx \Delta\omega_t \cdot T$, gdzie T — czas obserwacji. Im mniejsza jest liczba parametrów potrzebnych do opisania sygnału optycznego przechodzącego przez tor optyczny, tym bardziej koherentny jest układ.

Perspektywy optyki fourierowskiej

Przedstawione wyżej podejście otwiera nowe perspektywy przed przetwarzaniem sygnałów optycznych. Jedną z możliwości jest tzw. przelewanie częstości przestrzennych. Całkowita liczba informacyjnych stopni swobody danego sygnału i toru optycznego jest niezmiennikiem, ale przy jej zachowaniu można np. zwiększyć zdolność rozdzielczą obrazu w kierunku x, a zmniejszyć w kierunku y, „przelewając" częstości przestrzenne z kierunku y na kierunek x. Ta możliwość będzie wyzyskana zapewne w telewizji holograficznej, dla której trudną pod względem technicznym sprawą jest przesyłanie tak wielkiej liczby informacji, jaką niesie hologram. Ponieważ aparat wzrokowy człowieka wymaga większej zdolności rozdzielczej w poziomie niż w pionie, można uzyskać dobre dla oka obrazy przez „przelewanie" częstości przestrzennych sygnałów o zmniejszonej liczbie informacyjnych stopni swobody.

przelewanie częstości przestrzennych

W dalszej przyszłości należy się spodziewać wykorzystania — jako nośnika informacji — pola elektromagnetycznego o jeszcze krótszej fali — nadfioletu oraz miękkiego promieniowania rentgenowskiego, czyli wyzyskanie tego zakresu widma fal elektromagnetycznych, w którym własności falowe promieniowania jeszcze są wyraźne. Przydatność optyki fourierowskiej będzie wówczas niewątpliwa.

A. Kalestyński *Co to są komputery optyczne*, w „Człowiek i nauka" 1977; J. R. Meyer-Arendt *Wstęp do optyki*, Warszawa 1977; A. Piekara *Nowe oblicze optyki*, Warszawa 1976.

KRIOFIZYKA

Nadpłynność

Eugeniusz Trojnar

Ciecz kwantowa

W ciekłym helu w temperaturach bliskich zera bezwzględnego obserwuje się niezwykłe zjawiska, trudne do wytłumaczenia na gruncie fizyki klasycznej. Zjawiska te bowiem są przejawem działania praw kwantowych, a więc tych praw, które rządzą światem atomów i cząsteczek. W ciekłym helu prawa kwantowe przejawiają się w skali makroskopowej, dlatego taką ciecz nazywamy cieczą kwantową.

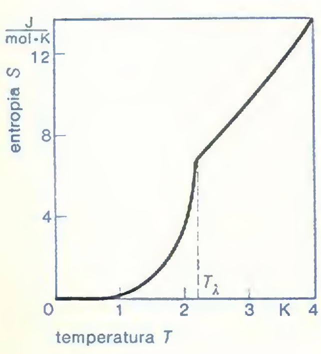
ciecz kwantowa

W ciałach o rozmiarach makroskopowych zjawiska kwantowe są przeważnie zamaskowane chaotycznymi ruchami cieplnymi atomów czy cząsteczek. W niskich temperaturach, gdy zanika bezładny ruch cieplny może dojść do wewnętrznego uporządkowania substancji i wtedy mogą się ujawnić zjawiska kwantowe. Nadpłynność jest właśnie wynikiem wysokiego stopnia uporządkowania wewnętrznego w ciekłym helu. Uporządkowanie to dotyczy ruchów atomów helu, a nie ich położeń. Porządkowanie położeń atomów w przestrzeni — to krystalizacja, a nadpłynny hel pod względem uporządkowania przestrzennego nie różni się od innych cieczy, o czym świadczą badania rentgenograficzne.

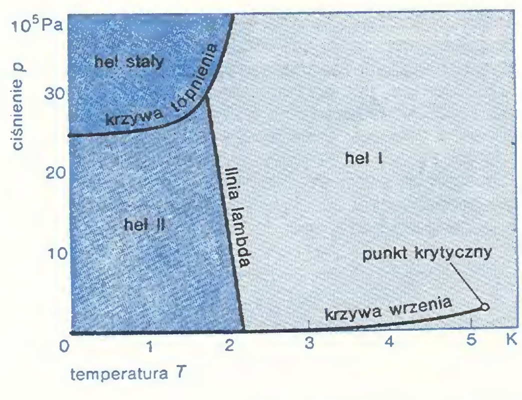
Miarą uporządkowania układu jest jedna z funkcji termodynamicznych, zwana entropią, przy czym im wyższy stopień uporządkowania układu, tym mniejsza jest jego entropia. Po przejściu ciekłego helu w stan nadpłynny jego entropia raptownie maleje, co dowodzi wewnętrznego podporządkowania się (rys. 1).

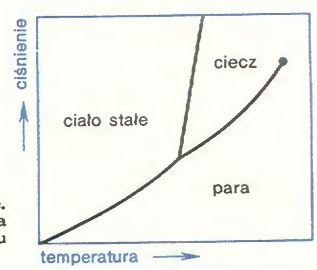
Hel jest jedyną w przyrodzie substancją nie zamarzającą pod normalnym ciśnieniem, aż do temperatury zera bezwzględnego. Z punktu widzenia fizyki klasycznej w temperaturze zera bezwzględnego każda substancja powinna znajdować się w stałym stanie skupienia, gdyż wszystkie cząsteczki powinny być w zupełnym bezruchu. Zgodnie z prawami fizyki kwantowej, całkowita utrata energii ruchu cząsteczek nie jest możliwa; nawet w temperaturze zera bezwzględnego cząsteczki wykonują pewien ruch drgający, tzw. drgania zerowe, a jeśli energia tych drgań jest wystarczająco duża, to ciało nie krystalizuje. A właś-

drgania zerowe w helu



Rys. 1. Zależność entropii ciekłego helu pod ciśnieniem pary nasyconej od temperatury





Rys. 2. Wykres stanu ⁴He. U dołu zamieszczono dla porównania wykres stanu normalnej substancji

nie w helu energia drgań zerowych jest duża i jednocześnie hel — jako gaz szlachetny — odznacza się słabym oddziaływaniem międzyatomowym. Oba te czynniki nie sprzyjają krystalizacji. Aby zestalić hel, należy zastosować zwiększone ciśnienie, czyli bardziej zbliżyć do siebie poszczególne atomy. Wartość ciśnienia krystalizacji zależy od temperatury, co ilustruje zamieszczony na rys. 2 wykres stanu (wykres fazowy). Górna krzywa na tym wykresie jest krzywą krystalizacji (topnienia), czyli krzywą równowagi fazy stałej i ciekłej. Dolna krzywa jest krzywą zależności prężności nasyconej pary helu od temperatury, czyli krzywą równowagi fazy ciekłej i gazowej — krzywą skraplania (wrzenia). Krzywa skraplania kończy się w punkcie krytycznym $T_{kr} = 5{,}2$ K, $p_{kr} = 2{,}16 \cdot 10^5$ Pa. Dla normalnych substancji obie te krzywe łączą się z sobą w punkcie potrójnym, gdzie wszystkie trzy fazy — stała, ciekła i gazowa — są w równowadze termodynamicznej.

wykres stanu helu

Hel nie ma punktu potrójnego, ponieważ krzywa wrzenia nie przecina nigdzie krzywej krystalizacji. Nie można więc helu zestalić przez obniżanie jego temperatury pod ciśnieniem pary nasyconej. Aby hel zestalić w pobliżu $T = 0$ K, należy zwiększyć ciśnienie do $24{,}5 \cdot 10^5$ Pa.

Przemiana lambda

Krzywą krystalizacji helu z krzywą skraplania łączy na wykresie fazowym tzw. linia lambda, rozdzielająca obszar cieczy na dwa obszary fazowe: obszar helu I (He I) i obszar helu II (He II). Pod względem własności fizycznych hel I jest normalną cieczą, niczym istotnym nie wyróżniającą się spośród innych cieczy. Nieco bardziej tylko, niż inne ciecze, przypomina gaz, gdyż ma stosunkowo małą lepkość i gęstość.

hel I

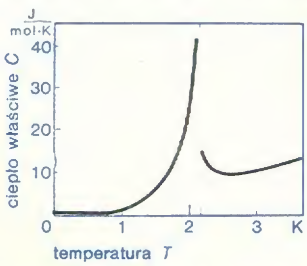
Hel II jest cieczą nadpłynną. Przemiana fazowa zmieniająca hel I w hel II, zwana przemianą lambda, zachodzi pod ciśnieniem pary nasyconej w temperaturze $T_\lambda = 2{,}17$ K. Ciśnienie odpowiadające tej temperaturze wynosi $0{,}05 \cdot 10^5$ Pa. Pod zwiększonym ciśnieniem temperatura przemiany w fazę nadpłynną jest niższa i w pobliżu krzywej krzepnięcia wynosi 1,76 K. Wartości ciśnień i odpowiadających im temperatur przemiany lambda wyznaczają na wykresie fazowym linię lambda.

hel II

Przemiana lambda jest innym rodzajem przemiany fazowej (przemianą fazową II rodzaju) niż np. skraplanie czy krzepnięcie. Nie towarzyszy jej oddawanie lub pobieranie ciepła, nie ma więc „utajonego" ciepła przemiany. Linia lambda nie jest też krzywą równowagi dwu faz, jak krzywa topnienia lub skraplania, gdyż hel II nie może istnieć w równowadze termodynamicznej z helem I; cała ciecz — w zależności od temperatury i ciśnienia — jest albo helem I albo helem II.

linia lambda

Pierwszym zaobserwowanym faktem doświadczalnym wskazującym na to, że w ciekłym helu dokonuje się jakaś przemiana, była zmiana temperaturowej zależności przenikalności elektrycznej, którą to zmianę wykrył w 1927 r. Polak, Mieczysław Wolfke, pracujący wtedy w Lejdzie. Przypuszczenie o zachodzącej przemianie potwierdziły pomiary ciepła właściwego, przeprowadzone przez W. H. Keesoma. Krzywa zależności ciepła właściwego (rys. 3) od temperatury w pobliżu $T_\lambda = 2{,}17$ K rozrywa się i wygina ku górze tak, że jej kształt przypomina grecką literę λ (lambda). Stąd pochodzi nazwa przemiany fazowej zachodzącej w ciekłym helu.



Rys. 3. Zależność ciepła właściwego ciekłego helu od temperatury

Fizycy początkowo nie docenili znaczenia tego odkrycia i nie zwrócili należytej uwagi na zmiany zachodzące w ciekłym helu. Bez echa przeszło również spostrzeżenie J. C. McLennana i jego współpracowników, że gdy temperatura helu spada poniżej 2,17 K, ciecz nagle przestaje wrzeć, tzn. pęcherzyki pary już się nie tworzą w jej objętości. Dopiero w 1938 r., a więc sześć lat po odkryciu przemiany lambda i trzydzieści lat po pierwszym skropleniu helu przez H. Kamerlingh-Onnesa, P. L. Kapica w Moskwie oraz J. F. Allen i A. D. Misener w Cambridge niezależnie od siebie odkryli zadziwiającą zdolność helu II do przepływania niemal bez tarcia przez bardzo wąskie szczeliny szerokości rzędu 10^{-5} cm. Tę właściwość helu II Kapica zaproponował nazwać nadpłynnością.

nadpłynność helu II

Podstawowe własności helu II

Objętość zwykłej cieczy przepływającej w jednostce czasu przez rurkę lub szczelinę jest odwrotnie proporcjonalna do lepkości cieczy, czyli do jej tarcia wewnętrznego. Lepkość wody wynosi około 10^{-3} Ns/m², lepkość helu I około 10^{-6} Ns/m², a lepkość helu II mierzona tą metodą może wynosić nawet 10^{-12} Ns/m², a więc milion razy mniej niż lepkość helu I. Dla cieczy zwyczajnych słuszny jest wzór Poiseuille'a:

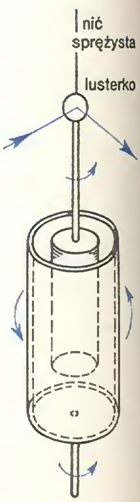
lepkość helu II

$$\frac{V}{t} = \frac{\pi a^2 \cdot \Delta p}{8\eta l}, \qquad (1)$$

gdzie V jest objętością cieczy przepływającej w czasie t przez rurkę o promieniu a i długości l pod wpływem różnicy ciśnień Δp; η jest współczynnikiem lepkości.

Dla helu II wyniki pomiarów lepkości metodą przepływu są niejednoznaczne, gdyż zależą od średnicy kapilary: im większa kapilara, tym mniejszą wartość lepkości otrzymuje się z pomiarów. Poza tym prędkość przepływu przez wąskie kapilary prawie nie zależy od różnicy ciśnień na obu jej końcach, chociaż powinna być do niej proporcjonalna.

Można zastosować inny sposób pomiaru lepkości: Jeżeli zapełni się cieczą przestrzeń między dwoma współosiowymi walcami i będzie się obracać walec zewnętrzny, to ciecz pociągana obracającym się walcem będzie oddziaływać na walec wewnętrzny i starać się go obrócić w tym samym kierunku (rys. 4). Moment obrotowy działający na walec wewnętrzny jest miarą lepkości cieczy. Lepkość helu II mierzona tą metodą niewiele różni się od lepkości helu I, a poniżej 1 K jest nawet od niej większa. Tak więc dwie wypróbowane metody pomiaru lepkości zastosowane do normalnych cieczy dają takie same wyniki, a zastosowane do helu II — wyniki bardzo się od siebie różniące. Pojęcie lepkości, dotyczące zwykłych cieczy, w odniesieniu do cieczy kwantowej traci swój zwykły sens.



Rys. 4. Układ do pomiaru lepkości metodą obracającego się walca

Drugą osobliwością helu II jest jego olbrzymie przewodnictwo cieplne. Przewodność cieplną $\varkappa$ definiujemy zwykle jako współczynnik proporcjonalności między strumieniem przewodzonego ciepła q i gradientem temperatury $\Delta T/\Delta l$, gdzie ΔT oznacza różnicę temperatur między punktami oddalonymi od siebie o Δl. Równanie przewodnictwa cieplnego można zapisać w postaci:

przewodnictwo cieplne helu II

$$q = -\varkappa \frac{\Delta T}{\Delta l}.$$

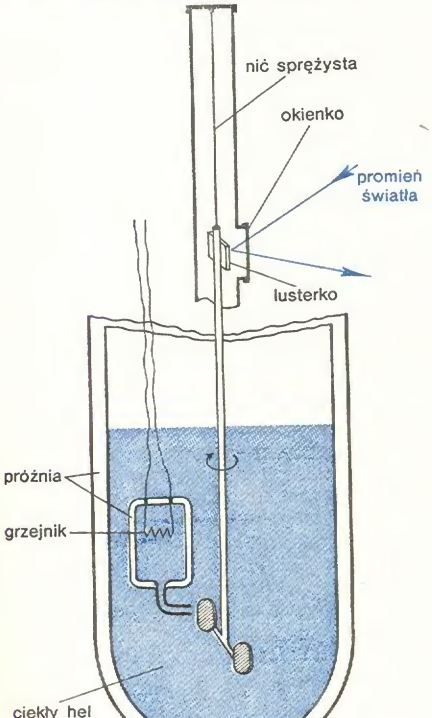
Znak „minus" pochodzi stąd, że strumień ciepła przepływa w kierunku przeciwnym do kierunku gradientu temperatury (gradient jest kierowany umownie w stronę wzrostu temperatury). W pewnych warunkach $\varkappa$ helu II jest tysiąckrotnie większe niż $\varkappa$ miedzi i wiele milionów razy większe niż $\varkappa$ helu I. W kąpieli nadpłynnego helu nie da się wytworzyć stałej różnicy tempera-

tur. Hel II nie wrze, ponieważ do wytworzenia się pęcherzyków pary wewnątrz cieczy potrzebne są niewielkie lokalne przegrzania (fluktuacje temperatury). Hel II paruje tylko z powierzchni swobodnej.

Podobnie jak pomiary lepkości, pomiary współczynnika przewodnictwa cieplnego helu II nie dają jednoznacznych wyników. Wartości $\varkappa$ zależą od geometrycznych szczegółów aparatury pomiarowej, poza tym nie ma proporcjonalności między strumieniem ciepła i gradientem temperatury. Największe wartości $\varkappa$ uzyskuje się wtedy, gdy na końcach bardzo wąskiej kapilary wytworzy się niewielka różnica temperatur. Przy przepływie helu II przez takie wąskie kapilary lub szczeliny uzyskuje się również najmniejsze wartości współczynnika lepkości. Nasuwa to myśl o związku między tymi dwoma zjawiskami. Najwidoczniej transport ciepła w helu II ma charakter konwekcyjny i jest związany z bezlepkościowym przepływem cieczy. W przypadku helu II nie można więc mówić o współczynniku przewodnictwa cieplnego w zwykłym znaczeniu.

doświadczenie Kapicy

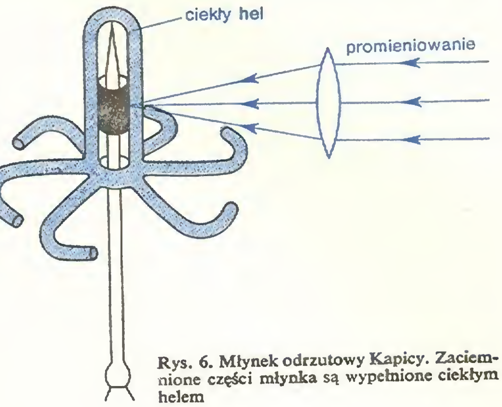
Aby wyjaśnić proces przenoszenia ciepła w helu II, przeprowadzono szereg doświadczeń. W jednym z nich, wykonanym przez Kapicę, małe naczynie Dewara w kształcie kolby z wąską szyjką i z grzejnikiem elektrycznym w środku zostało zanurzone w helu II. Naprzeciw wylotu szyjki umieszczono niewielką tarczę zawieszoną wraz z równoważącym ją ciężarkiem na wadze skręceń (rys. 5). Włączenie grzejnika do obwodu prądu powodowało odchylenie tarczy. Po wyłączeniu grzejnika tarcza wracała w poprzednie położenie. Tarcza zachowywała się więc tak, jakby była odpychana przez wypływający z naczynia strumień



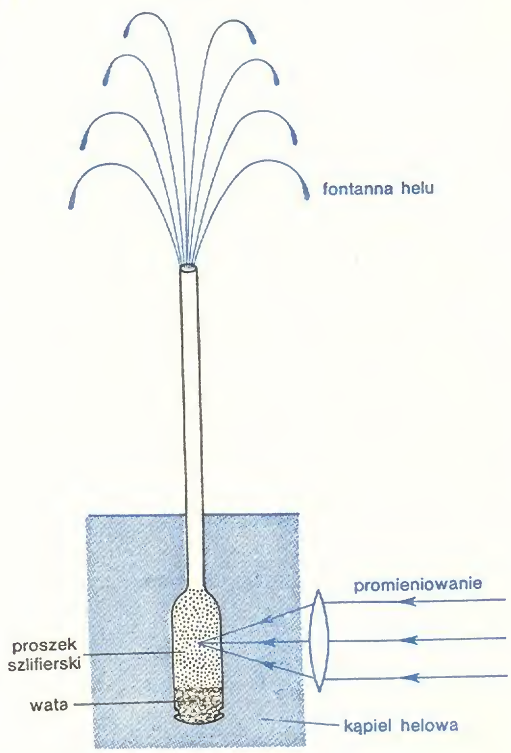
Rys. 5. Układ Kapicy do demonstracji przepływu składnika normalnego helu II

cieczy. Cieczy w naczyniu jednak nie ubywało, a Kapica na próżno starał się wykryć strumień płynący w przeciwną stronę, tj. wpływający do naczynia (np. po jego ściankach). W cieczy nadpłynnej mogą istnieć zatem dwa rodzaje ruchów: jeden ruch, jak w normalnej cieczy lepkiej, związany z przekazywaniem pędu ciałom w niej zanurzonym, i drugi ruch odbywający się bez wymiany pędu między cieczą i ciałem stałym. Oczywiście, gdyby naczynie nie było unieruchomione, to mogłoby się poruszać w kierunku przeciwnym do wypływającego strumienia, jak silnik odrzutowy.

W innym tego typu doświadczeniu Kapica wykorzystał siłę odrzutu wypływającego strumienia do uruchomienia urządzenia przypominającego po-



Rys. 6. Młynek odrzutowy Kapicy. Zaciemnione części młynka są wypełnione ciekłym helem



Rys. 7. Zjawisko fontannowe w helu II

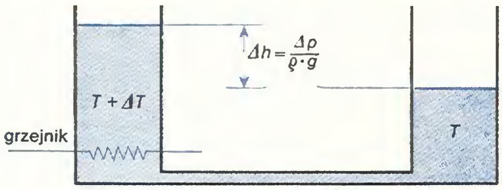
wszechnie znany młynek Segnera (rys. 6). Zgięte ramiona tego urządzenia są rurkami prowadzącymi do przestrzeni zawartej między podwójnymi ściankami cylindrycznego naczynka ze szkła. Młynek wisi na pionowej podpórce, wokół której może się swobodnie obracać. Całość jest zanurzona w nadpłynnym helu. Jeśli na ścianki młynka, wyczernione sadzą od strony wewnętrznej, pada smuga światła, to młynek się obraca. Ruch obrotowy młynka może trwać dowolnie długo, gdyż — podobnie jak w poprzednio opisanym doświadczeniu — cieczy nie ubywa między ściankami młynka. Młynek zatrzyma się dopiero po zaprzestaniu naświetlania, czyli przerwaniu ogrzewania cieczy znajdującej się między ściankami młynka. Wypływający z naczynia strumień powstaje więc w nagrzewanej cieczy nie powodując jej ubytku.

zjawisko fontannowe

Podczas innych doświadczeń, przeprowadzonych przez grupę badaczy w Cambridge, okazało się, że jeśli do naczynia, w którym ogrzewano hel II, prowadziła cienka kapilara lub wąskie szczeliny, to ciśnienie w tym naczyniu wzrastało. W najbardziej widowiskowym wariancie tego doświadczenia ciekły hel pod wpływem zwiększonego ciśnienia wytryskiwał ku górze w postaci fontanny (rys. 7). Przyrząd użyty w tym doświadczeniu składał się ze szklanej rurki napełnionej proszkiem szlifierskim i zatkanej kłębkiem waty. Rurka była zanurzona w helu II tak, że jej górny koniec wystawał nieco ponad poziom cieczy. Gdy na filtr proszku szlifierskiego skierowano smugę światła, z górnego końca rurki wytrysnął strumień cieczy.

zjawisko termo-mechaniczne

Mniej widowiskowa, lecz bardziej przejrzysta odmiana tego doświadczenia pokazana jest na rys. 8. Dwa naczynia zawierające hel II połączone są cienką



Rys. 8. Szkic ilustrujący zjawisko termomechaniczne i mechanokaloryczne

kapilarą. Ogrzewanie cieczy w jednym z tych naczyń powoduje podniesienie się jej poziomu w tym naczyniu, natomiast obniżenie — w drugim naczyniu. Różnica ciśnień Δp odpowiadająca tej różnicy poziomów jest proporcjonalna do wytworzonej różnicy temperatur ΔT. Zjawisko to otrzymało nazwę zjawiska termomechanicznego. Istnieje również zjawisko odwrotne, zwane zjawiskiem mechanokalorycznym: jeżeli w naczyniach połączonych kapilarą wytworzy się mechanicznie różnicę ciśnień, to ta różnica wywoła różnicę temperatur. Zależność między Δp i ΔT jest taka sama jak poprzednio.

zjawisko mechano-kaloryczne

Zakładając pełną odwracalność tych zjawisk i traktując układ naczyń jako maszynę cieplną, w której dostarczone ciepło zamienia się w całości na pracę podnoszenia cieczy na wyższy poziom, London znalazł następującą zależność:

$$\Delta p = \varrho S \Delta T,$$

gdzie ϱ oznacza gęstość helu II, zaś S jest entropią właściwą cieczy w naczyniu o wyższej temperaturze. Wyniki doświadczeń potwierdziły słuszność tego wzoru.

Model dwupłynowy

Opisane powyżej zjawiska można poglądowo wytłumaczyć posługując się tzw. modelem dwupłynowym wprowadzonym dla helu II przez L. Tiszę. W tym modelu zakłada się, że hel II jest mieszaniną dwu płynów o różnych własnościach fizycznych. Jeden ze składników helu to ciecz normalna podobnie jak hel I; drugi składnik jest pozbawiony lepkości i ma równą zeru entropię. Procentowa zawartość składnika normalnego zależy od temperatury. W temperaturze zera bezwzględnego cała ciecz jest nadpłynna. W temperaturze wyższej od zera pojawia się składnik normalny. W miarę wzrostu temperatury udział składnika normalnego zwiększa się coraz bardziej i w temperaturze lambda cała ciecz staje się normalna.

wyjaśnienie doświadczenia Kapicy

Posługując się tym modelem możemy wyjaśnić wynik doświadczenia Kapicy następująco. Ciecz ogrzewana wewnątrz zanurzonej kolby ma temperaturę wyższą niż otaczająca kąpiel, a więc ma większe stężenie składnika normalnego. Hel II stara się wyrównać wszędzie swój skład (podobnie jak wyrównuje się w naczyniu stężenie roztworu), dlatego do wnętrza kolby wpływa składnik nadpłynny nie oddziałujący ze ściankami naczynia ani tarczą wagi skręceń, natomiast z kolby wypływa strumień składnika normalnego i zachowuje się jak zwykła ciecz. Jednak, dopóki działa grzejnik, do wyrównania składu cieczy wewnątrz i na zewnątrz kolby nie dochodzi, gdyż pod wpływem ciepła wydzielonego w kolbie dokonuje się tam ciągła zamiana składnika nadpłynnego na normalny. Jeśli wlot do naczynia zanurzonego w helu II jest bardzo wąski lub zatkany porowatym korkiem (np. z proszku szlifierskiego), a wewnątrz naczynia temperatura jest wyższa niż w otaczającej kąpieli helowej, to składnik nadpłynny może łatwo do naczynia wpływać, natomiast składnik normalny przez taką przegrodę przedostaje się z trudnością. W rezultacie ciśnienie wewnątrz naczynia wzrasta. Nasuwa się tu analogia z ciśnieniem osmotycznym wywołanym różnicą stężeń roztworów (np. soli) rozdzielonych przegrodą przepuszczalną dla rozpuszczalnika, lecz nieprzepuszczalną dla substancji rozpuszczonej.

Jeżeli po obu stronach porowatej przegrody lub w dwu naczyniach połączonych cienką kapilarą wytworzymy w helu II różnicę ciśnień, to pod jej wpływem przez przegrodę lub kapilarę będzie przepływał niemal wyłącznie składnik nadpłynny. Ponieważ składnik nadpłynny nie przenosi z sobą ciepła (ściślej mówiąc, nie przenosi entropii), temperatura w tej części naczynia, do której wpływa składnik nadpłynny, obniża się.

Na gruncie modelu dwupłynowego zrozumiała staje się także rozbieżność wyników pomiarów lepkości dwiema różnymi metodami. Przy przepływie cieczy przez kapilarę pod wpływem różnicy ciśnień zasadniczą rolę odgrywa pozbawiony lepkości składnik nadpłynny, podczas gdy obracający się walec pociąga za sobą lepki składnik normalny, a ten z kolei oddziałuje na walec wewnętrzny.

Model dwupłynowy tłumaczy także doskonałe przewodnictwo cieplne helu II: nawet bardzo mała różnica temperatur wytworzona w helu II spowoduje natychmiastowy przepływ składnika nadpłynnego w kierunku wyższej temperatury i odpływ w kierunku przeciwnym składnika normalnego; w rezultacie temperatura się wyrówna.

stężenie składnika normalnego

Gęstość helu II jest sumą gęstości obu jego składników: $\varrho = \varrho_n + \varrho_s$. Wskaźnik n odnosi się do gęstości składnika normalnego, wskaźnik s — do gęstości składnika nadpłynnego. Stosunek ϱ_n/ϱ, czyli stężenie składnika normalnego, jest — jak już wspominaliśmy — rosnącą funkcją temperatury. Przebieg tej zależności wyznaczył doświadczalnie E. Andronikaszwili w 1946 r. Przyrząd, którym się posługiwał składał się ze stosu cienkich krążków mikowych, nasadzonych w niewielkich odstępach na wspólną oś (rys. 9). Układ krążków był zawieszony na sprężystej nici, mógł więc wykonywać drgania skrętne wokół swej osi. Umieszczone na osi lusterko umożliwiało obserwację ruchu drgającego.

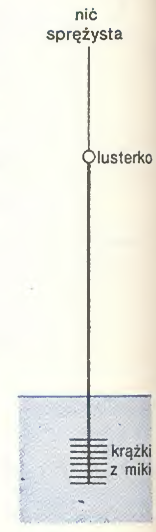
Okres τ_0 drgań skrętnych takiego układu krążków zależy od jego momentu bezwładności B_0 i stałej skręcenia nici k:

$$\tau_0 = 2\pi \sqrt{\frac{B_0}{k}}.$$

Po zanurzeniu w helu II układ krążków podczas drgań porywa z sobą składnik normalny, który więźnie między krążkami, i moment bezwładności układu zwiększa się do $B = B_0 + B_{He}$ gdzie B_{He} jest momentem bezwładności cieczy uwięzionej między krążkami i poruszającej się wraz z nimi, zależnym od jej gęstości. Okres drgań wyniesie teraz:

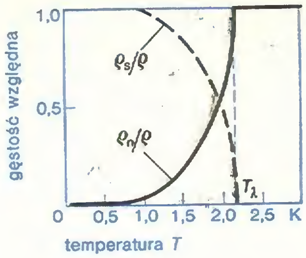
$$\tau = 2\pi \sqrt{\frac{B_0 + B_{He}}{k}}.$$

Składnik nadpłynny, jako pozbawiony lepkości, nie jest porywany przez poruszające się krążki. Mierząc



Rys. 9. Układ Andronikaszwiliego do pomiaru stężenia składnika normalnego helu II

więc okresy drgań układu w zależności od temperatury można znaleźć funkcję $\varrho_n/\varrho = f(T)$. Wykres tej funkcji przedstawia rys. 10.

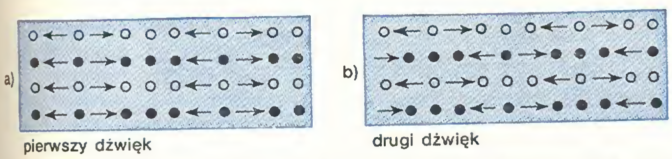


Rys. 10. Zależność stężenia składnika normalnego helu II od temperatury; $(\varrho_s+\varrho_n)/\varrho = 1$

Fale temperaturowe, czyli drugi dźwięk

Model dwupłynowy nie tylko wyjaśniał większość znanych w tym czasie faktów doświadczalnych związanych z helem II, ale także pozwalał przewidywać nowe zjawiska. Na podstawie tego modelu Tisza przewidział możliwość występowania w helu II fal temperaturowych, zwanych później drugim dźwiękiem.

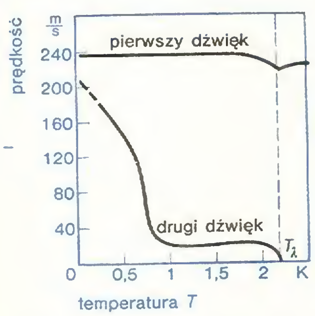
W nadpłynnym helu, podobnie jak w każdej innej cieczy, mogą rozprzestrzeniać się zwykle fale dźwiękowe, polegające na drganiach gęstości cieczy. Oba składniki helu II poruszają się wtedy jak jedna całość drgając w zgodnej fazie. Prowadzi to do kolejnych zgęszczeń i rozrzedzeń cieczy rozprzestrzeniających się ruchem falowym. Prócz tego zwykłego (pierwszego) dźwięku w nadpłynnym helu możliwy jest jeszcze inny ruch falowy, kiedy oba składniki drgają w przeciwnych kierunkach, tzn. z przesunięciem w fazie o kąt



Rys. 11. Wyjaśnienie powstawania pierwszego (a) i drugiego (b) dźwięku w helu II na podstawie modelu dwupłynowego. Zgęszczenie składnika nadpłynnego oznaczono jasnymi krążkami, zgęszczenie składnika normalnego — krążkami zaczernionymi

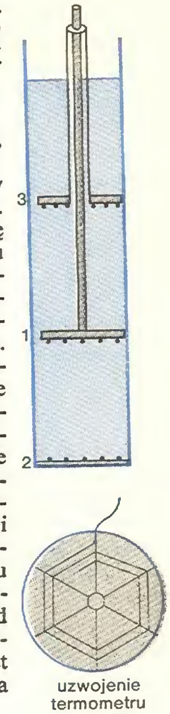
π (rys. 11). Nie są to więc drgania gęstości cieczy, lecz drgania stężenia obu jej składników. Inaczej mówiąc, w obranym dowolnie miejscu w cieczy zmienia się okresowo wartość stosunku ϱ_s/ϱ_n tak, że suma $\varrho_s+\varrho_n = \varrho$ pozostaje stała. Ponieważ zmiana stężenia jednego ze składników jest równoważna zmianie temperatury, takie zaburzenie rozprzestrzenia się jako fala temperaturowa. Prędkość drugiego dźwięku jest funkcją temperatury i jest mniejsza od prędkości pierwszego dźwięku (rys. 12). Powyżej punktu lambda,

prędkość drugiego dźwięku



Rys. 12. Zależność prędkości pierwszego i drugiego dźwięku od temperatury

gdy nie ma już składnika nadpłynnego, fala drugiego dźwięku istnieć nie może. Fala drugiego dźwięku nie mogłaby również istnieć w temperaturze dokładnie równej zeru bezwzględnemu, gdyż wtedy ciekły hel musiałby być w stanie niewzbudzonym, każde zaś



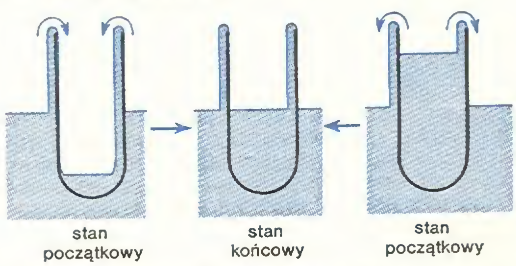
Rys. 13. Schemat układu Pieszkowa do pomiaru prędkości drugiego dźwięku. Termometr *2* na dnie rury rezonansowej służy do ustalenia położenia grzejnika *3*. Aby w rurze mogła powstać fala stojąca, odległość grzejnika od dna rury musi być równa wielokrotności połowy długości fali

wzbudzenie oznacza, że temperatura różni się od zera, chociażby nawet o niewielką wartość.

Prędkość drugiego dźwięku w funkcji temperatury po raz pierwszy zmierzył W. Pieszkow (1944). Omówimy tutaj późniejszą, udoskonaloną wersję jego doświadczenia. Jako źródło drugiego dźwięku służył grzejnik elektryczny o bardzo małej bezwładności cieplnej, zasilany sinusoidalnym prądem zmiennym. Odbiornikiem (detektorem) fal temperaturowych był termometr oporowy zasilany prądem stałym. Zmiany temperatury powodowały zmiany oporu termometru, a więc drgania spadku napięcia na nim, te zaś łatwo można wykryć i określić ich częstość. Pulsacje napięcia na termometrze miały dwukrotnie wyższą częstość niż zmiany prądu w grzejniku, można je więc było odróżnić od elektromagnetycznego tła pochodzącego od grzejnika. Zarówno grzejnik jak i termometr nawinięte były w postaci płaskich spirali z bardzo cienkiego drutu i umieszczone w rurze rezonansowej (rys. 13). Przy odpowiednim położeniu grzejnika drugi dźwięk odbijał się od dna rury i tworzył falę stojącą. Zmieniając odległość termometru od dna rury można było znaleźć położenia węzłów i strzałek tej fali, a więc określić jej długość. Jeśli znana jest długość fali λ i częstość drgań temperatury ν, można obliczyć prędkość drugiego dźwięku: $u_2 = \lambda\nu$.

Pełzająca warstwa i prędkość krytyczna

Warstwa cieczy pełzająca ruchem nadpłynnym po powierzchni ciała stałego — to jeszcze jedna osobliwość związana z helem II. Warstwa taka, o grubości rzędu 10^{-6} cm, tj. około 50 średnic atomowych, pokrywa każdą powierzchnię ciała stałego stykającego się z helem II i porusza się po niej bez tarcia pod wpływem siły grawitacji albo gradientu temperatury. Po ściankach kriostatu z helem II wspina się więc do góry w kierunku wzrastającej temperatury cienka nadpłynna warstwa i stopniowo wyparowuje stykając się z coraz cieplejszymi ściankami.



Rys. 14. Pełzanie warstwy helu II po ściankach naczynia

pełzanie warstwy helu

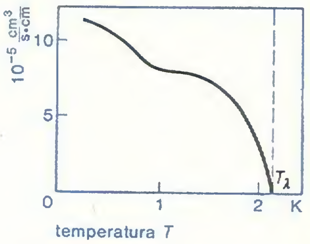
Jeżeli do kąpieli z helu II wstawimy pustą zlewkę tak, aby jej dno znalazło się poniżej poziomu otaczającej ją cieczy (rys. 14), to ciecz pełzając po ściankach zlewki będzie ją stopniowo napełniać aż do chwili, gdy wyrównują się poziomy cieczy wewnątrz i na zewnątrz zlewki. Jeśli teraz podniesiemy zlewkę, to hel będzie z niej wypełzać po ściankach aż do kolejnego wyrównania poziomów. Jeśli zaś wyjmiemy zlewkę z helem II z kąpieli, to z jej dna będą kapać krople cieczy i zlewka opróżni się. Tak więc He II w polu grawitacyjnym zawsze stara się zająć możliwie najniższy poziom.

Objętość cieczy (w cm^3), którą przenosi pełzająca warstwa w ciągu sekundy przez 1 cm długości obwodu zlewki, jest niezależna od różnicy poziomów cieczy, podobnie jak prędkość przepływu składnika nadpłynnego w cienkich kapilarach jest niezależna od różnicy ciśnień. Nasuwa się więc przypuszczenie, że

prędkość krytyczna

istnieje pewna krytyczna prędkość przepływu ograniczająca ilość przepływającej cieczy. Przepływ z prędkością większą niż krytyczna utraciłby swój beztarciowy charakter.

Rys. 15 przedstawia temperaturową zależność ilości przenoszonej cieczy przez 1 cm długości obwodu zlewki w ciągu sekundy. W przedziale temperatur od 1 do



Rys. 15. Prędkość pełzania warstwy helu II w zależności od temperatury. Na osi rzędnych naniesiona jest objętość cieczy przeniesionej w ciągu 1 s przez odcinek o długości 1 cm prostopadły do kierunku ruchu warstwy

1,5 K ilość ta jest prawie niezależna od temperatury i wynosi ok. $7{,}5\cdot10^{-5}$ cm³/(cm·s); przy grubości warstwy 10^{-6} cm prędkość krytyczna wynosi 75 cm/s. Jest to wielkość tego samego rzędu, co wyniki pomiarów prędkości przepływu przez cienkie kapilary.

Wzbudzenia elementarne w helu II

Posługując się modelem dwupłynowym do wyjaśnienia własności helu II nie powinniśmy zapominać, że model ten — to tylko poglądowy sposób opisu zjawisk występujących w cieczy kwantowej. W rzeczywistości hel II nie składa się z dwu cieczy różniących się między sobą własnościami fizycznymi, lecz stanowi jeden układ złożony z identycznych atomów.

Jedną z zasad mechaniki kwantowej jest zasada nierozróżnialności jednakowych cząstek układu. Zgodnie z tą zasadą wszystkie atomy helu II zawarte w jednym naczyniu stanowią jakby jedną gigantyczną molekułę, w której ruch jednego atomu nie może być niezależny od innych atomów. Nadpłynność jest zjawiskiem kolektywnym, właściwym całemu zbiorowisku atomów; nie można mówić o nadpłynności poszczególnych atomów. Nie ma więc żadnych podstaw, aby niektórym atomom helu II przypisywać inne własności niż pozostałym. Cała ciecz powinna być rozpatrywana jako jedna całość.

Autorem takiego podejścia do zagadnienia nadpłynności helu II był L. D. Landau. Swą teorię nadpłynności Landau opracował w 1941 r., a więc prawie w tym samym czasie, kiedy ukazała się publikacja Tiszy dotycząca modelu dwupłynowego. Teoria Landaua przewidywała również możliwość występowania drugiego dźwięku w helu II, przy czym jej wyniki lepiej zgodziły się z późniejszymi danymi doświadczalnymi, niż oszacowanie Tiszy. A oto podstawowe założenia teorii Landaua:

teoria nadpłynności Landaua

W temperaturze zera bezwzględnego ciekły hel znajduje się w stanie podstawowym, tzn. ma najniższą energię. Jest to energia stanu podstawowego, czyli energia drgań zerowych. Podobnie jak w przypadku molekuły znajdującej się w stanie podstawowym, energia ta nie jest równa zeru. W temperaturach wyższych od zera ciecz przechodzi w jeden ze swoich stanów wzbudzonych. Oznacza to, że w cieczy pojawia się jakiś ruch, inny niż drgania zerowe. Jeśli temperatura nie jest zbyt wysoka, to ciecz w stanie wzbudzonym ma małą energię (mierzoną od poziomu energii stanu podstawowego). Małej energii odpowiadają proste rodzaje ruchów, np. lokalne zgęszczenia cieczy przemieszczające się w jej objętości, czyli fala dźwiękowa.

Każda fala dźwiękowa niesie z sobą określoną energię ε i pęd p związane zależnością:

$$\varepsilon = up, \tag{2}$$

gdzie u jest prędkością dźwięku. Fali dźwiękowej można formalnie przyporządkować pewną cząstkę

fonony

obdarzoną energią ε i pędem p. Cząstka taka pod nazwą fononu jest dobrze znana z teorii ciała stałego jako kwant drgań sieci krystalicznej (kwant dźwięku).

O tym, że kwanty niskoenergetycznych wzbudzeń w helu II można utożsamić z kwantami wzbudzeń elementarnych w sieci krystalicznej, świadczy jednakowa temperaturowa zależność „sieciowego" ciepła właściwego i ciepła właściwego helu II w pobliżu zera bezwzględnego. Ciepło właściwe helu II, podobnie jak ciepło właściwe sieci krystalicznej ciała stałego, jest proporcjonalne do trzeciej potęgi temperatury.

rotony

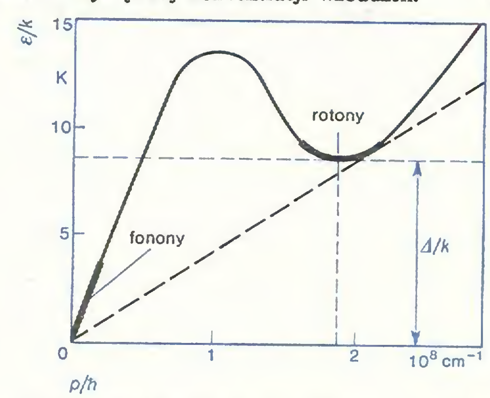
W temperaturach wyższych od 0,6 K ciepło właściwe C helu II odbiega od zależności $C \sim T^3$, co świadczy o tym, że istotną rolę zaczyna odgrywać inny typ wzbudzeń elementarnych. Te dodatkowe wzbudzenia Landau nazwał rotonami, wiążąc je z elementarnymi zawirowaniami w cieczy kwantowej. Rotony wcześniej nie były znane i po raz pierwszy wprowadzono je w związku z zagadnieniem nadpłynności. Zależność pomiędzy energią i pędem rotonu nie jest już zależnością liniową.

Funkcja charakteryzująca zależność energii od pędu nazywa się widmem energetycznym układu. Aby dopasować widmo energetyczne helu II do wyników pomiarów ciepła właściwego Landau zaproponował dla rotonu następującą zależność energii od pędu:

$$\varepsilon = \Delta + \frac{(p-p_0)^2}{2\mu}. \tag{3}$$

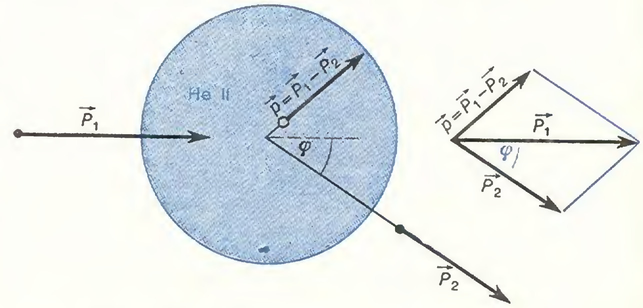
Δ oznacza tu najmniejszą wartość energii, jaką może mieć roton (wtedy jego pęd jest równy p_0), μ jest masą efektywną rotonu. Danym doświadczalnym odpowiadają następujące wartości $\Delta/k = 8{,}5$ K, $\mu = 0{,}16\, m_{He}$, $p_0/\hbar = 1{,}9\cdot10^8$ cm^{-1} (k — stała Boltzmanna, m_{He} — masa atomu He, $\hbar$ — stała Plancka dzielona przez 2π).

Rys. 16 przedstawia widmo wzbudzeń energetycznych He II. Pogrubionymi liniami oznaczono części widma o największej koncentracji wzbudzeń.



Rys. 16. Widmo energetyczne helu II. Zgrubienia na krzywej wskazują na najczęściej występujące wzbudzenia

Fonony i rotony (a także inne wzbudzenia występujące np. w ciele stałym) różnią się od cząstek ele-



Rys. 17. Zmiana pędu neutronu po rozproszeniu w helu II. $\vec{P}_1$, $\vec{P}_2$ pęd neutronu przed i po rozproszeniu, $\vec{p}$ pęd wytworzonego wzbudzenia

mentarnych takich jak swobodny elektron czy neutron tym, że są nierozłącznie związane z ośrodkiem, w którym powstają i w którym poruszają się: poza tym ośrodkiem wzbudzenia samodzielnie istnieć nie mogą. W odróżnieniu od „zwykłych" cząstek elementarnych wzbudzenia energetyczne nazywamy kwazicząstkami (niby-cząstkami).

kwazi-cząstki

W 1957 r. R. P. Feynman podał sposób doświadczalnego wyznaczenia widma wzbudzeń elementarnych w helu II. Zaproponował mianowicie, aby do tego celu wykorzystać rozpraszanie w ciekłym helu wiązki monoenergetycznych neutronów. Neutron przechodząc przez hel II może przekazać cieczy część energii i pędu, czyli może wytworzyć w cieczy wzbudzenie elementarne. W wyniku tego procesu neutron odchyli się o kąt φ od swego pierwotnego kierunku (rys. 17). Jeżeli E_1 i $\vec{P}_1$ oznaczają energię i pęd neutronu przed rozproszeniem, a E_2 i $\vec{P}_2$ — po rozproszeniu, to ze względu na prawo zachowania energii i pędu można napisać:

widmo energetyczne helu II

$$\left.\begin{aligned} \varepsilon &= E_1 - E_2, \\ \vec{p} &= \vec{P}_1 - \vec{P}_2, \end{aligned}\right\} \tag{4a}$$

czyli

$$\left.\begin{aligned} \varepsilon &= \frac{P_1^2}{2m} - \frac{P_2^2}{2m}, \\ p &= P_1^2 - P_2^2 - 2P_1P_2\cos\varphi, \end{aligned}\right\} \tag{4b}$$

gdzie ε i p jest energią i pędem wytworzonego wzbudzenia, m — masą neutronu. Mierząc P_1, P_2 i kąt φ można znaleźć zależność energii ε wzbudzenia od jego pędu p, czyli wyznaczyć widmo energetyczne helu II. Wyniki pomiarów potwierdziły słuszność propozycji Landaua dotyczącej kształtu krzywej widmowej.

Prześledzimy teraz, jak teoria Landaua tłumaczy zjawisko nadpłynności, tzn. przepływu bez tarcia. Rozpatrzmy hel II w temperaturze $T = 0$ K, przepływający przez kapilarę z prędkością v. Jeśli ciecz doznaje tarcia o ścianki rurki, jej energia kinetyczna rozprasza się w postaci ciepła. Nagrzewanie się cieczy oznacza przejście w stan wzbudzony. Ale układ kwantowy, jakim jest hel II, nie może zmieniać swej energii w sposób ciągły. Aby ciecz kwantowa mogła przejść ze stanu podstawowego w najbliższy stan wzbudzony, czyli zmienić swą energię o $\varepsilon = \varepsilon(p)$, musi powstać wzbudzenie elementarne, czyli kwazicząstka o energii ε i pędzie p. Energia i pęd wytworzonej kwazicząstki są tu określone w układzie odniesienia spoczywającym względem cieczy, a więc poruszającym się względem ścian rurki z prędkością v.

przepływ bez tarcia w teorii Landaua

W układzie odniesienia obserwatora związanego nieruchomo ze ścianami rurki energia wzbudzenia, czyli zmiana energii cieczy, wynosi $\varepsilon + \vec{p}\vec{v}$, zgodnie ze znanymi z mechaniki wzorami transformacyjnymi dla energii i pędu. Wartość tej zmiany energii cieczy musi być ujemna, gdyż dla obserwatora nieruchomego względem ścian rurki energia kinetyczna cieczy na skutek tarcia o ściany może się tylko zmniejszać. Mamy więc:

$$\vec{p}\vec{v} < 0. \tag{5}$$

Wyrażenie po lewej stronie znaku nierówności ma najmniejszą wartość wtedy, gdy wektor prędkości cieczy $\vec{v}$ i wektor pędu kwazicząstki $\vec{p}$ mają przeciwne zwroty. Wtedy

$$\varepsilon - pv < 0,$$

czyli

$$v > \frac{\varepsilon}{p}. \tag{6}$$

A więc minimalna wartość prędkości cieczy v_{min}, która jest konieczna do wytworzenia wzbudzenia elementarnego, nie może być mniejsza niż $(\varepsilon/p)_{min}$. Przy mniejszych prędkościach przepływu nie mogą powstać wzbudzenia w nadpłynnym helu i przepływ będzie się odbywał bez tarcia.

Zauważmy, że stosunek ε/p na wykresie zależności ε od p jest tangensem kąta α między osią odciętych a prostą przechodzącą przez początek układu i punkt o współrzędnych p, ε. Najmniejszy kąt $\alpha_{min} = (\varepsilon/p)_{min}$ utworzy styczna do krzywej $\varepsilon(p)$ w punkcie leżącym blisko minimum na rotonowej części widma (na rys. 16 styczna ta jest oznaczona linią przerywaną). Tangens najmniejszego kąta odpowiada prędkości $v_{min} = 60$ m/s. Jest to wartość ok. sto razy większa niż krytyczne prędkości przepływu obserwowane w doświadczeniach. Powód tej rozbieżności zostanie wyjaśniony później.

Aby mogła pojawić się nadpłynność, kąt α_{min} nie może być równy zeru dla żadnego punktu na krzywej $\varepsilon(p)$. Inaczej mówiąc, oś rzędnych nigdzie nie może być styczna do krzywej widma. Tak mogłoby się zdarzyć, gdyby np. widmo rotonów miało postać opisaną nie wzorem (3) lecz wzorem $\varepsilon = p^2/2m$, jak widmo cząstek swobodnych, albo gdyby we wzorze (3) nie było wyrazu Δ oznaczającego minimalną wartość dozwolonych energii rotonu. Wartość Δ nazywana jest przerwą energetyczną w widmie dozwolonych wartości energii rotonu.

Gdyby kąt α_{min} był równy zeru, to $\operatorname{tg}\alpha_{min} = v_{min} = 0$ i każda prędkość przepływu większa od zera powodowałaby utratę nadpłynności, czyli nadpłynność w ogóle nie mogłaby wystąpić. Warunek $\operatorname{tg}\alpha_{min} > 0$ nazywa się kryterium Landaua wystąpienia nadpłynności.

kryterium Landaua

Rozpatrzyliśmy przepływ helu II przez rurkę w temperaturze $T = 0$ K. W temperaturach wyższych od zera ciecz znajduje się w jednym ze swoich stanów wzbudzonych. Zamiast mówić, że ciecz jest w n-tym stanie wzbudzonym, możemy powiedzieć, że w cieczy powstało n elementarnych wzbudzeń. W tym ujęciu przejście cieczy w kolejny stan wzbudzony jest równoważne wytworzeniu jeszcze jednej kwazicząstki.

Jeśli temperatura cieczy nie jest zbyt wysoka, to koncentracja wzbudzeń w cieczy jest mała i wzbudzenia te można traktować jako niezależne od siebie. W temperaturach bliskich T_λ nie można już zaniedbać oddziaływań wzajemnych między wzbudzeniami i koncepcja wzbudzeń elementarnych staje się mało przydatna. Poniżej $T = 1{,}8$ K możemy jednak gaz wzbudzeń elementarnych uważać za gaz doskonały, tzn. możemy pominąć oddziaływania wzajemne cząstek tego gazu.

Kiedy wzbudzeń jest dużo, częściej się z sobą zderzają, czas ich życia Δt określony wzajemnymi zderzeniami jest więc krótki, a ponieważ z zasady nieokreśloności Heisenberga wynika, że $\Delta\varepsilon\Delta t = \hbar$, to nieokreśloność energii $\Delta\varepsilon$ wzbudzeń jest duża, porównywalna z ich własnymi energiami ε. Wtedy niecelowe jest posługiwanie się pojęciem wzbudzeń, którym nie można przyporządkować określonej energii.

Ponieważ wyprowadzając kryterium nadpłynności nigdzie nie wykorzystano założenia o nieistnieniu wzbudzeń elementarnych w cieczy, kryterium to pozostaje słuszne także i dla $T > 0$. A więc przy prędkościach przepływu mniejszych niż v_{min} nowe wzbudzenia tworzyć się nie będą. Jednak istniejące już w cieczy wzbudzenia mogą oddziaływać ze ściankami naczynia lub kapilary. Zachowują się one jak cząsteczki normalnego gazu, wymieniają więc ze ściankami swą energię i pęd. Gaz wzbudzeń elementarnych odpowiada więc temu, co w modelu dwupłynowym nazwaliśmy składnikiem normalnym.

gaz wzbudzeń elementarnych

Kwantowane wiry

Pozostała do wyjaśnienia sprawa, dlaczego obserwowane krytyczne prędkości przepływu są prawie sto razy mniejsze niż te, które wynikają z kryterium Landaua. Landau w swej teorii uwzględnił tylko dwa rodzaje wzbudzeń w helu II: fonony i rotony. W rzeczywistości prócz tych wzbudzeń elementarnych w nadpłynnym helu mogą tworzyć się wzbudzenia innego

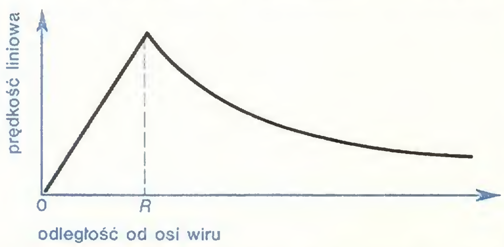
typu, tzw. kwantowane wiry. Są to wzbudzenia makroskopowe — w ruchu wirowym bierze udział duża liczba atomów helu. Energia wzbudzenia przypadająca na jeden atom jest mała — do wytworzenia wiru wystarczają więc małe prędkości przepływu.

Krytyczna prędkość związana z wytwarzaniem makroskopowego wiru jest odwrotnie proporcjonalna do średnicy rurki, przez którą przepływa hel II. W szerokich rurkach prędkość krytyczna jest bardzo mała i dlatego w takich rurkach trudno jest obserwować nadpłynność. Zależność prędkości krytycznej od średnicy rurki jest zrozumiała: im szersza rurka, tym łatwiej może się w niej wytworzyć makroskopowy wir.

wiry w płynach klasycznych

Wiry tworzą się również w płynach klasycznych, zarówno w cieczach, jak i w gazach. Są to jednak wiry o dużych, a często nawet ogromnych rozmiarach. Wir w wodzie można obserwować np. przy wypuszczaniu wody z wanny przez otwór odpływowy w dnie. W powietrzu niekiedy tworzą się wiry nazywane trąbami powietrznymi.

Wir obraca się wokół pewnej osi, która nie musi być linią prostą. Środkowa część wiru, tzw. jądro, obraca się jak bryła sztywna, ze stałą prędkością kątową ω, a więc prędkość liniowa v_w cząstek wewnątrz jądra wiru jest proporcjonalna do odległości r od osi obrotu, czyli $v_w = \omega r$. Na zewnątrz jądra ciecz lub gaz wiruje w ten sposób, że prędkość liniowa jest odwrotnie proporcjonalna do odległości od środka wiru $v_z = K/r$ (rys. 18). Na obwodzie jądra o promieniu R obie prędkości są sobie równe, czyli $\omega R = K/R$, skąd:



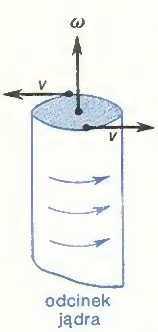
Rys. 18. Zależność prędkości liniowej v cząsteczek wirującego płynu od odległości r od osi wiru; R promień jądra wiru

$$K = \omega R^2. \qquad (17)$$

natężenia wiru

Wielkość K charakteryzuje dany wir i nazywa się natężeniem wiru. Często posługujemy się również pojęciem cyrkulacji lub wielkości od niej dwa razy mniejszej i nazywanej strumieniem wirowości. Strumień wirowości, określony jako $I = \pi K = \omega \cdot \pi R^2$, jest więc iloczynem wartości wektora wirowości $\vec{\omega}$ (czyli wektora prędkości kątowej) i powierzchni poprzecznego przekroju jądra. Wektor $\vec{\omega}$ jest równoległy do osi wiru, czyli prostopadły do wektora $\vec{v}$, zaś jego zwrot określa reguła śruby prawoskrętnej: jeśli obracamy śrubę w kierunku $\vec{v}$, to wkręca się ona w kierunku $\vec{\omega}$ (rys. 19).

strumień wirowości



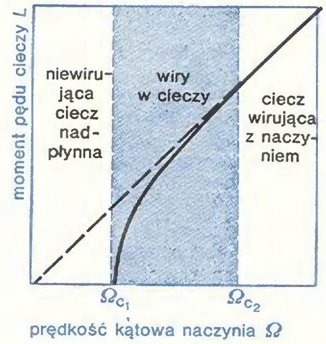
Rys. 19. Orientacja wektora wirowości (wektora prędkości kątowej jądra wiru) względem wektora prędkości liniowej

Strumień wirowości jest stały wzdłuż osi wiru. W cieczy doskonałej, tj. pozbawionej lepkości i nieściśliwej, jest on również stały w czasie, czyli wir nie może się wytworzyć, a jeśli już istnieje, to nie może zniknąć. Pod tym względem hel II nie może być uważany za ciecz doskonałą.

Moment pędu atomu helu krążącego dokoła jądra wiru jest równy $m_{He}vr$ (m_{He} oznacza masę atomu helu). Podobnie jak w wypadku elektronu krążącego wokół jądra atomowego, moment pędu nie może przybierać wartości dowolnej, lecz musi być krotnością $\hbar$. Ponieważ $v = K/r$, moment pędu atomu helu w wirze musi spełniać warunek: $m_{He}K = n\hbar$, czyli $I = n(h/2m_{He})$, gdzie n jest liczbą całkowitą. Wiry tworzące się w helu II są więc skwantowane, czyli ich strumień wirowości $I = \pi K$ jest całkowitą krotnością kwantu strumienia równego $h/2m_{He} = 0{,}5 \cdot 10^{-3}$ cm²/s.

kwanty strumienia wirowości

Wiry w helu II można wytworzyć przez wprowadzenie naczynia z helem w ruch obrotowy. Włókna wirowe są wtedy równoległe do osi obrotu naczynia i rozciągają się od dna naczynia do swobodnej powierzchni cieczy. Wiry zaczną się tworzyć, gdy prędkość kątowa Ω naczynia przekroczy pewną wartość krytyczną Ω_{c1} zależną od średnicy naczynia. W miarę wzrostu Ω coraz więcej wirów pojawia się w cieczy i moment pędu L cieczy wzrasta. Gdy Ω osiągnie drugą wartość krytyczną Ω_{c2}, cała ciecz będzie wirować jak ciecz normalna i $L = M\Omega$, gdzie M — masa cieczy (rys. 20).



Rys. 20. Zależność momentu pędu helu II od prędkości kątowej obracającego się naczynia

Próby doświadczalnego potwierdzenia kwantowania wirów w helu II przez długi czas kończyły się niepowodzeniem. Dopiero W. Vinen w 1961 r. wykazał istnienie kwantów strumienia wirowości. Badał on, jak zachowuje się drgająca struna, dokoła której istnieje cyrkulacja cieczy. Na strunę działa wtedy siła znana w hydrodynamice jako siła Magnusa. Siła ta jest prostopadła do kierunku wektora wirowości i do prędkości struny względem niewirującej cieczy, tj. cieczy odległej od jądra wiru. Jest ona także proporcjonalna do tych wielkości. Badania wykazały, że cyrkulacja helu II dokoła struny ma skwantowane wartości.

Najbardziej przekonywającego dowodu istnienia kwantów strumienia wirowości dostarczyli G.W. Reyfield i F. Reif, którzy mierzyli szybkość poruszania się jonów helu w cieczy nadpłynnej. Jon helu zachowuje się jak obca cząstka w helu II i podczas ruchu zderza się z wzbudzeniami cieplnymi, czyli doznaje tarcia. Celem doświadczenia było wyznaczenie liczby wzbudzeń elementarnych jako funkcji temperatury. Jony helu uzyskiwały prędkość początkową w polu elektrycznym i następnie poruszały się w cieczy nadpłynnej. W najniższej osiągniętej podczas tych doświadczeń temperaturze $T = 0{,}28$ K tarcie w helu II było tak małe, że jon wędrował bez zderzeń kilka cm, ale już w $T = 0{,}5$ K droga swobodna jonu wynosiła tylko 10^{-4} cm.

Najdziwniejsze wyniki uzyskano, gdy zwiększono natężenie pola elektrycznego, które nadawało jonom początkową energię kinetyczną. Okazało się, że prędkość nośnika ładunku maleje, gdy wzrasta jego energia. Oznaczało to, że ze wzrostem energii rośnie masa tworu niosącego ładunek; oszacowano, że ów twór musi się składać z wielu tysięcy atomów helu. Nasuwało się więc przypuszczenie, że jest to makroskopowy wir. Dalsze badania wykazały, że jony helu o wystarczająco dużej energii kinetycznej rzeczywiście powodują tworzenie się w helu II wirów w kształcie pierścieni, czyli podobnych do kółek z dymu papierosowego. Oś takiego wiru jest więc krzywą zamkniętą (kołem), a średnica kółka wirowego o energii 50 eV wynosi około 10^{-4} cm i rośnie ze wzrostem energii. Każdy z tych pierścieni wirowych niesie dokładnie jeden kwant strumienia wirowości równy $h/2m_{He}$. Tak więc kwantowanie, z którym zwykle spotykamy się na poziomie mikrokosmosu, tutaj zostało wykryte również w układzie makroskopowym.

wir makroskopowy

Koncepcje dotyczące kwantowania wirów w helu II odegrały ważną rolę w kształtowaniu się poglądów na wiry prądowe w nadprzewodnikach II rodzaju w stanie mieszanym. Takie wiry prądowe wytwarzają skwantowany strumień magnetyczny (→ Nadprzewodnictwo).

Zagadnienia nadpłynności a statystyka kwantowa

Wprowadzając koncepcję modelu dwupłynowego do zagadnienia helu II Tisza oparł się na teorii doskonałego gazu bozonów, tj. cząstek o spinie całkowitym. Nazwa tych cząstek pochodzi stąd, że podlegają one statystyce Bosego–Einsteina.

^{4}He jako gaz Bosego

Jądro atomu ^{4}He, składające się z parzystej liczby cząstek o spinach połówkowych, ma ogólny spin całkowity; atomy ^{4}He są więc bozonami. Oddziaływanie między atomami ciekłego helu jest bardzo słabe, ciekły hel może być więc w przybliżeniu potraktowany jako gaz. W wystarczająco niskiej temperaturze w doskonałym gazie bozonów powinna nastąpić kondensacja pewnej części gazu. Nie jest to jednak kondensacja w zwykłym sensie, tzn. skraplanie lub krystalizacja, lecz kondensacja w „przestrzeni" pędów. Innymi słowy, pewna liczba N_0 cząstek z ogólnej liczby N bozonów zawartych w objętości V powinna znaleźć się w stanie podstawowym, tj. najniższym stanie energetycznym z pędem równym zeru. Stosunek N_0/N jest funkcją temperatury. Przyjmując, że masa cząstki gazu Bosego jest równa masie atomu ^{4}He, otrzymuje się dla temperatury kondensacji wartość 3,14 K, a więc bardzo bliską temperaturze lambda. Zgodność jest bardzo dobra, zwłaszcza jeśli weźmiemy pod uwagę, że ciekły hel nie jest gazem doskonałym.

teoria Bogolubowa

Mikroskopową teorię niedoskonałego gazu bozonów opracował N. N. Bogolubow w 1947 r., posługując się metodą drugiego kwantowania (→ Teoria pola). W tej metodzie istotną sprawą jest dobór odpowiedniego hamiltonianu (tj. operatora energii), uwzględniającego międzycząstkowe oddziaływanie w układzie. Zakładając słabość oddziaływania i stosując odpowiednie przekształcenie hamiltonianu Bogolubow otrzymał widmo energetyczne układu zbliżone do widma Landaua, zwłaszcza w zakresie małych pędów. Zwróćmy uwagę, że Landau w swej teorii nigdzie nie korzystał z założeń dotyczących statystyki cząstek. Landau właściwie odgadł kształt widma energetycznego helu II, pasującego do danych termodynamicznych.

teoria Feynmana

Jeszcze inne podejście do zagadnienia nadpłynnego helu zaproponował R. P. Feynman. Korzystając z ogólnych zasad mechaniki kwantowej wprowadził on funkcję ψ_0, opisującą stan podstawowy helu II. Funkcja ψ_0 zależy tylko od położeń atomów helu w układzie i bardzo słabo zmienia się w przestrzeni od cząstki do cząstki. Stany wzbudzone opisuje funkcja ψ_E. Aby funkcja ψ_E mogła odpowiadać faktom doświadczalnym, musi spełniać szereg warunków. Na przykład ze względu na statystykę jąder ^{4}He, funkcja ψ_E musi być parzysta, a więc nie może zmieniać się przy zamianie atomów miejscami. Żądane warunki spełnia funkcja: $\psi_E = \sum_j f(r_j)\psi_0$. Sumowanie przeprowadza się po wszystkich atomach układu, a $f(r_j)$ jest funkcją tylko ich położeń r_j. Po podstawieniu funkcji ψ_E do równania Schrödingera i zastosowaniu metody wariacyjnej Feynman wykazał, że rozwiązaniem, przy którym energia układu jest najmniejsza, jest $f(r_j) = e^{(i/\hbar)pr_j}$, zaś wyrażenie na energię ma postać $\varepsilon(p) = \dfrac{p^2}{2mS(p)}$, gdzie p jest pędem wzbudzenia. Symbol $S(p)$ oznacza tzw. czynnik struktury cieczy. Kształtu funkcji $S(p)$ nie można obliczyć na podstawie ogólnych rozważań, ale można go wyznaczyć drogą analizy ugięć promieni rentgenowskich w cieczy. Znając $S(p)$ można obliczyć $\varepsilon(p)$, czyli widmo energetyczne układu. Otrzymane w ten sposób widmo niewiele różni się od widma Landaua.

Nadpłynność ^{3}He

Związek pomiędzy statystyką kwantową i nadpłynnością jest wyraźnie widoczny w wypadku lekkiego izotopu helu, a mianowicie ^{3}He. Izotop ten występuje w przyrodzie jako nikła domieszka ^{4}He; stanowi ona zaledwie ok. $10^{-3}\%$, nie wywiera więc żadnego zauważalnego wpływu na własności ^{4}He. Z powodu bardzo małej zawartości ^{3}He w naturalnych złożach wydobywanie go stamtąd jest nieopłacalne. ^{3}He otrzymuje się w reakcjach jądrowych przez bombardowanie atomów litu neutronami:

otrzymywanie izotopu ^{3}He

$$^6_3\text{Li} + n \rightarrow {}^3_1\text{H} + {}^4_2\text{He}.$$

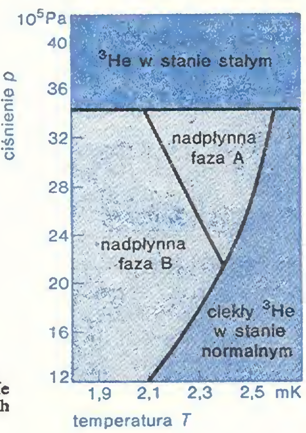
Powstający w czasie tej reakcji tryt ($^3_1$H) poprzez rozpad β przechodzi w $^3_2$He; okres połowicznego rozpadu wynosi 12 lat.

Jądro atomu $^3_2$He, złożone z trzech nukleonów, ma wypadkowy spin połówkowy, czyli jest fermionem. Układ takich cząstek podlega statystyce Fermiego–Diraca. Do układu fermionów stosuje się zasada Pauliego orzekająca, że w określonym stanie kwantowym nie może znaleźć się więcej niż jedna cząstka z danego układu. W stanie z najniższą energią nie może się więc znaleźć duża liczba cząstek, czyli kondensacja nie nastąpi. I rzeczywiście, oziębiany aż do 3 mK (0,003 K) lekki izotop helu nie przechodzi w stan nadpłynny, chociaż opierając się na zasadzie odpowiadających sobie stanów, przez analogię do $_2^4$He, można by się spodziewać, że dla $^3_2$He temperatura ambda wyniesie około 1,5 K.

Po opracowaniu teorii nadprzewodnictwa stało się jasne, że jeśli w układzie fermionów pomiędzy poszczególnymi cząstkami działają jakieś, chociażby nawet niewielkie, siły przyciągania dalekiego zasięgu, to w odpowiednio niskiej temperaturze może dojść do połączenia się fermionów w pary, które będą mieć sumaryczny spin całkowity. Takie pary (tzw. pary Coopera) nie są już fermionami i nie podlegają zakazowi Pauliego, może więc nastąpić kondensacja. Kondensacja w układzie elektronów przewodnictwa w metalach odpowiada przejściu w stan nadprzewodnictwa. Kondensat elektronowy może się wtedy poruszać w metalu bez tarcia, czyli bez oporu elektrycznego (→ Nadprzewodnictwo).

łączenie się atomów ^{3}He w pary

W ciekłym ^{3}He działają pomiędzy poszczególnymi atomami siły przyciągania o dalekim zasięgu. Są to siły van der Waalsa. W wystarczająco niskiej temperaturze te słabe siły mogą doprowadzić do łączenia się atomów ^{3}He w pary analogiczne do elektronowych par Coopera w nadprzewodnikach. Temperatura, w której ^{3}He przechodzi w stan nadpłynny, zależy od ciśnienia i zmienia się od 0,8 mK do 2,6 mK przy



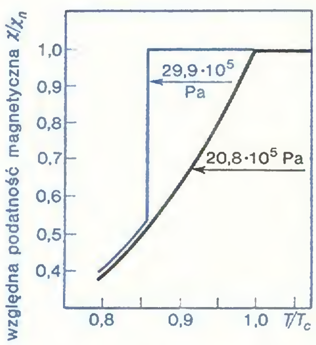
Rys. 21. Wykres stanu ^{3}He w przedziale najniższych temperatur

zmianie ciśnienia od zera do $33,5 \cdot 10^5$ Pa. Powyżej tego ciśnienia ^{3}He zestala się. Nadpłynność ^{3}He została doświadczalnie potwierdzona dopiero w ostatnich latach. Okazało się przy tym, że przy ciśnieniu zawartym w przedziale $(21,2\text{–}33,5) \cdot 10^5$ Pa ^{3}He ma dwie nadpłynne fazy, które oznaczono literami *A* i *B* (rys. 21). Jedna z tych faz (faza *A*) wykazuje wyraźną anizotropię własności fizycznych. Przejście z fazy nienadpłynnej w fazę *A* jest przemianą fazową II rodzaju (przemianą lambda), przejście między fazą *A* i *B* jest przemianą fazową I rodzaju.

faza *A* i *B*

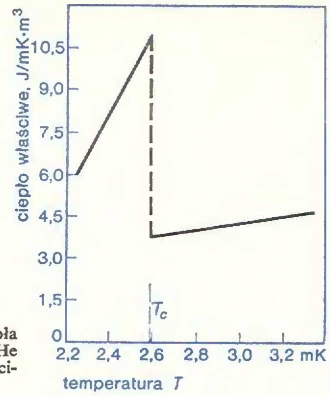
Położenie linii rozdzielających obszary poszczególnych faz ^{3}He na wykresie stanu w układzie ciśnienie–temperatura zależy od natężenia *H* zewnętrznego pola magnetycznego. Jądra ^{3}He mają (ze względu na nieparzystą liczbę nukleonów) różny od zera moment magnetyczny. Podany na rys. 21 wykres fazowy przedstawia położenie linii dla $H = 0$. Ze wzrostem natężenia pola linia rozdzielająca fazę normalną od nadpłynnej przesuwa się w stronę niższych temperatur. Przy łączeniu się atomów ^{3}He w pary ich jądrowe momenty magnetyczne nie kompensują się i dlatego nadpłynny ^{3}He (zwłaszcza faza *A*) wykazuje znaczną podatność magnetyczną. Podatność magnetyczna fazy *A* jest taka sama, jak w fazie normalnej, w fazie *B* podatność maleje ze spadkiem temperatury (rys. 22). A więc już chociażby ze względu na własności magnetyczne obie nadpłynne fazy ^{3}He różnią się od helu II. Różnice te występują również przy badaniu innych własności fizycznych. Ze względu na możliwość uporządkowania przestrzennego pod wpływem pola magnetycznego obie nadpłynne fazy ^{3}He można uważać za ciekłe kryształy (→ Ciekłe kryształy).

podatność magnetyczna ^{3}He



Rys. 22. Zależność względnej podatności magnetycznej ciekłego ^{3}He od temperatury przy dwu różnych wartościach ciśnienia. χ_n oznacza podatność magnetyczną ^{3}He w stanie normalnym

Badanie własności ^{3}He w bardzo niskich temperaturach jest trudnym zadaniem, toteż prace badawcze — chociaż dość zaawansowane — nie dały jeszcze pełnego obrazu zagadnienia. O trudnościach eksperymentowania w przedziale temperatur rzędu milikelwina niech świadczy takie porównanie: aby osiągnąć punkt przemiany lambda w ^{4}He, należy uzyskać temperaturę niższą ok. 140 razy od pokojowej, natomiast temperatura przemiany ^{3}He w stan nadpłynny



Rys. 23. Zależność ciepła właściwego ciekłego ^{3}He od temperatury (pod ciśnieniem 3,3 MPa)

będzie jeszcze tysiąckrotnie niższa niż temperatura przemiany lambda ^{4}He. Ponieważ ciepło właściwe fazy skondensowanej w tak niskich temperaturach jest znikomo małe (rys. 23), to nawet nieznaczne dopływy ciepła z zewnątrz mogą ogrzać ^{3}He powyżej jego punktu przemiany w stan nadpłynny.

Struktura nadpłynnych faz ^{3}He badana jest także od strony teoretycznej. Przypuszcza się, że fazy te odpowiadają łączeniu się w pary atomów ^{3}He znajdujących się w stanie *P*, tj. w stanie z orbitalnym momentem pędu $l = 1$. Modelem fazy *B* jest model z izotropową przerwą energetyczną opracowany przez R. Baliana i N. R. Werthamera, zaś model fazy *A* z anizotropią przerwy energetycznej zaproponowali P. W. Anderson i P. Morel. Opracowano sześć różnych typów modeli takiej anizotropowej fazy. Który z tych typów odpowiada rzeczywistości, zadecydują wyniki badań doświadczalnych, w szczególności badania jądrowego rezonansu magnetycznego. Badania doświadczalne dostarczą także materiału, na podstawie którego można będzie poznać widma energetyczne nadpłynnych faz ^{3}He.

model fazy *B* i model fazy *A*

Odrębnym i dość obszernym zagadnieniem, do którego istotny wkład wnieśli także polscy fizycy (Z. Galasiewicz), jest nadpłynność mieszanin obu izotopów helu: ^{3}He i ^{4}He. Taka mieszanina może być przykładem mieszaniny cieczy kwantowych typu Fermiego i typu Bosego. Sam proces rozpuszczania ^{3}He w ^{4}He jest procesem oziębiającym i dzięki temu ma ważne znaczenie w kriogenice przy chłodzeniu od 0,5 do 0,05 K. Dalsze chłodzenie odbywa się przez krystalizację ^{3}He pod zwiększonym ciśnieniem.

C. T. Lane *Nadpłynność*, Warszawa 1967; K. Mendelssohn *Na drodze do zera bezwzględnego*, Warszawa 1970.

Nadprzewodnictwo

Eugeniusz Trojnar

Nadprzewodnictwo, czyli nieskończenie wielkie przewodnictwo elektryczne odkrył w 1911 r. H. Kamerlingh-Onnes. Wkrótce po udanym skropleniu helu i osiągnięciu w ten sposób temperatur rzędu kilku kelwinów Kamerlingh-Onnes zajął się badaniem zmian oporu elektrycznego metali przy ich oziębianiu do niskich temperatur.

odkrycie nadprzewodnictwa

Wiadomo, że opór czystych metali wyraźnie zależy od temperatury, przy czym w temperaturach bliskich pokojowej zależność ta jest liniowa, zaś w niskich temperaturach odbiega od liniowości. W metalach zanieczyszczonych zależność oporu od temperatury jest znacznie słabsza, a w niektórych stopach, jak manganin czy konstantan, opór od temperatury prawie nie zależy. Kamerlingh-Onnes wybrał więc do badań rtęć jako metal, który w owym czasie najlepiej można było oczyścić (stosując kilkakrotną destylację). Ku jego zdziwieniu opór rtęci, zamiast stopniowo zmieniać się wraz z temperaturą, w pobliżu $T = 4$ K nagle zmalał do niewykrywalnie małej wartości. Później okazało się, że czystość metalu nie grała tu istotnej roli, gdyż opór zanieczyszczonej rtęci także nagle znikał po oziębieniu jej do $T < 4$ K. Od czystości metalu zależała tylko wartość oporu, jaką metal wykazywał przed przejściem w stan nadprzewodnictwa.

opór nadprzewodnika

Zagadnienie, czy opór nadprzewodnika spada rzeczywiście do zera, czy tylko do skończonej, chociaż

bardzo małej wartości, trudno rozstrzygnąć przy użyciu zwykłych przyrządów pomiarowych. Przyrządy zawsze wnoszą pewien błąd do pomiaru, nie możemy więc zmierzyć wielkości mniejszej od tego błędu. Możemy jedynie powiedzieć, że sama wielkość nie przekracza błędu pomiaru. Najdokładniejsza metoda pomiaru bardzo małego oporu polega na wzbudzeniu w obwodzie (np. w pierścieniu z nadprzewodnika, którego opór chcemy zmierzyć) prądu i pomiarze czasu jego zaniku. Jeżeli obwód ma skończony opór, to natężenie prądu będzie z czasem maleć, co można wykryć mierząc natężenie pola magnetycznego wytworzonego przez ten prąd. Takie doświadczenie z ołowianym pierścieniem nadprzewodzącym przeprowadził S. C. Collins; po upływie dwu i pół roku nie stwierdził on żadnego zauważalnego osłabienia prądu krążącego w pierścieniu. Uwzględniając dokładność pomiaru natężenia pola magnetycznego i indukcyjności pierścienia oszacowano, że opór właściwy nadprzewodzącego ołowiu nie mógł być większy niż 10^{-25} $\Omega \cdot m$. Jest to niezwykle mała wartość, około 10^{17} razy mniejsza niż opór właściwy najlepszych przewodników w temperaturze pokojowej, jakimi są srebro i miedź. Możemy więc uważać, że opór nadprzewodnika jest dokładnie równy zeru.

Obraz kwantowy

Zjawisko nadprzewodnictwa, podobnie jak i zjawisko nadpłynności (→ Nadpłynność), jest przejawem działania praw fizyki kwantowej w skali makroskopowej. Między tymi zjawiskami istnieją pewne analogie; nadprzewodnictwo można potraktować jako nadpłynność cieczy złożonej z naładowanych cząstek. Tą cieczą jest układ elektronów przewodnictwa w metalach. W obu wypadkach, zarówno nadpłynności helu II, jak i nadprzewodnictwa, mamy do czynienia z możliwością przepływu bez tarcia, czyli bez strat energii.

warunki występowania nadprzewodnictwa

Nadprzewodnictwo występuje tylko w niskich temperaturach. Najwyższa dotychczas notowana temperatura, w której stwierdzono jeszcze istnienie nadprzewodnictwa, wynosiła 23,2 K. Niskie wartości temperatur, w których może pojawiać się nadprzewodnictwo, świadczą o tym, że oddziaływanie wywołujące to zjawisko jest bardzo słabe. W wyższych temperaturach energia cieplna kT (k — stała Boltzmanna, T — temperatura bezwzględna) jest już zbyt duża i niszczy uporządkowanie nadprzewodzące. Energia oddziaływania wywołującego nadprzewodnictwo przypadająca na jedną oddziałującą cząstkę jest więc rzędu kT_k, gdzie T_k jest temperaturą krytyczną, powyżej której nadprzewodnictwo znika.

Pomimo że możliwość przepływu trwałego prądu elektrycznego bez strat odkryto jeszcze w 1911 r., to na wyjaśnienie istoty tego zjawiska na podstawie mikroskopowych własności materii trzeba było poczekać aż do 1957 r., a więc niemal pół wieku. Ten czas wydaje się bardzo długi, zwłaszcza jeśli porównać go z czasem, który upłynął od wykrycia nadpłynności (1940 r.) do jej objaśnienia przez L. D. Landaua (1943 r.).

Nadpłynność można było objaśnić uwzględniając fakt, że atomy helu są bozonami i mogą występować w stanie kondensacji Bosego. Elektrony przewodnictwa w metalu są jednak fermionami, podlegają więc regule zakazu Pauliego i duża ich liczba nie może się kondensować w jednym stanie energetycznym z pędem równym zeru. Gdy jednak dwa elektrony z przeciwnymi spinami i pędami połączą się w parę, to taka para będzie mieć ogólny spin całkowity i nie będzie już podlegać zakazowi Pauliego.

Przypuszczenie, że elektrony mogą łączyć się w pary, wyraził R. A. Ogg jeszcze przed opracowaniem teorii nadprzewodnictwa, ale wydawało się ono wtedy (1946 r.) nieprawdopodobne, gdyż trudno było sobie wyobrazić elektrony powiązane w pary wbrew siłom odpychania kulombowskiego. Dopiero H. Fröhlich w 1950 r. wskazał na rolę, jaką w nadprzewodnictwie gra oddziaływanie elektronów z siecią krystaliczną. Wykrycie w tym samym roku efektu izotopowego potwierdziło przypuszczenia Fröhlicha. Mechanizm tworzenia się par elektronowych przy udziale sieci wyjaśnił L. Cooper w 1956 r., dlatego takie pary nazywamy często parami Coopera.

pary Coopera

Pary elektronów nie są jednak bozonami w ścisłym znaczeniu. Doświadczenie wykazuje, że odległość, z jakiej oddziałują z sobą elektrony pary, jest znacznie większa niż średnia odległość między poszczególnymi parami. W takim układzie, złożonym z przenikających się wzajemnie par cząstek, daje o sobie znać wewnętrzna struktura pary, a więc fakt, że jest ona złożona z fermionów. Znajduje to odbicie w widmie wzbudzeń energetycznych układu elektronów nadprzewodnika. Kondensat bozonowy, jakim jest hel II, ma widmo wzbudzeń energetycznych typu Bosego, tzn. wzbudzenia (kwazicząstki) są także bozonami, ponieważ zarówno fonony, jak i rotony w nadpłynnym helu mają spin równy zeru; takie kwazicząstki mogą tworzyć się pojedynczo. Natomiast kwazicząstki typu Fermiego (ze względu na zachowanie spinu) powstają wyłącznie parami, np. elektron i dziura lub dwa wzbudzenia elektronowe w wypadku rozerwania pary Coopera. Widmo typu Fermiego charakteryzuje kondensat elektronowy. Wzbudzenia układu elektronów w nadprzewodniku zachowują się jak elektrony przewodnictwa w normalnym metalu.

elektronowe ciepło właściwe

W wyjaśnianiu problemu widma wzbudzeń elektronowych w nadprzewodnikach dużą rolę odegrały staranne pomiary ciepła właściwego. Ciepło właściwe układu elektronów przewodnictwa w metalu normalnym jest liniową funkcją temperatury, w wypadku zaś nadprzewodnika elektronowe ciepło właściwe jest proporcjonalne do $e^{-\Delta/kT}$. Taka zależność wskazuje na istnienie przerwy energetycznej Δ (pasma zabronionych wartości energii) w widmie wzbudzeń elektronowych w nadprzewodniku. Istnienie przerwy energetycznej potwierdziły później badania pochłaniania promieniowania elektromagnetycznego i ultradźwięków, oraz — najbardziej przekonywająco — zjawiska tunelowe.

Mikroskopowa teoria nadprzewodnictwa uwzględniająca wyniki dotychczasowych badań została stworzona w 1957 w. przez J. Bardeena, L. Coopera i J. Schrieffera, którzy za tę pracę otrzymali w 1973 r. nagrodę Nobla. Od pierwszych liter nazwisk twórców jest ona często nazywana teorią BCS. Od czasu powstania teoria ta jest ciągle rozwijana i udoskonalana przez wielu badaczy. Obecnie nauka o nadprzewodnictwie jest bardzo rozbudowaną dziedziną fizyki współczesnej i w tym artykule możemy rozpatrzyć tylko niektóre jej aspekty.

teoria BCS

Korelacja w układzie elektronów

Elektrony przewodnictwa w metalu

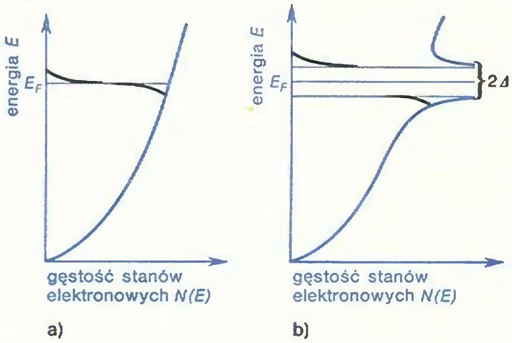
Wyjaśnijmy teraz bardziej szczegółowo przyczyny znikania oporu elektrycznego. W metalach prąd elektryczny jest to przepływ elektronów przewodnictwa (dokładniejsze omówienie przewodnictwa w metalu zawiera hasło → Metale), tj. tych elektronów, które nie są związane z poszczególnymi atomami, lecz są wspólne dla całej sieci krystalicznej. Często nazywamy je elektronami swobodnymi, chociaż nie są one w pełni swobodne, gdyż — po pierwsze — znajdują się w zmieniającym się okresowo w przestrzeni potencjale sieci krystalicznej, a po drugie — mogą przybierać tylko dozwolone, dyskretne wartości energii, przy czym — zgodnie z regułą zakazu Pauliego — ten sam poziom energetyczny mogą zajmować co najwyżej dwa elek-

elektrony swobodne

trony różniące się orientacją spinu. W danym stanie kwantowym może więc znaleźć się tylko jeden eletron układu.

gęstość stanów elektronowych

Gęstość stanów elektronowych N, czyli liczba ΔZ stanów kwantowych przypadających na dany przedział energii ΔE w jednym molu metalu (albo w jednostce objętości), jest funkcją energii E. Gęstość stanów definiujemy jako $N(E) = \Delta Z(E)/\Delta E$. Wykres zależności N od E dla pasma przewodnictwa jest przedstawiony na rys. 1a. Ze względu na wygodę w późniejszym posługiwaniu się wykresem oś energii kierujemy pionowo, zaś oś gęstości stanów — poziomo.



Rys. 1. Zależność gęstości stanów elektronowych w zależności od energii: a) w metalu normalnym, b) w nadprzewodniku (dla wyrazistości przerwa energetyczna mocno powiększona). Czarną linią oznaczono stany zajęte w $T > 0$

W temperaturze $T = 0$ układ elektronów przewodnictwa ma najniższą energię. Elektrony zajmują wszystkie nisko leżące stany, a najwyższy z zajętych stanów nazywa się poziomem Fermiego (E_F). Mówimy, że układ elektronów przewodnictwa jest w swym stanie podstawowym (niewzbudzonym). Energia całego układu jest oczywiście sumą energii elektronów we wszystkich stanach zajętych.

wzbudzenia termiczne (kwazicząstki)

W $T > 0$ niektóre elektrony wzbudzają się termicznie na wyższe poziomy energetyczne i zwalniają stany pod poziomem Fermiego. Elektrony te możemy traktować jako wzbudzenia termiczne układu elektronów. Podobnie można traktować zwolnione stany (dziury) pod poziomem Fermiego. W $T > 0$ następuje więc rozmycie poziomu Fermiego. Szerokość przedziału rozmycia jest rzędu kT. W temperaturze pokojowej stanowi to około 0,025 eV. Dla porównania podamy, że wysokość poziomu Fermiego nad dnem pasma przewodnictwa wynosi kilka eV (do 10 eV).

Elektrony zajmujące niskie stany energetyczne nie mogą zmieniać swej energii, gdyż wszystkie sąsiednie stany dozwolone są już zajęte. Nie mogą więc one być przyspieszane w polu elektrycznym, czyli nie biorą udziału w przepływie prądu. Pod wpływem pola mogą zmieniać swą energię tylko te elektrony, które zajmują stany w przedziale rozmycia, tj. w pobliżu poziomu Fermiego.

Elektrony wzbudzone termicznie, podobnie jak i dziury, są kwazicząstkami, bo ich własności zależą od sieci krystalicznej, w której się poruszają. Ich ruch w sieci podlega prawom mechaniki kwantowej i daje się opisać funkcjami falowymi modulowanymi z okresem sieci. W niezaburzonej, ściśle okresowej sieci krystalicznej fale elektronowe powinny się rozchodzić bez przeszkód, czyli nie powinny ulegać rozproszeniom.

opór elektryczny w sieci krystalicznej

Opór elektryczny nie może więc istnieć w takiej sieci. Defekty struktury i drgania cieplne jonów sieci zaburzają jednak okresowość struktury krystalicznej i rozpraszają fale elektronowe. W procesie rozproszenia elektrony przekazują sieci energię nabytą w polu elektrycznym. Opór elektryczny jest więc wynikiem zaburzeń w okresowości struktury krystalicznej. Ze spadkiem temperatury słabną drgania jonów sieci i opór metali maleje. Ta zaś część oporu, która jest spowodowana defektami struktury lub domieszkami obcych atomów nie zależy od temperatury i pozostaje nawet w najniższych temperaturach, jeśli metal nie przechodzi w stan nadprzewodnictwa. Jest to tzw. opór resztkowy metali.

rozpraszanie elektronów na fononach

Drgania cieplne sieci krystalicznej są skwantowane a kwanty drgań nazywają się fononami (→ Dynamika sieci krystalicznej). Można więc mówić o rozproszeniach elektronów na fononach jako o przyczynie tej części oporu elektrycznego, która zależy od temperatury. Opór wzrasta z temperaturą, gdyż zwiększa się gęstość gazu fononowego i zderzenia elektronów z fononami są częstsze. W niektórych metalach w odpowiednio niskiej temperaturze elektrony nie doznają rozproszeń na fononach i defektach sieci. Wskazówką, gdzie szukać przyczyny tego zjawiska było odkrycie efektu izotopowego.

efekt izotopowy

Efekt izotopowy polega na tym, że różne izotopy tego samego pierwiastka nadprzewodzącego (np. Sn lub Hg) mają różne temperatury przejścia w stan nadprzewodzący (T_k), przy czym dla danego pierwiastka zachodzi przybliżona zależność: $T_k \cdot \sqrt{M} = \text{const}$, gdzie M oznacza masę atomową izotopu. Okazało się więc, że dla nadprzewodnictwa, które jest zależne od układu elektronów w metalu, istotna jest również masa jonów tworzących sieć krystaliczną. Należy zatem uwzględnić ruch tych jonów, czyli ich drgania. Przyczyn nadprzewodnictwa trzeba przeto szukać w oddziaływaniu elektronów przewodnictwa z drganiami sieci jonowej, czyli z fononami. Ciekawe, że właśnie to oddziaływanie, którego wynikiem jest występowanie oporu elektrycznego, w pewnych warunkach może prowadzić do jego znikania. Tym można wytłumaczyć fakt, że najlepsze przewodniki elektryczności, jakimi są srebro, złoto czy miedź, nie przechodzą w stan nadprzewodnictwa (przynajmniej w tym przedziale temperatur, w którym były dotychczas przebadane); oddziaływanie elektronów z siecią jest w tych metalach zbyt słabe.

Pary Coopera

Jeśli w układzie fermionów (czyli cząstek podlegających regule zakazu Pauliego) między poszczególnymi cząstkami działają siły przyciągania dalekiego zasięgu, to mogą one doprowadzić do łączenia się cząstek o przeciwnych spinach w pary, które mają ogólny spin całkowity i nie podlegają już regule Pauliego. Pary takie mogą więc utworzyć kondensat cząstek znajdujących się na tym samym poziomie energetycznym. Od nazwiska twórcy tej koncepcji fermiony związane w pary nazywają się parami Coopera.

Prowadzące do nadpłynności łączenie się cząstek w pary Coopera wydaje się być powszechnym zjawiskiem w układach fermionów. Proces ten zachodzi np. w ciekłym ^{3}He. Przypuszcza się, że zachodzi on również w gwiazdach neutronowych. Najprawdopodobniej pulsary należą do gwiazd nadpłynnych.

Pomiędzy elektronami przewodnictwa w metalu, jako cząstkami o ładunku ujemnym, działają siły wzajemnego odpychania. Z powodu ekranowania elektronów dodatnimi ładunkami otaczających je jonów sieci krystalicznej, siły te mają krótki zasięg, rzędu kilku odległości międzyatomowych. Jak wskazuje efekt izotopowy, źródeł dalekozasięgowych sił wzajemnego przyciągania się elektronów należy szukać w oddziaływaniu elekronów za pośrednictwem sieci krystalicznej, tzw. oddziaływania elektronowo-fononowego.

oddziaływanie elektronowo-fononowe

Przyciągające oddziaływanie między elektronami za pośrednictwem sieci możemy sobie wyobrazić w następujący sposób. Elektron porusza się przez sieć i deformuje ją na skutek oddziaływania kulombowskiego. Deformacja polega na zgęszczaniu dodatnich jonów sieci w pobliżu elektronu. To zgęszczenie dodatnich jonów oddziałuje z kolei na inny elektron.

W rezultacie oba elektrony przyciągają się wzajemnie za pośrednictwem sieci. Zgęszczenie jonów oddziaływa na elektrony na znacznie większe odległości niż bezpośrednie oddziaływanie między elektronami, jest więc oddziaływaniem dalekiego zasięgu (rzędu kilkuset lub kilku tysięcy odległości międzyatomowych).

Kwantowy opis zjawiska oddziaływania wzajemnego między elektronami poprzez sieć jest następujący. Elektron poruszający się w krysztale wzbudza energetycznie sieć jonową na wyższy poziom, wytwarza więc kwant wzbudzenia sieci, czyli fonon. Ten fonon zostaje pochłonięty przez inny elektron i sieć wraca do poprzedniego stanu. Siła przyciągania między elektronami jest zatem siłą wymiany fononu. Siły wymiany znane są w fizyce, np. w teorii jądra atomowego, gdzie siła przyciągania się nukleonów jest siłą wymiany mezonu π.

wymiana fononów

Tak zwane wirtualne, krótkożyjące fonony wymiany są niezależne od fononów — wzbudzeń cieplnych sieci krystalicznej. Te ostatnie pojawiają się tylko w $T > 0$, natomiast fonony wymiany są emitowane i pochłaniane przez elektrony także w $T \approx 0$.

Przerwa energetyczna

Energia fononu nie może być dowolnie wielka, gdyż długość fali dźwięku rozchodzącego się w sieci krystalicznej nie może być mniejsza niż stała sieci, czyli odległość między sąsiednimi węzłami (energia fononu ε jest związana z długością fali dźwiękowej λ zależnością: $\varepsilon = hu/\lambda = h\nu$, gdzie u oznacza prędkość dźwięku, h — stałą Plancka, ν — częstość dźwięku). Maksymalna energia fononu $\varepsilon_{max} = h\nu_{max}$ w metalach wynosi ok. 10^{-2} eV. W procesie łączenia się w pary Coopera mogą więc brać udział tylko te elektrony, które mają energię mieszczącą się w przedziale o szerokości $h\nu_{max}$ po obu stronach poziomu Fermiego. Jest to wąski przedział w porównaniu z wysokością E_F poziomu Fermiego nad dnem pasma przewodnictwa ($E_F \approx 10$ eV).

Wiązanie się elektronów w pary prowadzi do zmiany rozkładu stanów energetycznych w pobliżu poziomu Fermiego i do pojawienia się pasma zabronionych wartości energii, czyli do wytworzenia się przerwy w widmie wzbudzeń elektronowych (rys. 1b). Przerwa ta oddziela stan podstawowy od stanów wzbudzonych. Szerokość przerwy energetycznej nie może być większa niż $h\nu_{max}$; zgodnie z teorią BCS, w temperaturze $T = 0$ jest ona równa:

szerokość przerwy energetycznej

$$2\Delta(0) = 3{,}52kT_k = 4h\nu_{max}e^{-1/g}, \quad (1)$$

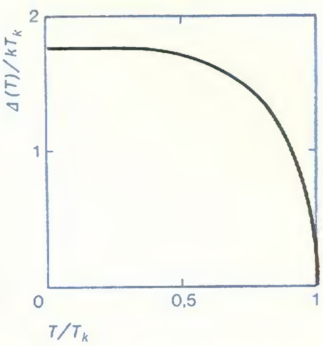
gdzie k oznacza stałą Boltzmanna, g jest stałą materiałową zależną od wielkości oddziaływania elektron–fonon (V) i od gęstości stanów elektronowych $N(E_F)$ w normalnym metalu w pobliżu poziomu Fermiego. Z teorii wynika, że stała $g = VN(E_F)$ nie może być większa niż 0,5 (na ogół bywa mniejsza).

Przerwa energetyczna w widmie wzbudzeń elektronowych istnieje również w izolatorach i półprzewodnikach, gdzie oddziela pasmo walencyjne od pasma przewodnictwa. Tam jednak istnienie przerwy energetycznej jest uwarunkowane budową krystaliczną substancji, natomiast w nadprzewodnikach przerwa energetyczna powstaje w wyniku pojawienia się uporządkowania w układzie elektronów przewodnictwa. Niszcząc to uporządkowanie (np. przez ogrzanie nadprzewodnika powyżej T_k) usuwa się przerwę energetyczną.

W temperaturze zera bezwzględnego elektrony niezwiązane w pary zajmują stany poniżej przerwy energetycznej; nad przerwą stany są nieobsadzone. Wszystkie pary Coopera stanowiące kondensat elektronowy są w tym samym stanie energetycznym odpowiadającym poziomowi Fermiego. W $T > 0$ wzbudzenia cieplne powodują osłabienie korelacji między elektronami i rozerwanie się części par. Układ elektronów uzyskuje więc dodatkową energię wzbudzenia. Podobnie jak w wypadku helu II, pojawienie się wzbudzeń energetycznych w układzie elektronów możemy potraktować jako tworzenie się kwazicząstek o określonej energii. Ponieważ te kwazicząstki zachowują się tak samo jak elektrony przewodnictwa, nazywamy je niekiedy elektronami normalnymi. Wzbudzenia elektronowe zajmują stany leżące nad przerwą energetyczną.

wpływ temperatury

Do rozerwania pary Coopera i utworzenia dwu wzbudzeń elektronowych potrzebna jest energia równa szerokości przerwy, czyli wynosząca 2Δ. Energia o wartości 2Δ jest energią wiązania pary. Ze wzrostem temperatury coraz więcej par ulega rozerwaniu i jednocześnie maleje szerokość przerwy energetycznej (rys. 2). Energia wiązania pary jest tym większa, im więcej elektronów jest związanych w pary, czyli im lepsza jest korelacja między elektronami. W $T = T_k$ szerokość przerwy maleje do zera i wszystkie pary ulegają rozerwaniu.



Rys. 2. Zależność szerokości przerwy energetycznej w nadprzewodniku od temperatury

kolektywny charakter

Tworzenie się par Coopera jest zjawiskiem kolektywnym, w którym bierze udział jednocześnie duża liczba cząsteczek, dlatego nie należy sobie wyobrażać pary Coopera jako dwóch wyodrębnionych, złączonych ze sobą elektronów. Pojedyncza para nie może istnieć, należy w zasadzie mówić o korelacji między elektronami o przeciwnych pędach i spinach. Korelacja ta prowadzi do obniżenia energii układu o $2\Delta(T)$ na każdą parę korelujących z sobą elektronów. W miarę wzrostu temperatury coraz liczniejsze wzbudzenia cieplne naruszają tę korelację, maleje więc względna liczba par oraz energia wiązania pary. W pobliżu T_k proces ten przybiera gwałtowny charakter.

średnia odległość między parami Coopera

Spróbujmy jeszcze oszacować, jaka jest średnia odległość między parami Coopera. Ponieważ w 1 cm³ metalu jest około 10^{22} atomów, to — przyjmując jeden elektron przewodnictwa na atom — otrzymamy tyleż elektronów przewodnictwa. W procesie łączenia się w pary mogą brać udział tylko te elektrony, których energia leży w wąskim przedziale w pobliżu poziomu Fermiego, co stanowi około 10^{-4} ogólnej liczby elektronów przewodnictwa, czyli około 10^{18} elektronów na 1 cm³. Średnia odległość między parami wyniesie więc około 10^{-6} cm. Jest to około sto razy mniej, niż wynosi odległość, z jakiej oddziałują z sobą elektrony pary (czyli średni rozmiar jednej pary). W zasięgu korelacji jednej pary mieści się wiele innych korelujących z sobą elektronów; poszczególne pary nakładają się na siebie i obraz pojedynczej pary zaciera się.

Brak oporu elektrycznego

Zastanówmy się teraz, jak w koncepcji par Coopera tłumaczy się zjawisko bezoporowego przepływu prądu elektrycznego w nadprzewodniku, czyli nadpłynność w układzie elektronów. Pojawienie się oporu elektrycznego oznacza, że elektrony w metalu ulegają rozproszeniu, czyli zmieniają swój pęd w zderzeniach z fononami lub defektami struktury krystalicznej. W procesie łączenia się w pary Coopera uczestniczą elektrony o przeciwnych pędach, np. K i $-K$, tak

że sumaryczny pęd pary — gdy prąd nie płynie — jest równy zeru. Para biorąca udział w przepływie prądu ma sumaryczny pęd różny od zera, równy np. $2P$, czyli elektrony pary mają pędy $K+P$ i $-K+P$. Gdyby jeden z elektronów pary uległ rozproszeniu, jego pęd zmieniłby się o wartość Q, wynosiłby więc $-K+P+Q$; taki elektron nie mógłby już korelować z elektronem o pędzie $K+P$, czyli para uległaby rozerwaniu. To zaś zwiększyłoby energię układu o 2Δ, co jest dla układu niekorzystne (każdy układ stara się mieć jak najniższą energię). Elektrony związane w parę nie biorą zatem udziału w procesie rozpraszania, czyli ich przepływ odbywa się bez tarcia.

krytyczna gęstość prądu

Gdyby jednak gęstość prądu wzrosła do takiej wartości, że energia kinetyczna nośników prądu nadprzewodzącego przekroczyłaby spadek energii wynikły z korelacji, to korelacja uległaby zniszczeniu, gdyż przestałaby być energetycznie korzystna. Gęstość prądu, przy której w nadprzewodnikach pojawia się opór, nazywa się krytyczną gęstością prądu. Nasuwa się tu analogia z krytyczną prędkością przepływu helu II.

Kwantowanie strumienia magnetycznego

prąd stacjonarny w nadprzewodniku

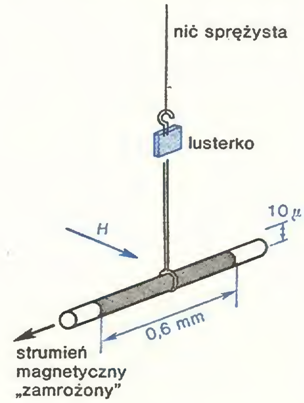
W wyniku braku oporu prąd elektryczny wzbudzony w zamkniętym obwodzie nadprzewodzącym może krążyć dowolnie długo bez jakichkolwiek zauważalnych strat. Taki trwały prąd można wzbudzić np. w pierścieniu z nadprzewodnika w następujący sposób: podnosi się temperaturę pierścienia powyżej jego punktu przejścia w stan normalny T_k i umieszcza się go w polu magnetycznym o liniach sił prostopadłych do płaszczyzny pierścienia; następnie obniża się temperaturę do $T < T_k$ i usuwa pole magnetyczne. Zmiana zewnętrznego pola magnetycznego wzbudzi w nadprzewodzącym pierścieniu niezanikający prąd elektryczny, który — zgodnie z regułą Lenza — przeciwdziałać będzie tej zmianie, tj. będzie podtrzymywać stały w czasie strumień magnetyczny w obszarze otoczonym przez pierścień (prąd wzbudzony w stanie normalnym szybko zanika).

Pierścień nadprzewodzący wraz z krążącym w nim prądem jest układem stacjonarnym, czyli trwałym w czasie. Według mechaniki kwantowej, stany stacjonarne są związane z pewnymi warunkami kwantowymi. Na przykład stacjonarne orbity elektronów w modelu atomu Bohra są związane z kwantowaniem momentu pędu elektronu. Można więc przypuszczać, że w pierścieniu nadprzewodzącym możliwe są tylko pewne dyskretne stany prądowe, określone warunkiem kwantowania jakiejś wielkości fizycznej związanej z prądem krążącym w pierścieniu. Taką wielkością podlegającą kwantyzacji jest strumień magnetyczny zawarty w otworze pierścienia. Po raz pierwszy przypuszczenie to wypowiedział F. London w 1950 r. Doszedł on do wniosku, że strumień magnetyczny uwięziony w otworze nadprzewodzącego pierścienia lub wydrążonego walca jest krotnością kwantu strumienia równego $\Phi_0 = h/q$, gdzie q jest ładunkiem elektrycznym nośnika prądu nadprzewodzonego. London sądził, że nośnikami prądu nadprzewodzonego są pojedyncze elektrony ($q = e$), czyli że $\Phi_0 = 4{,}14 \cdot 10^{-15}$ Wb. Po opracowaniu teorii BCS wyjaśniło się, że kwant strumienia magnetycznego, zwany także fluksonem, jest o połowę mniejszy, gdyż nośnikami prądu nadprzewodzonego są pary Coopera ($q = 2e$), a więc $\Phi_0 = h/2e = 2{,}07 \cdot 10^{-15}$ Wb. Jest to bardzo mała wielkość w porównaniu z tymi wartościami strumienia, z którymi najczęściej spotykamy się w praktyce. Jeśli założymy, że otwór w nadprzewodniku, w którym został uwięziony jeden flukson, ma powierzchnię przekroju poprzecznego równą 1 mm², to gęstość strumienia w tym otworze, tzn. indukcja magnetyczna B, będzie równa 10^{-9} T, czyli około sto tysięcy razy mniejsza niż gęstość strumienia ziemskiego pola magnetycznego. Tak słabe pole magnetyczne, jak pole ziemskie, wystarcza, aby po jego usunięciu uwięzić we wspomnianym otworze 10^5 kwantów strumienia magnetycznego. Zmiana tego strumienia o jeden lub kilka kwantów byłaby właściwie niewykrywalna.

flukson

pomiar fluksonu

Pomimo tak małej wartości, kwant strumienia został jednak doświadczalnie zmierzony. Nadprzewodnikiem, w którym uwięziono strumień magnetyczny, była cylindryczna warstewka ołowiu o długości 0,6 mm naparowana na kawałku kwarcowej kapilary o średnicy zewnętrznej 10 μm. Kwarcową kapilarę z warstewką ołowiu zawieszono za środkową jej



Rys. 3. Szkice układu do pomiaru kwantu strumienia magnetycznego

część na sprężystej nitce, do której przymocowano malutkie lusterko (rys. 3). Po zamrożeniu w nadprzewodniku stałego strumienia magnetycznego przyłożono dodatkowe słabe pole pod kątem prostym do rurki i mierzono moment skręcający wywierany na rurkę ze strony dodatkowego pola. Moment ten był proporcjonalny do wielkości zamrożonego strumienia. Stwierdzono, że ten strumień jest zawsze równy wielokrotności kwantu strumienia magnetycznego $\Phi_0 = h/2e$, co było potwierdzeniem teoretycznych przewidywań oraz koncepcji dotyczącej łączenia się elektronów w pary.

Kwantowanie strumienia magnetycznego jest jedną z podstawowych własności nadprzewodników i odgrywa bardzo ważną rolę w stanie mieszanym (o czym w dalszej części hasła) oraz w zjawiskach Josephsona. Analogiczne zjawisko kwantowania wielkości makroskopowej występuje w nadpłynnym helu, gdzie kwantuje się strumień wektora wirowości. Kwantowanie strumienia magnetycznego dowodzi wysokiego stopnia uporządkowania w układzie nośników prądu nadprzewodzonego. Pary Coopera nie mogą poruszać się niezależnie jedna od drugiej; wszystkie poruszają się zgodnie, a funkcje falowe opisujące ruch tych par mają jednakową długość i są uzgodnione fazowo czyli koherentne (spójne), przy czym koherencja obejmuje cały obszar nadprzewodnika.

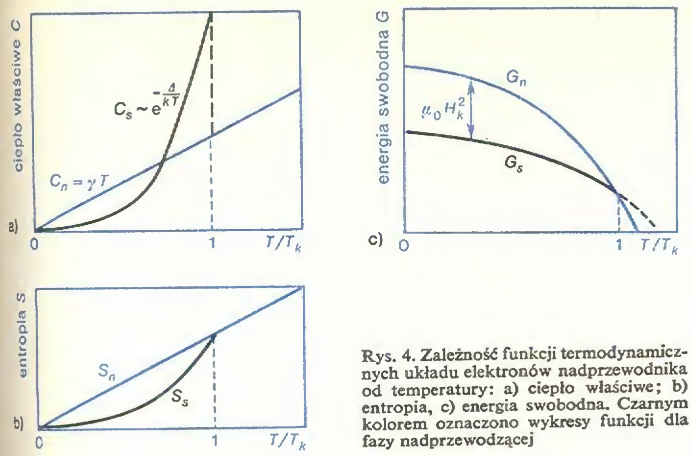
Entropia i energia swobodna nadprzewodnika

zależność entropii od temperatury

O tym, że po przejściu w stan nadprzewodnictwa zwiększa się uporządkowanie w układzie elektronów świadczy również zależność entropii nadprzewodnika od temperatury (rys. 4b). Poniżej punktu przejścia w stan nadprzewodzący T_k następuje szybki spadek entropii. Wartość entropii S układu elektronów przewodnictwa możemy znaleźć mierząc elektronowe ciepło właściwe C nadprzewodnika w funkcji temperatury i całkując otrzymaną zależność:

$$\int_0^T \frac{C(T)}{T}\, dT = S(T) - S(0). \tag{2}$$

Na mocy trzeciej zasady termodynamiki $S(0) = 0$.



Rys. 4. Zależność funkcji termodynamicznych układu elektronów nadprzewodnika od temperatury: a) ciepło właściwe; b) entropia, c) energia swobodna. Czarnym kolorem oznaczono wykresy funkcji dla fazy nadprzewodzącej

Znając entropię w podobny sposób możemy obliczyć energię swobodną:

$$\int_0^T S(T)dT = G(T) - G(0). \quad (3)$$

Energia swobodna w temperaturze $T = 0$, czyli $G(0)$ nie jest dla nas istotna, gdyż interesuje nas tylko zmiana energii swobodnej po przejściu w stan nadprzewodnictwa, czyli różnica $G_n(T) - G_s(T)$ (znaczki n i s odnoszą się do stanu normalnego i nadprzewodzącego). Jak widać na wykresie rys. 4c, dla $T < T_k$ stan nadprzewodzący jest stanem stabilnym, gdyż ma niższą energię swobodną niż stan normalny.

Energia kondensacji

Różnica energii swobodnych w stanie normalnym i nadprzewodzącym, czyli $G_n(T) - G_s(T)$, jest funkcją temperatury i nazywa się energią kondensacji. Jest to właśnie ta część energii, którą traci układ elektronów tworząc kondensat złożony z par Coopera. Jak już wspominaliśmy, w pary łączą się nie wszystkie elektrony, lecz tylko te, które w stanie normalnym zajmują przedział energetyczny o szerokości rzędu $^1/_2\Delta(T)$ w pobliżu poziomu Fermiego. Jeśli gęstość jednoelektronowych stanów w tym przedziale w normalnym stanie oznaczymy przez $N(E_F)$, to liczba utworzonych par wyniesie ok. $^1/_4 N(E_F)\Delta(T)$ na jednostkę objętości nadprzewodnika. Energia wiązania jednej pary jest równa $2\Delta(T)$, zatem energia kondensacji w $T = 0$ wyniesie

energia kondensacji

$$G_n(0) - G_s(0) = {}^1/_2 N(E_F)\Delta^2(0) \quad (4)$$

na jednostkę objętości (albo na atom lub na mol – zależnie od tego, jak określono $N(E_F)$). Stanowi to ok. 10^{-8} eV na atom, czyli ok. 10^{-4} J/cm^3 (są to tylko wartości orientacyjne: w zależności od rodzaju materiału mogą się one wahać o rząd wielkości lub więcej). Jest to bardzo mała energia w porównaniu z energią kinetyczną układu elektronów (ok. 10 eV na atom) lub z energią oddziaływania kulombowskiego. Mała wartość tej energii była główną przyczyną trudności przy opracowywaniu teorii nadprzewodnictwa, gdy trzeba było rozpatrywać małą różnicę dwu dużych wielkości. Należało więc uwzględnić tylko te efekty które były istotne dla nadprzewodnictwa, a pozostałe — jednakowe dla obu stanów — odrzucić przy rozpatrywaniu różnicy ich energii.

Odległość korelacji

Na granicy nadprzewodnika z materiałem normalnym (lub z próżnią) musi znikać uporządkowanie nadprzewodzące, a więc parametr uporządkowania musi zmniejszyć się od swej maksymalnej wartości, jaką ma w głębi nadprzewodnika, do zera na jego granicy. Za parametr uporządkowania możemy przyjąć np. względną gęstość par Coopera albo względną szerokość przerwy energetycznej, tzn. stosunek jej wartości lokalnej do wartości w głębi nadprzewodnika. Parametr uporządkowania nie znika nagle, lecz maleje stopniowo w warstwie o skończonej grubości. Grubość tej warstwy nazywamy odległością korelacji lub zasięgiem korelacji i oznaczamy literą ξ. Jest ona w przybliżeniu równa odegłości, z jakiej oddziałują z sobą (korelują) elektrony pary.

odległość (zasięg) koleracji

Odległość korelacji jest bardzo ważnym parametrem charakteryzującym materiał nadprzewodnika. Dla czystych metali odległość korelacji jest rzędu 10^{-4} cm, a więc przekracza około 10 000 razy odległość międzyatomową w krysztale. Elektrony nadprzewodnika korelują więc swe ruchy na stosunkowo dużej odległości. Domieszki lub defekty w nadprzewodniku zakłócają korelację międzyelektronową i zmniejszają odległość korelacji, dlatego stopy metali mają zwykle mały zasięg korelacji (niekiedy sto razy mniejszy niż czyste metale).

Materiały nadprzewodzące

Nadprzewodnictwo wykazuje wiele pierwiastków, głównie metalicznych, a także wiele stopów i związków chemicznych. Rozmieszczenie nadprzewodników w układzie okresowym pierwiastków przedstawiono w tabeli na następnej stronie. Liczby w tabeli w rubrykach pierwiastki oznaczają temperaturę przejścia pierwiastków w stan nadprzewodzący; litera p oznacza nadprzewodnictwo pod ciśnieniem, litera f — nadprzewodnictwo cienkiej warstwy osadzonej z par danego pierwiastka na podłoże chłodzone ciekłym helem, litery α, β, γ — fazy krystalograficzne. Niektóre pierwiastki nadprzewodzą tylko pod zwiększonym ciśnieniem (przechodzą wtedy w inną modyfikację krystalograficzną) lub w postaci cienkich warstw o niekrystalicznej strukturze, uzyskanych przez naparowanie wyjściowego materiału na podłoże oziębione ciekłym helem. Nadprzewodnictwo takich warstw świadczy o tym, że struktura krystaliczna nie jest konieczným warunkiem pojawienia się nadprzewodnictwa. Nie są nadprzewodnikami metale ferromagnetyczne i metale pierwszej grupy układu okresowego (oprócz cezu pod ciśnieniem). Wśród nadprzewodników spotyka się pierwiastki krystalizujące w różnych układach, przy czym różne modyfikacje krystalograficzne tego samego pierwiastka mają na ogół różne temperatury przejścia w stan nadprzewodnictwa. Dowodzi to, że nadprzewodnictwo nie jest cechą określonych atomów, lecz zależy od struktury sieci krystalicznej.

pierwiastki

W obecnym stanie wiedzy nie możemy przewidzieć, czy dany pierwiastek może być nadprzewodnikiem, czy nie. Możemy tylko wykluczyć nadprzewodnictwo ferromagnetyków, gdyż silne wewnętrzne pole magnetyczne w tych substancjach niszczy uporządkowanie nadprzewodzące na rzecz uporządkowania magnetycznego. Nie ma jednak zasadniczych przeszkód dla pojawienia się nadprzewodnictwa np. w metalach pierwszej grupy układu okresowego. Być może, temperatury ich przejścia w stan nadprzewodnictwa są zbyt niskie, niższe niż te, w których były one badane.

Obecnie znamy 38 pierwiastków i około tysiąca stopów i związków, które stają się nadprzewodnikami. Lista materiałów nadprzewodzących nie jest jeszcze zakończona, gdyż nadal odkrywane są coraz to nowe nadprzewodniki. Wśród związków nadprzewodzących wiele jest takich, w których żaden ze składników z osobna nie jest nadprzewodnikiem (np. CuS). Wśród nadprzewodników spotykamy nawet substancje organiczne, jak np. anilina czy pirydyna, w postaci monomolekularnych warstw rozdzielonych chalkogenidkami niektórych metali.

stopy i związki

Rozmieszczenie nadprzewodników w układzie okresowym pierwiastków

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
H																	He
Li	Be 0,03 9,6(f)											B	C	N	O	F	Ne
Na	Mg											Al 1,2 5(p)	Si 7(p)	P 5,4(p)	S	Cl	Ar
K	Ca	Sc	Ti 0,39	V 5,3	Cr	Mn	Fe	Co	Ni	Cu	Zn 0,88	Ga 1,09 6,2β 7,8γ	Ge 5,5(p)	As	Se 7(p)	Br	Kr
Rb	Sr	Y 2,5(p)	Zr 0,55	Nb 9,2	Mo 0,92 5(f)	Tc 7,7	Ru 0,5	Rh	Pd	Ag	Cd 0,56	In 3,4	Sn 3,7	Sb 2,6(p)	Te 3,3(p)	J	Xe
Cs 1,5(p)	Ba 5(p)	La 4,9α 6,1β 12(p)	Hf 0,16	Ta 4,5	W 0,012 4(f)	Re 1,7 7(f)	Os 0,7	Ir 0,14	Pt	Au	Hg 4,1α 3,9β	Tl 2,4	Pb 7,2	Bi 3,9(p) 7,3(p) 6(f)	Po	At	Rn
Fr	Ra	Ac															

Ce 1,7(p)	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
Th 1,37	Pa 1,3	U 0,2α 1,8γ	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lw

Nadprzewodnik w polu magnetycznym

Zjawisko Meissnera

Przez długi czas po odkryciu nadprzewodnictwa sądzono, że nadprzewodnik to jedynie doskonały przewodnik, tzn. przewodnik pozbawiony oporu elektrycznego. Dopiero w 1933 r., a więc 22 lata po odkryciu znikania oporu, W. Meissner i R. Ochsenfeld za pomocą małej ceweczki pomiarowej wyznaczyli rozkład linii sił pola magnetycznego dokoła kuli nadprzewodzącej i odkryli drugą podstawową własność nadprzewodnika: linie sił pola zewnętrznego zawsze omijają nadprzewodzącą bryłę. Niezależnie od tego, czy pole magnetyczne zostało nałożone przed przejściem w stan nadprzewodnictwa, czy po przejściu, indukcja magnetyczna w obszarze nadprzewodnika jest zawsze równa zeru. Nadprzewodnik jest więc nie tylko doskonałym przewodnikiem, ale także doskonałym diamagnetykiem.

odkrycie Meissnera i Ochsenfelda

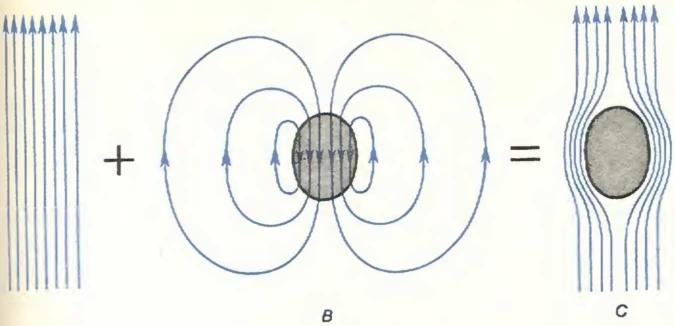
Bryła substancji, która byłaby tylko doskonałym przewodnikiem, po nałożeniu pola magnetycznego i następnym oziębieniu poniżej temperatury przejścia T_k w stan bezoporowy zachowałaby to pole w swej objętości nawet po usunięciu pola zewnętrznego, podobnie jak pierścień nadprzewodzący lub wydrążony walec zachowuje pole magnetyczne w swym otworze. Zgodnie z regułą Lentza, każda zmiana pola zewnętrznego po przejściu bryły w stan bezoporowy wzbudzałaby na jej powierzchni trwały prąd elektryczny, podtrzymujący wewnątrz bryły stałą w czasie wartość indukcji magnetycznej. Indukcja magnetyczna wewnątrz doskonałego przewodnika byłaby więc stale równa tej jej wartości, jaka istniała w chwili przejścia w stan bezoporowy.

Nadprzewodnik tym różni się od przewodnika doskonałego, że w chwili przejścia w stan bezoporowy wypycha ze swej objętości istniejące tam poprzednio pole magnetyczne, tak że $B = 0$ niezależnie od początkowej jej wartości. Takie zachowanie się nadprzewodnika nie daje się wyjaśnić na gruncie elektrodynamiki klasycznej. Teorię uwzględniającą zachowanie się nadprzewodników w polu magnetycznym, czyli zjawisko Meissnera, opracowali bracia F. i H. Londonowie w 1935 roku.

wypychanie pola magnetycznego przez nadprzewodnik

Wypychanie strumienia magnetycznego z wnętrza nadprzewodzącej bryły odbywa się w ten sposób, że w chwili przejścia w stan nadprzewodnictwa na powierzchni nadprzewodnika wzbudza się trwały prąd elektryczny, który wytwarza własne pole magnetyczne kompensujące do zera pole magnetyczne istniejące poprzednio wewnątrz bryły. Ten powierzchniowy prąd elektryczny nazywa się prądem Meissnera lub prądem ekranującym, gdyż jak gdyby ekranuje

on wnętrze nadprzewodnika od pola zewnętrznego. Na zewnątrz nadprzewodzącej bryły pole magnetyczne pochodzące od prądu Meissnera nakłada się



Rys. 5. Nadprzewodnik w polu magnetycznym: *A* linie sił pola zewnętrznego przed umieszczeniem w nim nadprzewodnika, *B* linie sił pola magnetycznego wytworzonego przez prąd Meissnera po umieszczeniu nadprzewodnika w polu zewnętrznym, *C* linie sił wypadkowego pola magnetycznego powstałego przez nałożenie pola prądów Meissnera na pole zewnętrzne

na pole pierwotne i tworzy wspólne pole wypadkowe; linie tego pola wypadkowego opływają nadprzewodzącą bryłę.

Rys. 5 przedstawia elipsoidalną bryłę nadprzewodnika umieszczoną w jednorodnym polu magnetycznym. Widać, jak nadprzewodnik zmienia to pole zewnętrzne i powoduje, że największa gęstość linii sił pola (czyli najsilniejsze pole) występuje na „równiku" elipsoidy (rys. 5c). Im większa długość elipsoidy w kierunku pola w porównaniu z jej średnicą równikową, tym mniejszy stopień zniekształcenia pola zewnętrznego. Gdy elipsoida jest nieskończenie długa lub gdy mamy nieskończenie długi walec o osi równoległej do linii sił pola, zniekształcenie pola zewnętrznego w ogóle nie nastąpi. Nieskończenie długi walec nadprzewodzący, równoległy do linii sił pola, wraz z prądem Meissnera na jego powierzchni możemy potraktować jak nieskończenie długi, gęsto nawinięty solenoid. Pole magnetyczne takiego solenoidu, wytworzone przez prąd płynący w jego uzwojeniach, jest jednorodne wewnątrz solenoidu i równe zeru na zewnątrz. Nieskończenie długi solenoid nie zaburza więc postronnego pola zewnętrznego. W praktyce za nieskończenie długi walec możemy uważać taki walec, którego długość jest bardzo wielka w porównaniu z jego średnicą.

Zjawisko Meissnera ma bardzo istotne znaczenie z punktu widzenia termodynamiki. Dzięki temu zjawisku przejście w stan nadprzewodnictwa jest przejściem odwracalnym, może więc być potraktowane jako przemiana fazowa. Stan nadprzewodnictwa cechuje zerowa wartość indukcji magnetycznej niezależnie od tego, w jaki sposób stan ten został osiągnięty: czy najpierw oziębiono substancję do temperatury niższej od temperatury przemiany w stan nadprzewodnictwa, a później przyłożono pole magnetyczne, czy też najpierw umieszczono substancję w polu magnetycznym, a potem ją oziębiono. Stan nadprzewodnictwa jest więc odrębną termodynamiczną fazą substancji. Doskonałego przewodnika nie moglibyśmy w ten sposób traktować, gdyż jego stan po utracie oporu zależałby od tego, czy był oziębiany w polu magnetycznym, czy nie.

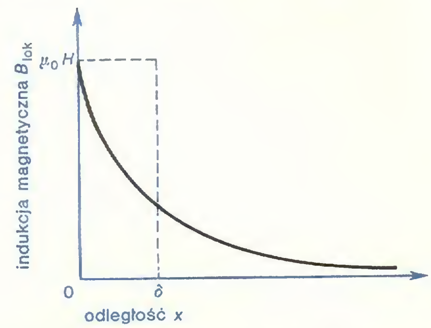
stan nadprzewodnictwa

Głębokość wnikania

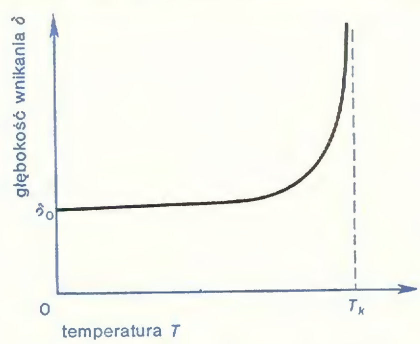
Prąd Meissnera, kompensujący wewnątrz nadprzewodnika zewnętrzne pole magnetyczne, płynie po powierzchni nadprzewodnika w warstwie o pewnej grubości. W tej powierzchniowej warstwie indukcja magnetyczna jest różna od zera. Możemy więc uważać, że pole zewnętrzne wnika w powierzchniową warstwę nadprzewodnika na pewną głębokość, nazywaną po prostu głębokością wnikania.

efektywna głębokość wnikania

W warstwie powierzchniowej indukcja magnetyczna zmienia się od wartości, jaka istnieje na zewnątrz, do zera w głębi nadprzewodnika. Jeśli, jak to wynika z elektrodynamiki Londonów, indukcja maleje wykładniczo ze wzrostem odległości x od powierzchni, czyli $B(x) = B(0)e^{-\delta/x}$, to w odległości $x = \delta$ zmaleje ona e $\approx$ 2,718... razy w stosunku do tej wartości, jaką miała na powierzchni próbki. Odległość δ można uważać za efektywną głębokość wnikania, tj. taką odległość, do której indukcja nie zmienia się wcale, a po jej przekroczeniu (czyli dla $x > \delta$) indukcja jest już zerem. Krzywą wykładniczą zastępujemy w ten sposób linią łamaną (rys. 6).



Rys. 6. Zależność lokalnej indukcji magnetycznej B_{lok} wewnątrz nadprzewodnika od odległości x od powierzchni nadprzewodnika



Rys. 7. Zależność głębokości wnikania pola od temperatury

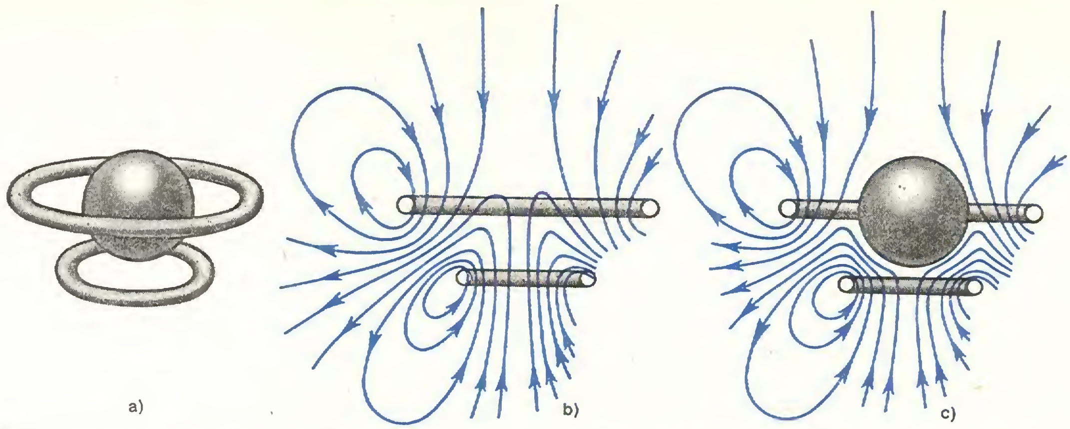
zależność δ od temperatury

Głębokość wnikania δ zależy od temperatury; gdy temperatura zbliża się do T_k, głębokość wnikania rośnie do nieskończoności (rys. 7). Oznacza to, że w miarę zbliżania się do T_k pole magnetyczne stopniowo wnika coraz głębiej w nadprzewodnik i w chwili przejścia w stan normalny cała objętość nadprzewodnika jest już zajęta przez pole. Głębokość wnikania w temperaturze zera bezwzględnego δ_0 jest ważną stałą charakteryzującą dany nadprzewodnik. Podobnie jak zasięg korelacji ξ, głębokość wnikania δ_0 zależy od czystości materiału, z tym, że δ_0 zwiększa się z koncentracją domieszek. Dla czystych metali $\delta_0 \approx 10^{-6}$ cm; dla niektórych stopów może być nawet o dwa rzędy większa.

Odpychanie nadprzewodnika przez pole magnetyczne

Doskonały diamagnetyzm jest przyczyną wypychania nadprzewodnika z obszaru najsilniejszego pola. Odwrotnie bywa w przypadku ferromagnetyków: są one wciągane w obszar najsilniejszego pola (np. rdzeń żelazny jest wciągany do cewki, w której płynie prąd elektryczny). Powszechnie znane jest doświadczenie z kulą nadprzewodzącą unoszącą się nad pierścieniami, w których krążą niezanikające prądy elektrycz-

ne (wzbudzone w przeciwnych kierunkach); prądy te wytwarzają pole magnetyczne odpychające nadprzewodzącą kulę. Rysunek 8 ilustruje omawiane zjawisko, rys. 8b pokazuje konfigurację pola magnetycznego wytworzonego przez oba pierścienie przed umieszczeniem nad nimi kuli, rys. 8c — po umieszczeniu kuli

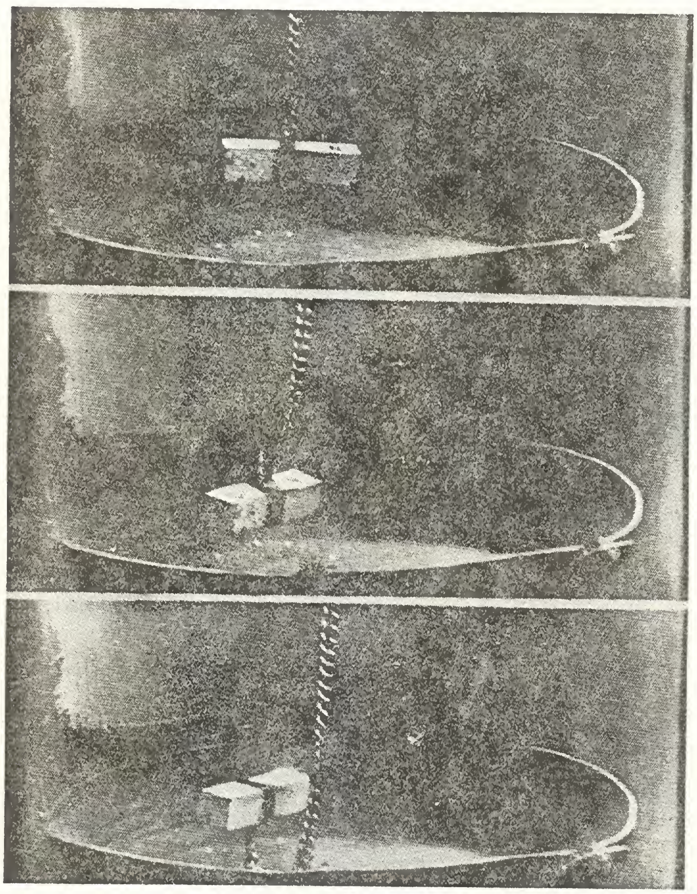


Rys. 8. Unoszenie kuli nadprzewodzącej przez pole magnetyczne: a) widok ogólny, b) linie sił pola magnetycznego wytworzonego przez niezanikające prądy płynące w pierścieniach, c) linie sił po umieszczeniu kuli nadprzewodzącej nad pierścieniami

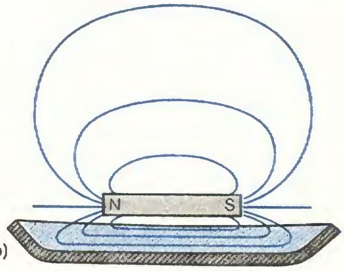
nad pierścieniami. Pole magnetyczne utrzymuje kulę równoważąc siłę ciężkości.

pływający magnes

Innym wariantem takiego doświadczenia jest tzw. pływający magnes (rys. 9). Magnes sztabkowy opuszczony jest nad czaszą nadprzewodzącą. Linie sił pola magnetycznego pochodzące od tego magnesu unoszą go nad czaszą. Zjawisko to tłumaczymy odpychaniem nadprzewodzącej czaszy przez linie sił pola magnetycznego. Jeśli czasza jest sztywno umocowana w kriostacie, to siła odpychania unosi magnes.



a)

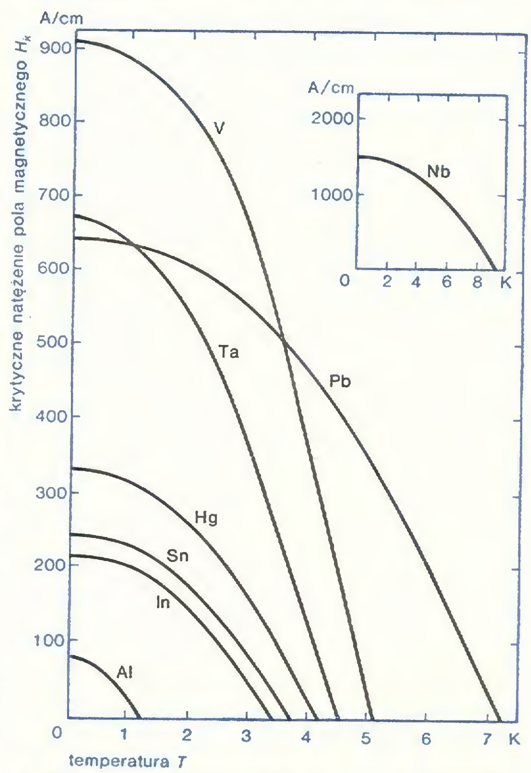


b)

Rys. 9. Magnes unoszący się nad czaszą nadprzewodzącą: a) zdjęcie, b) linie sił pola magnesu

Krytyczne pole magnetyczne

Jeśli natężenie pola magnetycznego, w którym umieszczono nadprzewodnik, przekroczy pewną wartość



Rys. 10. Zależność krytycznych natężeń pól magnetycznych od temperatury dla niektórych pierwiastków nadprzewodzących

krytyczną H_k — nadprzewodnictwo zniknie, aby powrócić znowu, gdy natężenie pola zmniejszy się do wartości $H < H_k$. Krytyczna wartość natężenia pola magnetycznego dla danego nadprzewodnika zależy od temperatury (rys. 10). Zależność tę można w przybliżeniu wyrazić następującym wzorem:

$$H_k(T) = H_k(0)[1-(T/T_k)^2]. \tag{5}$$

wykres stanu nadprzewodnika

Wykres $H_k(T)$ jest wykresem stanu dla nadprzewodnika. Krzywa $H_k(T)$ rozdziela ćwierćpłaszczyznę T–H na dwa obszary fazowe: obszar fazy nadprzewodzącej (pod krzywą) i obszar fazy normalnej (reszta ćwierćpłaszczyzny). Krzywa $H_k(T)$ jest krzywą równowagi fazowej, tzn. w stanie odpowiadającym punktowi (T_0, H_0) leżącemu na krzywej $H_k(T)$ mogą istnieć obie fazy obok siebie w równowadze. Wynika to stąd, że dla $T < T_k$ obu fazom odpowiadają różne wartości entropii (rys. 4b), a więc przejście w stan nadprzewodnictwa w polu magnetycznym jest przemianą fazową I rodzaju. Utajone ciepło przemiany wynosi $Q = T(S_n - S_s)$. Jeśli $H = 0$, to przemiana zachodzi w $T = T_k$ i ciepło przemiany jest wtedy równe zeru, gdyż $S_n(T_k) = S_s(T_k)$. W nieobecności pola magnetycznego przejście w stan nadprzewodnictwa jest więc przemianą fazową II rodzaju. Świadczy o tym także kształt krzywej zależności ciepła właściwego od temperatury (rys. 5a) zbliżony do kształtu litery λ.

Namagnesowanie nadprzewodnika

Jak widać na rys. 5b, nadprzewodnik w polu magnetycznym magnesuje się, czyli przybiera określony moment magnetyczny. Jako diamagnetyk, nadprzewodnik magnesuje się w kierunku przeciwnym do przyłożonego pola magnetycznego, a więc jego moment magnetyczny jest ujemny. Oczywiście, moment magnetyczny nadprzewodnika pochodzi od niezanikających prądów Meissnera, krążących po jego powierzchni. Jak wiadomo, płaska pętla prądu elektrycznego ma dipolowy moment magnetyczny równy $M = IA$, gdzie I jest natężeniem prądu w pętli, A — polem płaskiej powierzchni otoczonej tą pętlą. Na moment magnetyczny nadprzewodnika składają się momenty dipolowe wszystkich płaskich pętli prądowych, na jakie można by podzielić powierzchniowy prąd Meissnera.

zależność namagnesowania od pola

W ustalonych warunkach moment magnetyczny nadprzewodzącej bryły jest proporcjonalny do jej objętości, można więc posługiwać się pojęciem namagnesowania, czyli momentu magnetycznego przypadającego na jednostkę objętości. Początkowo namagnesowanie wzrasta liniowo ze wzrostem natężenia pola zewnętrznego, gdyż zwiększa się gęstość powierzchniowa prądów Meissnera, ekranujących wnętrze bryły od tego pola. Gdy natężenie pola przekroczy krytyczną wartość H_k, nadprzewodnictwo znika i namagnesowanie spada do zera. Taka zależność namagnesowania od natężenia pola magnetycznego jest typowa dla nadprzewodników i wyróżnia je spośród innych materiałów magnetycznych. W przeciwieństwie do namagnesowania, brak oporu elektrycznego nie zawsze stanowi wiarygodne kryterium nadprzewodnictwa, gdyż niekiedy bywa wynikiem obecności domieszek materiału nadprzewodzącego w nienadprzewodzącej próbce. Domieszki te mogą utworzyć nadprzewodzące ścieżki, zwierające „na krótko" resztę materiału.

Zmiana energii swobodnej w polu magnetycznym

Magnesowanie się nadprzewodnika w polu magnetycznym (czyli wzbudzanie prądów Meissnera) zwiększa jego energię swobodną o $\frac{1}{2}\mu_0 H^2$ na jednostkę objętości. Ponieważ zmianę energii dG, spowodowaną magnesowaniem, obliczamy z zależności:

$$dG = -MdH, \quad \text{przeto} \quad G(H_0) - G(0) = -\int_0^{H_0} MdH.$$

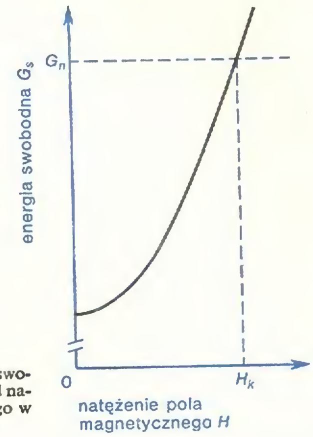
Dla nadprzewodnika w stanie Meissnera $\mu = -1$, czyli $M = -HV$; stąd:

$$G_s(H_0) = G_s(0) + \frac{1}{2}\mu_0 H_0^2 V, \tag{6}$$

gdzie V jest objętością bryły nadprzewodzącej. Jeśli pominiemy słaby paramagnetyzm (lub diamagnetyzm) normalnego metalu (a ferromagnetyki nie są nadprzewodnikami), to energia swobodna w stanie normalnym nie będzie zależeć od natężenia pola magnetycznego, czyli $G_n(H_k) = G_n(0)$.

zależność energii swobodnej od pola

Gdy natężenie pola zewnętrznego osiągnie wartość H_k, energia swobodna w stanie nadprzewodzącym wzrośnie na tyle, że zrówna się z energią swobodną



Rys. 11. Zależność energii swobodnej nadprzewodnika od natężenia pola magnetycznego w stałej temperaturze

w stanie normalnym (rys. 11). Przy dalszym zwiększaniu natężenia pola utrzymywanie stanu nadprzewodnictwa byłoby już niekorzystne, gdyż $G_s(H) > G_n(H)$ dla $H > H_k$, dlatego przy $H = H_k$ nadprzewodnik przechodzi w stan normalny. Krytyczne natężenie pola magnetycznego H_k jest to takie natężenie, przy którym gęstości energii swobodnych w obu fazach, normalnej i nadprzewodzącej, są sobie równe. To stwierdzenie można uznać za definicję krytycznego natężenia pola magnetycznego H_k. Mamy więc:

krytyczne natężenie pola magnetycznego

$$G_s(H_k) = G_s(0) + \frac{1}{2}\mu_0 H_k^2 V = G_n(H_k),$$

a ponieważ $G_n(H_k) = G_n(0)$, to

$$G_n(0) - G_s(0) = \frac{1}{2}\mu_0 H_k^2 V. \tag{7}$$

Wynika z tego, że w nieobecności pola magnetycznego zmiana (spadek) energii spowodowanej po przejściu w stan nadprzewodnictwa w wyniku obniżenia temperatury (czyli energia kondensacji) wynosi $\frac{1}{2}\mu_0 H_k^2(T)$ na jednostkę objętości. Porównując wyrażenia (4) i (7) otrzymujemy zależność:

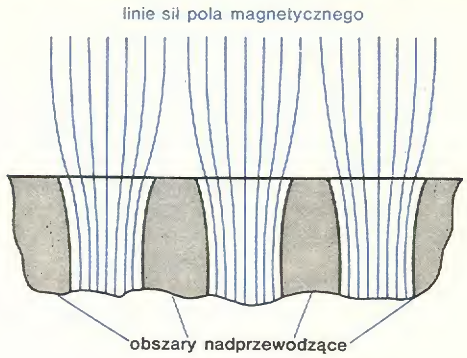
$$H_k(T) = \Delta(T)\big(N(E_F)/\mu_0\big)^{1/2}. \tag{8}$$

Stan pośredni

Jeśli próbka nadprzewodnika nie jest walcem równoległym do linii sił zewnętrznego pola magnetycznego, to próbka taka zmienia w swoim otoczeniu rozkład linii sił, a na jej powierzchni natężenie pola nie wszędzie ma jednakową wartość. Gdy nadprzewodnik ma kształt elipsoidy obrotowej skierowanej osią symetrii wzdłuż linii sił pola zewnętrznego (jak na rys. 6), natężenie pola ma większą wartość na równiku elipsoidy, a mniejszą na biegunach. Jeśli więc natężenie pola zwiększa się z czasem, to najwcześniej krytyczną wartość osiągnie ono na równiku elipsoidy i pas równikowy najpierw powinien przejść w stan normalny. Nie jest to jednak możliwe, gdyż wtedy pole wniknęłoby w obszar normalny i natężenie pola zmalałoby

obszary normalne i nadprzewodzące

tam do wartości $H < H_k$; obszar pasa równikowego musiałby więc odzyskać nadprzewodnictwo, a to przywróciłoby wyjściową sytuację. Jedynym możliwym rozwiązaniem jest podział całej objętości próbki na warstwowe obszary fazy normalnej rozdzielone



Rys. 12. Obszary normalne i nadprzewodzące w stanie pośrednim

obszarami fazy nadprzewodzącej (rys. 12). Pole magnetyczne przenika próbkę poprzez obszary normalne. Kształt obszarów normalnych wychodzących na powierzchnię próbki możemy obserwować wizualnie po posypaniu próbki proszkiem ferromagnetycznym; proszek osiądzie na obszarach normalnych, gdyż tam $H \neq 0$ (il. 94, tabl. 24).

natężenie krytyczne dla stanu pośredniego

Stan, w jakim znajduje się próbka po podziale na współistniejące obok siebie obszary fazy normalnej i nadprzewodzącej, nazywamy stanem pośrednim. Natężenie zewnętrznego pola magnetycznego H_a, przy którym próbka przechodzi w stan pośredni, zależy od jej kształtu i orientacji względem linii sił, czyli od współczynnika rozmagnesowania D:

$$H_a = H_k(1-D). \tag{9}$$

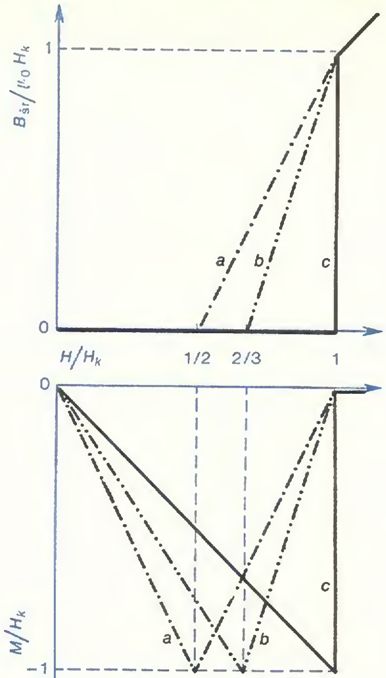
Dla kuli $D = {}^1/_3$ i $H_a = {}^2/_3 H_k$, dla walca prostopadłego do pola $H_a = {}^1/_2 H_k$; walec równoległy do pola nie zmienia rozkładu linii sił i przechodzi bezpośrednio w stan normalny, gdy $H_a = H_k$.

Makroskopowe rozmiary obszarów poszczególnych faz w stanie pośrednim świadczą o istnieniu dodatniej energii związanej z powierzchnią rozgraniczającą te obszary; w przeciwnym razie korzystniejszy byłby bardzo drobny podział, gdyż wtedy przy powierzchni próbki byłoby mniejsze skrzywienie linii sił pola przenikającego przez obszary normalne. W miarę wzrostu natężenia pola od wartości $H_k(1\ \ D)$ do H_k zwiększa się w próbce udział fazy normalnej kosztem fazy nadprzewodzącej. Gdy natężenie pola osiągnie wartość H_k, obszary nadprzewodzące znikają zupełnie.

Wewnątrz próbki w stanie pośrednim indukcja magnetyczna w obszarach fazy nadprzewodzącej jest równa zeru, a w obszarach fazy normalnej $B = \mu_0 H_k$. Indukcja średnia dla całej próbki, tj. uśredniona po całej objętości próbki, wynosi $B_{śr} = \mu_0 H_k \cdot x_n$, gdzie x_n jest częścią objętości próbki zajętą przez fazę normalną. $B_{śr}$ jest liniową funkcją H, czyli udział fazy normalnej jest proporcjonalny do H.

moment diamagnetyczny w stanie pośrednim

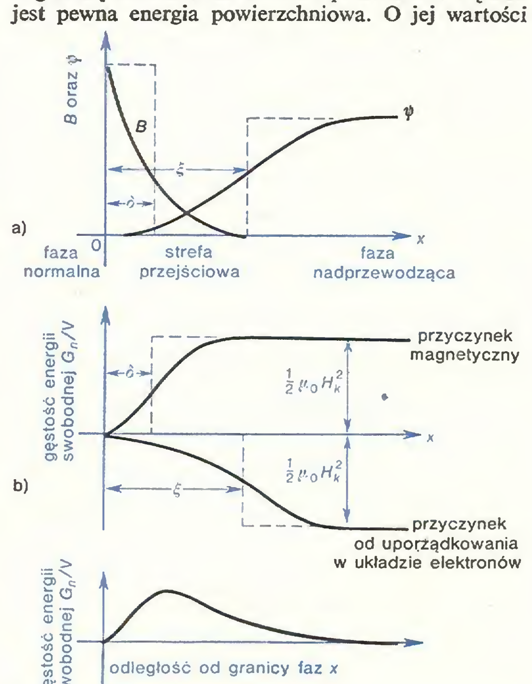
Moment diamagnetyczny próbki w stanie pośrednim jest momentem magnetycznym nadprzewodzących obszarów, do których pole nie wnika (tylko one są ekranowane od pola). Rys. 13 przedstawia zależności indukcji średniej i momentu magnetycznego od natężenia pola zewnętrznego dla trzech próbek nadprzewodnika o różnym współczynniku rozmagnesowania. Zauważmy, że pole powierzchni zawartej między krzywą momentu magnetycznego a osią H jest jednakowe dla wszystkich próbek niezależnie od ich współczynnika rozmagnesowania, jeśli tylko mają one jednakową objętość i jednakową wartość H_k. Pole to przedstawia pracę wykonaną przy magnesowaniu próbki, czyli wzrost energii próbki po nałożeniu pola magnetycznego $H \geqslant H_k$. Wiemy już, że ten przyrost energii jest równy ${}^1/_2 \mu_0 H_k^2$ na jednostkę objętości próbki.



Rys. 13. Zależność średniej indukcji magnetycznej i namagnesowania od natężenia pola magnetycznego dla próbek nadprzewodnika o różnym współczynniku kształtu D; a — kula, $D = {}^1/_3$; b — walec poprzeczny do linii sił pola, $D = {}^1/_2$; c — walec równoległy do linii sił pola, $D = 0$

Energia powierzchniowa i dwa typy nadprzewodnictwa

Z granicą rozdziału faz w stanie pośrednim związana jest pewna energia powierzchniowa. O jej wartości



Rys. 14. Szkic ilustrujący powstawanie energii powierzchniowej między fazą normalną i nadprzewodzącą

decyduje stosunek wzajemny obu charakterystycznych wielkości w nadprzewodniku, tj. głębokości wnikania δ pola magnetycznego i zasięgu korelacji ξ. Rozpatrzmy, jakie jest pochodzenie tej energii.

Linia pionowa na rys. 14 oznacza granicę rozdzielającą obszar fazy nadprzewodzącej od obszaru fazy normalnej. Parametr uporządkowania nadprzewodzącego nie zmienia się nagle na tej granicy, lecz maleje stopniowo w miarę zbliżania się do obszaru fazy normalnej, przy czym odległość, na której następuje zmiana parametru uporządkowania od wartości maksymalnej do zera jest w przybliżeniu równa zasięgowi korelacji ξ.

W głębi fazy nadprzewodzącej uporządkowanie zmniejsza energię swobodną nadprzewodnika o energię kondensacji, czyli o wartość równą $^1/_2\mu_0 H_k^2$ na jednostkę objętości. Zmiana gęstości energii (czyli energii na jednostkę objętości) w warstwie przygranicznej, wynikająca z uporządkowania elektronów w układzie, jest przedstawiona na rys. 14b. Ponieważ stopień uporządkowania w warstwie przygranicznej jest mniejszy niż w głębi obszaru fazy nadprzewodzącej — gęstość energii swobodnej nie zmniejsza się w tej warstwie o $^1/_2\mu_0 H_k^2$, a więc warstwa ta wnosi efektywny wkład do energii fazy nadprzewodzącej, wynoszący $^1/_2\mu_0 H_k^2$ na jednostkę objętości warstwy, czyli $^1/_2\mu_0 H_k^2\xi$ na jednostkę powierzchni granicznej (skuteczna grubość warstwy wynosi ξ).

Istnieje jeszcze jeden czynnik wpływający na gęstość energii swobodnej w warstwie przygranicznej: jest to pole magnetyczne wnikające w obszar fazy nadprzewodzącej na głębokość skuteczną δ. Pole magnetyczne o natężeniu H zwiększa gęstość energii swobodnej fazy nadprzewodzącej o $^1/_2\mu_0 H^2$, ale w warstwie przygranicznej o grubości δ pole nie powoduje wzrostu gęstości energii, gdyż wnika w tę warstwę. Jest to równoważne zmniejszeniu energii swobodnej obszaru fazy nadprzewodzącej o wartość $^1/_2\mu_0 H_k^2\delta$ na jednostkę powierzchni granicy rozdziału faz (natężenie pola przyjmujemy tu za równe H_k, gdyż taką wartość ma ona na granicy międzyfazowej w stanie pośrednim; jest to warunek stabilności granicy).

W głębi fazy nadprzewodzącej oba wkłady w energię swobodną (wkład od pola magnetycznego i wkład od energii kondensacji) znoszą się wzajemnie i gęstość energii fazy nadprzewodzącej jest w stanie pośrednim równa gęstości energii fazy normalnej, tak jak w polu o natężeniu krytycznym. Na granicy faz przeważa zwykle jeden z wkładów, gdyż na ogół długości ξ i δ nie muszą być sobie równe (rys. 14c). Dodatkowa energia związana z granicą rozdziału faz jest więc równa $^1/_2\mu_0 H_k^2(\xi-\delta)$ na jednostkę powierzchni granicznej. Jeśli $\xi > \delta$, to energia powierzchniowa jest dodatnia i takie nadprzewodniki nazywane są nadprzewodnikami I typu; jeśli $\xi < \delta$, to energia powierzchniowa jest ujemna i nadprzewodniki takie należą do drugiej grupy, czyli do nadprzewodników II typu.

nadprzewodniki I typu

nadprzewodniki II typu

Nadprzewodnikami I typu są prawie wszystkie pierwiastki nadprzewodzące (z wyjątkiem Nb, V i Tc), oraz niektóre stopy sporządzone na bazie takich metali, jak In, Al, Hg, Sn, przez rozpuszczenie w nich niewielkiej domieszki drugiego składnika. Większość stopów nadprzewodzących, a w szczególności stopy na bazie metali przejściowych, oraz pierwiastki Nb, V i Tc należą do nadprzewodników II typu. Nadprzewodniki obu typów różnią się między sobą pod względem termodynamicznym i — w związku z tym — inaczej zachowują się w polu magnetycznym. Opisywane tutaj dotychczas własności magnetyczne dotyczyły nadprzewodników I typu. Nadprzewodniki I typu zostały najwcześniej zbadane, gdyż stosunkowo łatwo jest otrzymać jednorodne próbki czystych metali. Badania tych nadprzewodników dostarczyły informacji, na podstawie których Bardeen, Cooper i Schrieffer opracowali w 1957 r. mikroskopową teorię nadprzewodnictwa.

Nadprzewodniki II typu

Teorię nadprzewodników II typu opracował A. Abrikosow również w 1957 r., ale potwierdzające ją badania doświadczalne udało się przeprowadzić dopiero kilka lat później, gdy otrzymano wystarczająco jednorodne próbki nadprzewodzących stopów (teoria dotyczyła tylko jednorodnych nadprzewodników). Zainteresowanie nadprzewodnikami II typu wzrosło wyraźnie po 1961 r., kiedy odkryto, że niejednorodne próbki tych nadprzewodników mogą przenosić prąd elektryczny o dużej gęstości (rzędu 10^5 A/cm²) nawet w silnych polach magnetycznych (rzędu kilku tesli).

Teoria Abrikosowa opierała się na wcześniejszej fenomenologicznej teorii Ginzburga i Landaua, tj. teorii nie wnikającej w przyczyny nadprzewodnictwa, lecz opisującej zjawisko od strony makroskopowej. Później Gorkow wykazał, że teoria ta wynika również z teorii mikroskopowej. Od pierwszych liter nazwisk jej twórców teorię tę nazywamy obecnie teorią GLAG, podobnie jak teoria mikroskopowa nazywa się teorią BCS. Teoria GLAG podaje ścisłe kryterium podziału nadprzewodników na dwa typy: jest nim wartość parametru wprowadzonego w tej teorii, zwanego parametrem $\varkappa$; jeśli $\varkappa < 1/\sqrt{2}$, to nadprzewodnik należy do pierwszego typu, jeśli $\varkappa > 1/\sqrt{2}$ — do drugiego typu. Parametr $\varkappa$ jest w przybliżeniu równy stosunkowi δ/ξ.

teoria GLAG

Oprócz nadprzewodników doskonałych obu wspomnianych typów istnieją nadprzewodniki niedoskonałe, czyli niejednorodne lub zdeformowane. Podział nadprzewodników na doskonałe i niedoskonałe nie jest zbyt ścisły, gdyż w praktyce prawie zawsze spotykamy się z pewną niedoskonałością nadprzewodnika. Za miarę doskonałości próbki nadprzewodnika przyjęto uważać stopień odwracalności jego przemiany fazowej w polu magnetycznym. Ze względu na zdolność przenoszenia prądów o dużej gęstości niektóre niedoskonałe nadprzewodniki II typu odgrywają bardzo ważną rolę w zastosowaniach praktycznych. Niektórzy nazywają takie nadprzewodniki nadprzewodnikami III typu, ale nazwa ta nie wydaje się uzasadniona, gdyż nie chodzi tu o jakiś nowy typ nadprzewodnictwa. Bardziej stosowną byłaby nazwa „twarde nadprzewodniki", nawiązująca do ich własności mechanicznych.

nadprzewodniki niejednorodne lub zdeformowane

Zachowanie się nadprzewodników II typu w polu magnetycznym

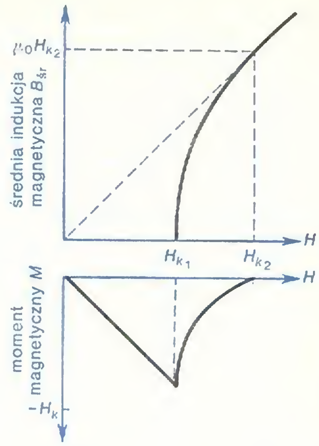
Prześledźmy, jak zmienia się moment magnetyczny M nadprzewodnika II typu ze zmianą natężenia zewnętrznego pola magnetycznego H (rys. 15). Dla uproszczenia załóżmy, że próbka ma współczynnik rozmagnesowania równy zeru i nie zawiera dziur.

W słabym polu magnetycznym próbka nadprzewodnika II typu zachowuje się tak samo, jak próbka nadprzewodnika I typu, tzn. wypycha całkowicie ze swego wnętrza strumień magnetyczny, czyli wykazuje pełne zjawisko Meissnera. Gdy jednak natężenie pola zewnętrznego osiągnie wartość H_{k1}, czyli tzw. pierwszą wartość krytyczną, pole magnetyczne wnika w próbkę w postaci pojedynczych włókien strumienia. Włókien strumienia magnetycznego nie należy mylić z liniami sił pola. Linie sił pola — to tylko formalne krzywe, wzdłuż których ustawiają w polu swój moment magnetyczny małe próbne dipole. Gęstość linii sił pola jest umowna, istotne jest tylko, aby była proporcjonalna do gęstości strumienia. Włókna strumienia magnetycznego są natomiast rzeczywistymi nitkokształtnymi obszarami nadprzewodnika, w których jest skupione pole magnetyczne. Te włókniste obszary przenikają próbkę nadprzewodnika na wskroś. Każde takie włókno zawiera jeden flukson, czyli jeden kwant strumienia magnetycznego. Lokalna gęstość strumienia (czyli lokalna indukcja magne-

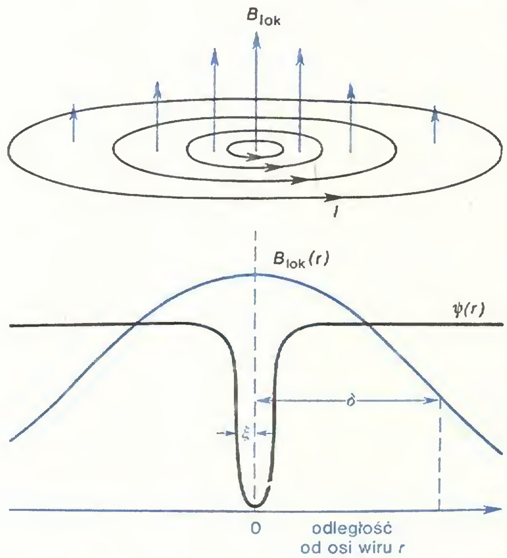
pierwsza wartość krytyczna

włókna strumienia magnetycznego (fluksoidy)

tyczna) jest największa w środku włókna i maleje w miarę wzrostu odległości od środka (rys. 16). Rdzeń włókna ma średnicę rzędu 2ξ. Dookoła osi



Rys. 15. Zależność średniej indukcji magnetycznej i namagnesowania nadprzewodnika II typu od natężenia pola magnetycznego



Rys. 16. Rozkład lokalnej indukcji magnetycznej i parametru uporządkowania dokoła osi wiru prądowego w nadprzewodniku II typu

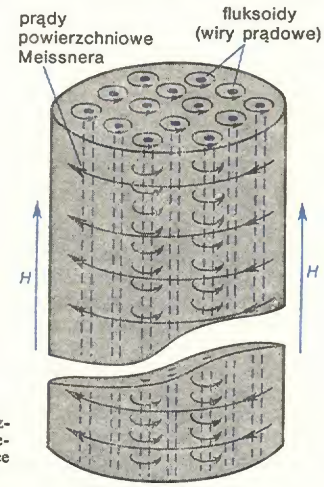
rdzenia cyrkulują prądy nadprzewodnictwa o gęstości malejącej ze wzrostem odległości od rdzenia. Właśnie te prądy wytwarzają pole magnetyczne w rdzeniu włókna. Włókno strumienia magnetycznego, zwane także fluksoidem, jest więc włóknem podobnym do włókna wirowego w cieczy, z tym że tutaj wokół osi włókna krążą elektrony nadprzewodnictwa a nie cząsteczki płynu. Ponieważ wektor lokalnej indukcji B_{lok} we włóknie strumienia magnetycznego ma kierunek zewnętrznego pola magnetycznego H, włókna nadają próbce dodatni moment magnetyczny zmniejszający jej diamagnetyzm.

W miarę wzrostu natężenia zewnętrznego pola magnetycznego coraz więcej fluksoidów przenika próbkę nadprzewodnika i w związku z tym coraz bardziej zmniejsza się diamagnetyczny moment próbki. Podobnie jak w stanie czysto nadprzewodzącym, ten diamagnetyczny moment jest momentem prądów Meissnera, krążących po powierzchni próbki w kierunku odwrotnym do kierunku krążenia wirów prądowych w poszczególnych fluksoidach (rys. 17).

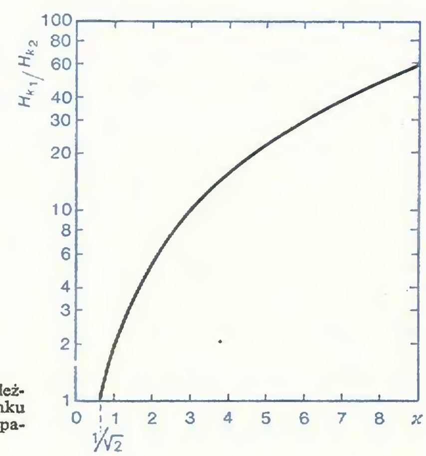
druga wartość krytyczna

Gdy natężenie zewnętrznego pola magnetycznego osiąga drugą krytyczną wartość H_{k_2}, fluksoidy gęsto wypełniają próbkę, rozkład indukcji magnetycznej po przekroju poprzecznym próbki staje się jednorodny i równy $\mu_0 H_{k_2}$ — moment magnetyczny próbki zmniejsza się do zera i próbka przechodzi w stan normalny. Stosunek H_{k_2}/H_{k_1} dla danego nadprzewodnika zależy od wartości parametru $\varkappa$ (rys. 18). Na powierzchni próbki pozostanie jednak jeszcze bardzo cienka nadprzewodząca warstwa o grubości rzędu zasięgu kolelacji ξ. Warstwa ta znika dopiero wtedy, gdy natężenie pola magnetycznego osiągnie trzecią wartość krytyczną $H_{k_3} = 1{,}7 H_{k_2}$. W warstwie powierzchniowej może krążyć słaby prąd nadprzewodnictwa, moment magnetyczny tego prądu jest jednak zbyt mały w porównaniu z poprzednim momentem magnetycznym ekranujących prądów Meissnera, aby go można było zmierzyć tą samą metodą.

trzecia wartość krytyczna



Rys. 17. Prądy powierzchniowe i skwantowane wiry prądowe w próbce nadprzewodnika II typu



Rys. 18. Zależność stosunku H_{k_1}/H_{k_2} od parametru $\varkappa$

nadprzewodnictwo warstwy powierzchniowej

Nadprzewodząca warstwa powierzchniowa ekranuje jednak wnętrze próbki od słabego zmiennego pola elektromagnetycznego nałożonego na pole stałe $H < H_{k_3}$, dlatego można ją wykryć metodą pomiaru indukcji wzajemnej dwu koncentrycznych cewek, wewnątrz których znajduje się badana próbka. Warstwa powierzchniowa zmniejsza także opór elektryczny próbki, jeśli stały prąd pomiarowy jest wystarczająco słaby, aby nie zniszczyć nadprzewodnictwa warstwy. Nadprzewodnictwo warstwy powierzchniowej utrzymuje się do H_{k_3} tylko na tych obszarach powierzchni próbki, które są równoległe do linii sił pola zewnętrznego. W wypadku istnienia składowej pola prostopadłej do powierzchni próbki nadprzewodnictwo powierzchniowe znika, gdy $H_{k_2} < H < H_{k_3}$. Na obszarach powierzchni próbki prostopadłych do H nadprzewodnictwo powierzchniowe powyżej H_{k_2} istnieć nie może. Nie może ono istnieć także wtedy, gdy próbka jest pokryta warstwą nienadprzewodzącego metalu (np. warstwą miedzi).

Stan mieszany

W polu magnetycznym o natężeniu H większym od pierwszej wartości krytycznej a mniejszym od drugiej ($H_{k1} < H < H_{k2}$), próbka nadprzewodnika II typu nie jest ani w stanie czysto nadprzewodzącym, ani w stanie normalnym. Nie jest też w stanie pośrednim, gdyż w objętości próbki nie da się wyodrębnić makroskopowych obszarów poszczególnych faz. Całą objętość próbki zajmuje jedna faza termodynamiczna, która niezbyt trafnie nazywa się fazą mieszaną, mimo iż nie jest ona mieszaniną różnych faz. Fazę tę nazywa się także fazą Szubnikowa. Przemiany fazowe związane z przejściem ze stanu nadprzewodzącego w stan

faza mieszana



Rys. 19. Wykres fazowy nadprzewodnika II typu

mieszany i ze stanu mieszanego w stan normalny przy zmianie natężenia pola zewnętrznego są przemianami II rodzaju. Przypominamy, że przejście nadprzewodnika I typu ze stanu nadprzewodzącego w stan normalny w polu magnetycznym jest przemianą fazową I rodzaju. Wykres fazowy dla nadprzewodnika II typu jest przedstawiony na rys. 19.

uporządkowanie włókien strumienia magnetycznego

W monokrystalicznej, pozbawionej defektów próbce nadprzewodnika II typu faza mieszana charakteryzuje się przestrzennie uporządkowanym rozkładem włókien wirowych strumienia magnetycznego tzw. fluksoidów. W płaszczyźnie poprzecznej do linii sił pola magnetycznego siatka fluksoidów może mieć układ trójkątny lub kwadratowy. Ilustracja 95 z tabl. 25 przedstawia obraz trójkątnej siatki fluksoidów otrzymany pod mikroskopem elektronowym przez Essmanna i Trauble'a po naniesieniu proszku ferromagnetycznego na powierzchnię próbki wykonanej ze stopu ołów-antymon. Ziarenka proszku skupiły się w obszarach najsilniejszego pola, czyli w miejscach, gdzie rdzenie wirów wychodzą na powierzchnię próbki.

Podobnie jak lokalna indukcja magnetyczna, uporządkowany przestrzenny rozkład w stanie mieszanym wykazuje również parametr uporządkowania nadprzewodzącego (np. względna gęstość elektronów nadprzewodnictwa).

Zależność indukcji średniej $B_{śr}$ (tj. uśrednionej po całej objętości próbki) od natężenia zewnętrznego pola magnetycznego H jest przedstawiona na rys. 15. W stanie Meissnera pole nie wnika do próbki i $B_{śr} = 0$. Przy $H = H_{k_1} B_{śr}$ gwałtownie wzrasta, co oznacza, że w próbkę wnika dużo fluksoidów. Dalej $B_{śr}$ wzrasta stopniowo ze wzrostem H, zależność $B_{śr}(H)$ nie jest jednak liniowa. Indukcja $B_{śr}$ jest proporcjonalna do H dopiero w stanie normalnym, tj. powyżej H_{k_2}. Taka postać funkcji $B_{śr}(H)$ dla nadprzewodników II typu przypomina zależność momentu pędu nadpłynnego helu od prędkości kątowej obracającego się naczynia, w którym zawarty jest ten hel. W obu tych wypadkach w określonych warunkach pojawiają się podobne struktury wirowe.

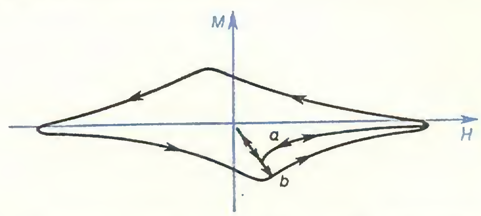
Zjawiska nieodwracalne

Histereza magnetyczna

Proces namagnesowania nadprzewodnika II typu w polu magnetycznym jest procesem odwracalnym tylko w wypadku doskonałych próbek. Krzywa namagnesowania $M(H)$ przebiega wtedy w obu kierunkach jednakowo i nie tworzy pętli histerezy. W rzeczywistości próbki nadprzewodników wykazują pewne odstępstwo od doskonałości i namagnesowanie próbki nie jest jednoznaczną fukcją natężenia pola magnetycznego, gdyż zależy od kierunku zmian pola (rys. 20).

wpływ defektów struktury

Defekty struktury w próbce nadprzewodnika utrudniają ruch fluksoidów — zarówno ich wnikanie w próbkę przy wzroście H, jak i opuszczanie próbki przy zmniejszaniu się H. Defekty mogą w próbce tworzyć lokalne obszary o mniejszej wartości H_{k2}; obszary takie przechodzą w stan normalny przy mniejszych wartościach natężenia pola i stanowią pułapkę (studnię potencjału) dla fluksoidów. Fluksoidy więzną



Rys. 20. Histereza magnetyczna niejednorodnego nadprzewodnika II typu

w tych miejscach, podobnie jak w otworach w próbce, i mogą je opuścić dopiero pod naporem innych fluksoidów (pomiędzy fluksoidami o tym samym kierunku wektora wirowości z małej odległości działają siły wzajemnego odpychania), albo pod wpływem siły oddziaływania prądu elektrycznego na fluksoidy, jeśli przez próbkę taki prąd płynie. Niekiedy miejsca uwięźnięcia opuszcza naraz duża liczba fluksoidów, co przejawia się gwałtowną zmianą strumienia magnetycznego w próbce, czyli tzw. skokiem strumienia.

rozpraszanie energii

Nieodwracalność procesu namagnesowania nadprzewodnika jest przyczyną rozpraszania energii pola magnetycznego. Miarą strat energetycznych podczas jednego cyklu namagnesowania jest wielkość pola powierzchni ograniczonej pętlą histerezy. W okresowo zmieniającym się polu magnetycznym nadprzewodnik II typu w stanie mieszanym rozprasza energię tego pola w postaci ciepła, i to zarówno wtedy, gdy zmienia się zewnętrzne pole magnetyczne, jak i wtedy, gdy zmienia się pole magnetyczne pochodzące od prądu zmiennego płynącego przez nadprzewodnik.

Krytyczna gęstość prądu

Dla zastosowań technicznych jest szczególnie ważne, aby przez przewód z nadprzewodnika można było przepuszczać bez strat prąd elektryczny o możliwie największym natężeniu. W stanie Meissnera prąd płynie tylko po powierzchni nadprzewodnika. Prąd stały o natężeniu I płynący przez drut o promieniu r wytwarza na jego powierzchni pole magnetyczne o natężeniu $H = I/2\pi r$. Jeśli natężenie tego pola osiągnie wartość H_{k1}, drut z nadprzewodnika II typu przejdzie w stan mieszany i prąd rozłoży się po całym przekroju przewodu. Fluksoidy w próbce będą mieć postać zamkniętych współśrodkowych pierścieni kołowych, zgodnie z kształtem linii sił pola magnetycznego wytworzonego przez prąd. Ponieważ energia włókna jest proporcjonalna do jego długości, fluksoidy zmniejszają swą długość kurcząc się w kierunku

kurczenie się i znikanie fluksoidów

osi i znikają w pobliżu osi. Na ruch fluksoidów wpływa także oddziaływanie między fluksoidami i prądem płynącym przez próbkę.

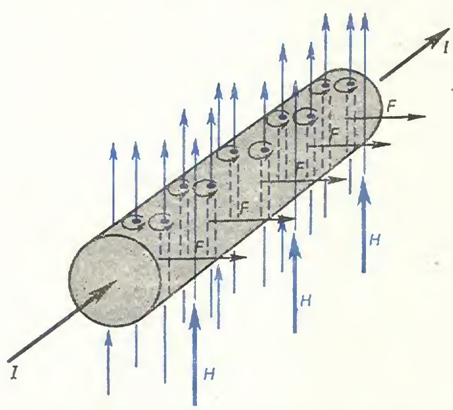
Ruch włókien strumienia magnetycznego i ich znikanie na osi drutu powoduje wydzielanie się ciepła, dlatego z przepływem prądu przez nadprzewodnik w stanie mieszanym wiąże się rozpraszanie energii. Aby uniknąć strat energetycznych, należy w jakiś sposób zatrzymać poruszające się fluksoidy, np. przez wytwarzanie w drucie defektów, na których fluksoidy będą się zaczepiać. W ten sposób można przez drut z nadprzewodnika przepuścić bez strat prąd stały o natężeniu znacznie większym niż $I_{k1} = 2\pi r H_{k1}$. Maksymalny prąd stały, który może przepływać bez strat przez daną próbkę drutu, nazywamy prądem nasycenia albo prądem krytycznym. Ponieważ natężenie tego prądu zależy od przekroju poprzecznego drutu, lepiej posługiwać się pojęciem krytycznej gęstości prądu j_k. Krytyczna gęstość prądu stałego dla nadprzewodnika nie jest stałą materiałową (jak np. H_{k2}), gdyż zależy od defektów w strukturze próbki, czyli od sposobu jej obróbki.

krytyczna gęstość prądu

Płynięcie strumienia

Rozpatrzmy teraz bliżej oddziaływanie wzajemne między fluksoidami i stałym prądem elektrycznym przepływającym przez próbkę. Załóżmy, że drut z twardego nadprzewodnika, przez który płynie prąd stały, znajduje się w zewnętrznym stałym polu magnetycznym skierowanym prostopadle do osi drutu (w takich warunkach pracuje większość uzwojeń w urządzeniach elektrycznych prądu stałego) i że natężenie pola jest wystarczające do wytworzenia w drucie stanu mieszanego.

Prąd elektryczny płynący wzdłuż drutu będzie oddziaływał na fluksoidy siłą skierowaną prostopadle zarówno do kierunku prądu, jak i do osi fluksoidu



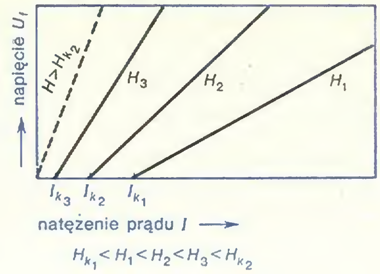
Rys. 21. Szkic ilustrujący ruch fluksoidów w nadprzewodniku z prądem w obecności zewnętrznego pola magnetycznego

(rys. 21). Pochodzenie tej siły, zwanej siłą Lorentza, możemy łatwo zrozumieć, jeśli przypomnimy sobie, że pole magnetyczne oddziałuje na poruszające się ładunki elektryczne (czyli na prąd) siłą prostopadłą do linii sił pola i do kierunku ruchu ładunków. Siła ta jest siłą wzajemnego oddziaływania prądu i pola magnetycznego. Fluksoidy pod wpływem działającej na nie siły Lorentza będą się poruszać w poprzek drutu. Nie jest to jednak ruch przyśpieszony, gdyż towarzyszy mu tarcie, czyli lepkość, pochodząca od prądów wirowych, wzbudzonych w próbce tymże właśnie ruchem strumienia magnetycznego. Równowaga hamującej siły lepkości i siły Lorentza określa prędkość ruchu fluksoidów.

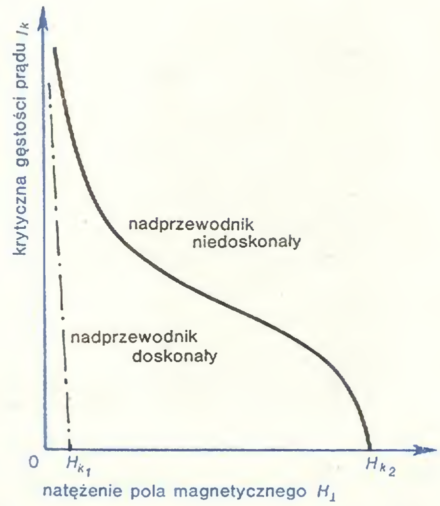
poprzeczny ruch fluksoidów

Ponieważ ruch fluksoidów związany jest z pokonywaniem siły tarcia, czyli z wykonywaniem pracy przez prąd elektryczny, to wzdłuż drogi prądu wystąpi spadek napięcia U_f, co jest równoważne pojawieniu się oporu elektrycznego $R_f = I \cdot U_f$. Ponieważ opór ten jest wywołany ruchem płynięcia fluksoidów, nazywa-

opór płynięcia



Rys. 22. Zależność napięcia indukowanego ruchem fluksoidów od natężenia prądu i zewnętrznego pola magnetycznego



Rys. 23. Zależność krytycznej gęstości prądu w nadprzewodniku II typu od natężenia pola magnetycznego prostopadłego do kierunku prądu

my go oporem płynięcia strumienia magnetycznego, albo krócej — oporem płynięcia.

Rys. 22 przedstawia zależność spadku napięcia U w nadprzewodniku II typu od natężenia przepływającego przezeń prądu. Dopóki gęstość prądu nie osiągnie wartości krytycznej j_k, spadek napięcia jest równy zeru. Wartość j_k dla danej próbki zależy od natężenia zewnętrznego pola magnetycznego H, a także od jego kierunku względem drutu. Zależność $j_k(H)$ dla poprzecznego pola magnetycznego jest najważniejszą charakterystyką twardego nadprzewodnika z punktu widzenia jego przydatności do celów praktycznych. Na rys. 23 przedstawione są dwie takie charakterystyki dla tego samego materiału: jedna dla próbki jednorodnej i odprężonej, druga dla próbki z defektami struktury krystalicznej. Charakterystyki takie podaje się zazwyczaj dla $T = 4{,}2$ K.

charakterystyka nadprzewodnika II typu

Zjawiska nieodwracalne w nadprzewodnikach odgrywają bardzo ważną rolę w zastosowaniach nadprzewodników dla celów praktycznych, szczególnie w elektrotechnice. Niestety, jak dotąd, teoria nadprzewodnictwa nie obejmuje zjawisk nieodwracalnych (np. oddziaływań fluksoidów z defektami struktury krystalicznej); w tym zakresie posługujemy się jedynie uproszczonymi modelowymi przedstawieniami i fenomenologicznym ujęciem zagadnienia.

Zagadnienie wysokotemperaturowego nadprzewodnictwa

Sprawą doniosłej wagi, zarówno ze względów poznawczych, jak i praktycznych, jest rozstrzygnięcie

temperatura krytyczna

zagadnienia, czy nadprzewodnictwo może występować również w temperaturach wysokich. Zagadnienie to można rozpatrywać opierając się na teorii BCS. Wyzyskując zależność (1) otrzymujemy, że $T_k \approx$ $\approx (\theta)e^{-1/g}$. Symbol θ oznacza tutaj tzw. temperaturę Debye'a, czyli temperaturę charakterystyczną dla wzbudzeń fononowych. Temperatura ta zależy od budowy krystalicznej danej substancji; jej sens jest taki, że $k\theta$ jest maksymalną energią fononu w danej sieci krystalicznej. Parametr g w wykładniku potęgi jest iloczynem stałej oddziaływania elektron–fonon–elektron oraz gęstości stanów na poziomie Fermiego w stanie normalnym. T_k zależy więc od dwu parametrów, z których jeden charakteryzuje samą sieć krystaliczną, a drugi — wielkość oddziaływania elektronów z tą siecią. Oddziaływanie to nie może być zbyt silne, gdyż wtedy sama sieć stałaby się niestabilna i musiałaby przejść w inną modyfikację krystalograficzną ze słabszym oddziaływaniem. Parametr g nie może więc przybierać zbyt dużych wartości. Dla większości znanych nadprzewodników $g < 1/2$. Najwyższą wartość g, bo ok. $1/2$, ma ołów, ale jego temperatura Debye'a jest niska ($\theta = 95$ K) i dlatego T_k dla ołowiu jest równa 7,2 K. Czynnik eksponencjalny $e^{-1/g}$ zmniejsza więc wartość θ co najmniej o rząd. Trzeba tu zwrócić uwagę, że dla tak dużych g, jak w ołowiu, wzór (1) nie stosuje się ściśle, gdyż został wyprowadzony przy założeniu, że $g \ll 1$.

Temperatury Debye'a dla większości ciał stałych mieszczą się w przedziale 10^2–10^3 K. Spośród nadprzewodników najwyższą temperaturę Debye'a (równą 990 K) ma $Be_{22}Rn$. Jednak gęstość stanów elektronowych w tym związku jest mała (a więc mały parametr g) i stąd $T_k = 9{,}6$ K.

wodór metaliczny

Istnieją uzasadnione przypuszczenia że zestalony wodór pod wysokim ciśnieniem (rzędu 2,6 Mbar) może przejść w fazę metaliczną, której temperatura Debye'a wynosiłaby ok. $3{,}5 \cdot 10^3$ K. Gdyby ta metaliczna faza wodoru okazała się nadprzewodzącą, to nawet przy słabym oddziaływaniu elektronowo-fononowym temperatura krytyczna nie byłaby niższa od 30 K, a może nawet sięgałaby 200 K w przypadku silnego oddziaływania. Można się spodziewać, że nawet po usunięciu ciśnienia metaliczna faza zestalonego wodoru pozostanie w metastabilnym stanie. Takie metastabilne fazy są znane; np. diament w niskich temperaturach ma wyższą energię swobodną niż grafit.

Przy fononowym mechanizmie pojawienia się nadprzewodnictwa niskie temperatury przejścia w stan nadprzewodzący wynikają z własności sieci krystalicznej, która pośredniczy w oddziaływaniu elektronów z sobą. Dla uzyskania wyższych temperatur T_k potrzebne byłoby oddziaływanie inne, znacznie silniejsze niż elektronowo-fononowe. Przypuszcza się, że nadprzewodnictwo może się pojawić również w wyniku przyciągania się elektronów na skutek wymiany wirtualnych ekscytonów. Ekscytony są elementarnymi wzbudzeniami układu elektronów. Wzbudzenia te przypuszczalnie rozchodzą się ze słabym tłumieniem w ośrodkach niemetalicznych, np. w półprzewodnikach. Zatem, w celu utworzenia par Coopera związanych siłami wymiany ekscytonu trzeba, aby metal znajdował się w ciasnym kontakcie z półprzewodnikiem lub dielektrykiem, co można osiągnąć np. w układach warstwowych. Energia ekscytonów może być znacznie większa niż energia fononów, a odpowiadająca tej energii temperatura charakterystyczna jest równa $\theta_e = \hbar\omega/k \approx 10^4$ K, co — po uwzględnieniu czynnika eksponencjalnego — odpowiadałoby $T_k = 10^3$ K. Ekscytonowe nadprzewodnictwo mogłoby więc istnieć w temperaturze pokojowej, a może nawet i wyższej.

ekscytony

przykład Little'a

Ciekawy przykład mechanizmu prowadzącego do wysokotemperaturowego nadprzewodnictwa w cząsteczkach organicznych podał W. A. Little. Cząsteczka ma kształt przewodzącego łańcucha z bocznymi polaryzującymi się gałęziami. Elektrony przewodnictwa w łańcuchu mogłyby przyciągać się wzajemnie za pośrednictwem polaryzujących się odgałęzień łańcucha. Jest to oddziaływanie kulombowskie, a więc energia tego oddziaływania może być duża. Według ocen Little'a, krytyczna temperatura w tym modelu powinna wynosić 2400 K. Głównym zarzutem wysuniętym przeciw takiemu poglądowi było stwierdzenie, że w jednowymiarowym łańcuchu nadprzewodnictwo nie może istnieć, gdyż fluktuacje gęstości elektronów doprowadziłyby do jego zaniku.

zakres stosowalności teorii BCS

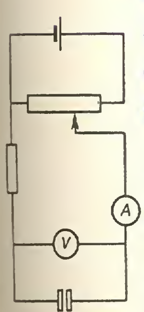
Ponieważ przy ocenie T_k korzystaliśmy z zależności wynikającej z teorii BCS, warto by się jeszcze zastanowić nad zakresem stosowalności tej teorii. Teoria BCS w ogólnym sformułowaniu nie może uwzględniać szczegółów budowy konkretnej substancji, dlatego w poszczególnych wypadkach można otrzymać wyniki odbiegające nieco od przewidywań tej teorii. Wynika to stąd, że samo oddziaływanie elektronowo-fononowe zależy od wielu czynników, między innymi od widma fononowego danej substancji oraz od ekranowania elektronów dodatnimi jonami. Oddziaływanie to nie zawsze jest słabe, jak założono w teorii. Istotne jest jednak, że teoria BCS jest na tyle ogólna, że w jej ramach można dokonywać wielu uzupełnień i poprawek nie naruszających podstaw samej teorii.

Uwzględnienie zależności przyciągającego oddziaływania między elektronami od szczegółów budowy sieci itp. zmienia postać wzoru (1), wiążącego T_k z temperaturą charakterystyczną dla pośredniczących w oddziaływaniu między elektronami wzbudzeń, nie na tyle jednak, aby można było oczekiwać dla T_k wartości znacznie większych, niż wynikające z tego wzoru.

A. C. ROSE-INNES, E. H. RHODERICK *Nadprzewodnictwo*, Warszawa 1973; L. ŚNIADOWER *Półprzewodniki nadprzewodzące*, Post. Fiz. 23, 157 (1972).

Zjawiska tunelowe w nadprzewodnikach

Eugeniusz Trojnar



Rys. 1. Schemat obwodu do wyznaczania charakterystyk tunelowych

Tunelowanie elektronów normalnych

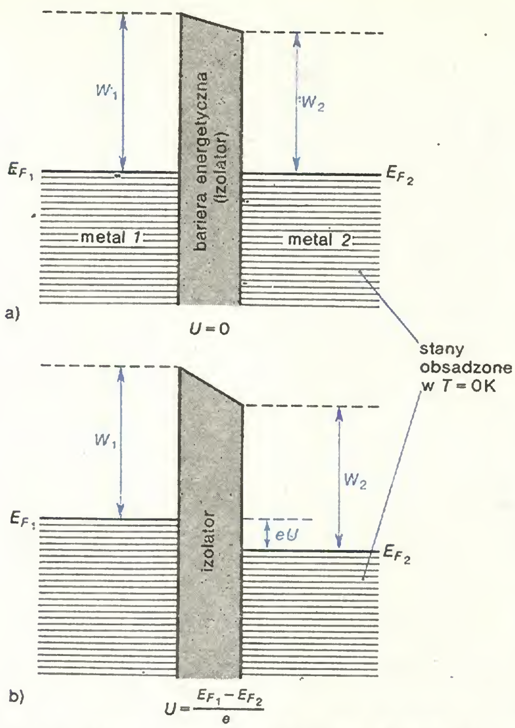
Wyobraźmy sobie obwód elektryczny składający się ze źródła prądu stałego, kondensatora płaskiego, oporników i mierników, jak na rys. 1. W normalnych warunkach trudno oczekiwać, że przez kondensator popłynie prąd stały. Jeśli jednak odległość między okładkami kondensatora zmniejszy się do ok. 5 nm to amperomierz zarejestruje przepływ prądu przez kondensator — mimo istnienia warstwy izolatora między okładkami. Przepływ tego prądu jest spowodowany kwantowym zjawiskiem tunelowania cząstek przez barierę energetyczną. Wytłumaczymy to zjawisko posługując się modelem elektronowych poziomów energetycznych w metalu.

Kawałek metalu można rozpatrywać jako jamę potencjału dla elektronów przewodnictwa. Elektron, aby wyjść z metalu, musiałby pokonać siłę przyciągania go przez sieć jonową, czyli wykonać pracę wyjścia. Według mechaniki klasycznej musiałby mieć na to wystarczającą energię.

Na rys. 2 przedstawiono schematy poziomów energetycznych w dwu zbliżonych do siebie kawałkach metalu (np. w okładzinach kondensatora). Kreski poziome oznaczają poziomy energetyczne zajęte przez elektrony w temperaturze $T = 0$ K. Najwyższy

poziom Fermiego

z tych poziomów nazywa się poziomem Fermiego. Wysokość kreski pionowej ponad poziomem Fermiego przedstawia wartość pracy wyjścia (*W*). Po-



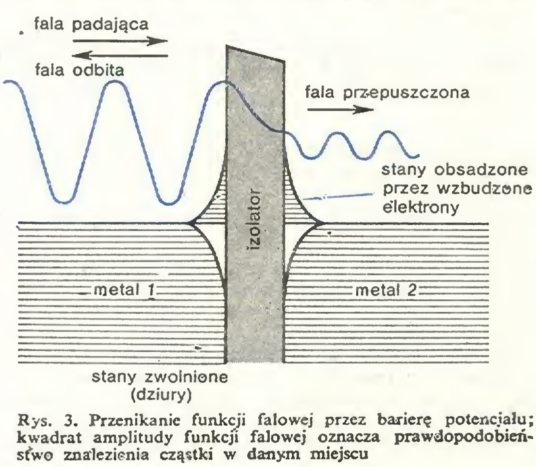
Rys. 2. Elektronowe poziomy energetyczne w układzie złożonym z dwu metali rozdzielonych warstwą izolatora (*W* — praca wyjścia): a) $U = 0$; poziomy Fermiego E_F leżą naprzeciw siebie; b) przyłożone napięcie powoduje przesunięcie poziomów Fermiego względem siebie o eU (e — ładunek elektronu)

bariera potencjału

między obu kawałkami metalu istnieje więc bariera energetyczna nazywana także barierą potencjału. Aby w klasyczny sposób pokonać tę barierę, czyli przejść przez nią, elektron musiałby mieć energię co najmniej równą wysokości tej bariery (czyli równą pracy wyjścia).

W temperaturze $T > 0$ K niektóre elektrony wzbudzają się na wyższe poziomy, leżące ponad poziomem Fermiego, i wtedy prawdopodobieństwo przejścia przez barierę wzrasta. W wysokiej temperaturze może nawet występować emisja elektronów z metalu zwana termoemisją. Wzbudzone elektrony zwalniają stany leżące poniżej poziomu Fermiego (rys. 3).

Fizyka kwantowa dopuszcza możliwość przejścia elektronu przez barierę potencjału także w temperaturze zbliżonej do zera bezwzględnego, chociaż wtedy



Rys. 3. Przenikanie funkcji falowej przez barierę potencjału; kwadrat amplitudy funkcji falowej oznacza prawdopodobieństwo znalezienia cząstki w danym miejscu

żaden z elektronów nie zajmuje wystarczająco wysokiego poziomu energetycznego. W fizyce kwantowej każdej cząstce przyporządkowuje się określoną falę (stan, w jakim znajduje się dana cząstka jest opisany równaniem falowym), przy czym kwadrat amplitudy tej fali w pewnym punkcie przestrzeni oznacza prawdopodobieństwo znalezienia tam cząstki. Fala częściowo odbija się od bariery, a częściowo przenika przez nią, dlatego istnieje różne od zera prawdopodobieństwo znalezienia cząstki również po drugiej stronie bariery (rys. 3). Cząstka przenika przez barierę, jakby przechodziła przez tunel w barierze — stąd nazwa takich zjawisk.

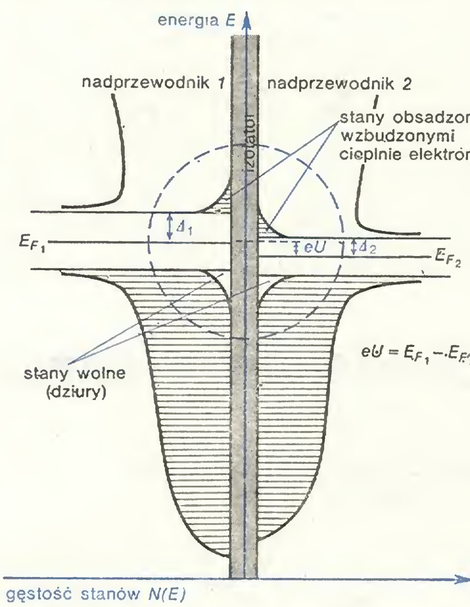
zjawiska tunelowe

W praktyce bardzo trudno jest umieścić dwa kawałki metalu w małej (ok. 5 nm) odległości, i zarazem nie dopuścić do ich zetknięcia, dlatego oba kawałki rozdziela się warstwą izolatora. Układ warstwowy do badania zjawisk tunelowych w nadprzewodnikach można wykonać w następujący sposób: na szklane podłoże naparowuje się w próżni warstwę nadprzewodzącego metalu (np. Pb), następnie warstwę tę poddaje się działaniu powietrza, aby ją utlenić na powierzchni (metal trudno utleniający się można pokryć cienką warstewką jakiegoś innego izolatora), a później naparowuje się na nią ten sam albo inny metal nadprzewodzący (np. Sn).

Zjawiska tunelowe w nadprzewodnikach można podzielić na dwa typy: tunelowanie jednocząstkowych wzbudzeń (elektronów normalnych) i tunelowanie elektronowych par Coopera (→ nadprzewodnictwo). Tunelowanie elektronowych par przewidział teoretycznie w 1962 r. B. D. Josephson i dlatego zjawiska związane z tym tunelowaniem noszą nazwę zjawisk Josephsona. Za prace dotyczące zjawisk tunelowych w nadprzewodnikach (i półprzewodnikach) B. Josephson, I. Giaever i L. Esaki otrzymali w 1973 r. nagrodę Nobla.

wykresy gęstości stanów

Omówimy najpierw tunelowanie elektronów normalnych. Dla wyjaśnienia zjawisk posłużmy się rysunkami podobnymi do rys. 2, z tym że teraz nie będziemy oznaczać wysokości bariery, natomiast obwiedziemy krzywymi zakończenia kresek obrazujących zajęte poziomy energetyczne. Długość tych



Rys. 4. Wykresy gęstości stanów dwu nadprzewodników rozdzielonych warstwą izolatora

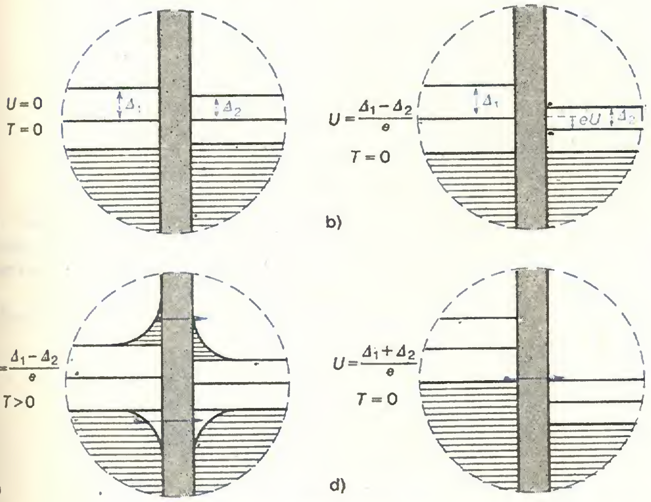
poziomych kresek odpowiada liczbie stanów (poziomów energetycznych) przypadających na jednostkowy przedział energetyczny, w którym leży dana kreska. W ten sposób otrzymamy wykresy gęstości stanów

elektronowych $N(E)$ dla obu nadprzewodników rozdzielonych barierą (rys. 4).

przerwa energetyczna

Po obu stronach poziomu Fermiego w nadprzewodnikach istnieje pasmo wzbronionych wartości energii dla elektronów, czyli tzw. przerwa w widmie wzbudzeń elektronowych. Szerokość połowy tego pasma wzbronionego oznaczamy grecką literą delta (Δ). Gęstość stanów elektronowych w nadprzewodniku wzrasta do nieskończoności w miarę zbliżania się do górnej lub dolnej krawędzi przerwy energetycznej.

Jeśli oba nadprzewodniki rozdzielone warstwą izolatora mają ten sam potencjał elektryczny, to ich



Rys. 5. Środkowa część wykresu gęstości stanów (obwiedziona kółkiem na rys. 4) przy różnych warunkach zewnętrznych. Strzałki niebieskie wskazują kierunek tunelowania elektronów

przesunięcie poziomów Fermiego

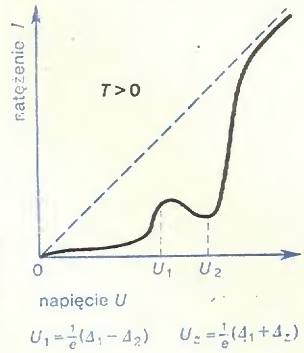
poziomy Fermiego leżą dokładnie naprzeciw siebie (rys. 5a). Przyłożenie napięcia U do bariery, czyli połączenie każdego z nadprzewodników z odpowiednim biegunem źródła napięcia stałego, spowoduje przesunięcie się poziomów Fermiego względem siebie o przedział energetyczny eU (e jest ładunkiem elektronu), przy czym wyższy poziom Fermiego ma metal połączony z biegunem ujemnym (rys. 5b). W temperaturze $T > 0$ (rys. 5c) niektóre stany powyżej przerwy energetycznej są obsadzone przez wzbudzone cieplnie elektrony, natomiast część stanów poniżej przerwy pozostaje wolna.

Prąd przez barierę może płynąć tylko wtedy, gdy naprzeciw obsadzonych stanów po jednej stronie bariery są odpowiednie wolne stany po drugiej stronie. Dopóki przesunięcie poziomów Fermiego jest mniejsze od różnicy połówek przerw energetycznych $\Delta_1 - \Delta_2$ (odpowiada to napięciu na barierze $U > (\Delta_1 - \Delta_2)/e$), prąd płynący przez barierę jest bardzo słaby, gdyż udział w nim biorą tylko nieliczne wzbudzone cieplnie elektrony, a gęstość wolnych stanów leżących naprzeciw stanów obsadzonych jest niewielka. Jednakże gdy napięcie wzrośnie do $U = (\Delta_1 - \Delta_2)/e$, górne (albo dolne — w zależności od polaryzacji złącza) krawędzie przerw energetycznych obu nadprzewodników znajdą się naprzeciw siebie i prąd tunelujących elektronów zwiększy się, ponieważ w pobliżu krawędzi przerwy energetycznej gęstość stanów znacznie wzrasta (rys. 5c). Dalsze zwiększenie napięcia spowoduje przesunięcie się względem siebie górnych (albo dolnych) krawędzi przerw energetycznych i w wyniku tego — zmniejszenie prądu, gdyż znów zmaleje gęstość wolnych stanów znajdujących się naprzeciw stanów zajętych (rys. 4; $U > (\Delta_1 - \Delta_2)/e$).

Gdy napięcie na barierze wzrośnie do wartości $U = (\Delta_1 + \Delta_2)/e$, naprzeciw siebie znajdą się znów krawędzie przerw energetycznych, tym razem górna i dolna, i prąd tunelowy szybko wzrośnie (rys. 5d). Przy dalszym zwiększaniu się napięcia coraz więcej wolnych stanów znajdować się będzie naprzeciw stanów zajętych i prąd będzie wzrastał ze wzrostem napięcia.

ujemny opór elektryczny

Charakterystyka napięciowo-prądowa $I = f(U)$ układu dwu nadprzewodników rozdzielonych cienką warstwą izolatora wykazuje więc obszar ujemnego oporu różniczkowego dI/dU w przedziale napięć od



Rys. 6. Charakterystyka prądowo-napięciowa złącza tunelowego dwu nadprzewodników (zależność prądu tunelowego elektronów normalnych od napięcia na złączu)

$(\Delta_1 - \Delta_2)/e$ do $(\Delta_1 + \Delta_2)/e$ (rys. 6). Podobną charakterystykę mają półprzewodnikowe diody tunelowe, co nasuwa myśl o wyzyskaniu nadprzewodników jako materiału na tunelowe diody nadprzewodnikowe.

Ponieważ prąd przy napięciu $U < (\Delta_1 - \Delta_2)$ jest prądem wzbudzonych cieplnie elektronów, to w temperaturze $T = 0$ K prąd ten nie popłynie. Nie popłynie również gdy napięcie $U < (\Delta_1 + \Delta_2)$, ponieważ w temperaturze $T = 0$ K nie ma wolnych stanów poniżej przerwy energetycznej. Wysokość garbu na charakterystyce $I = f(U)$ zależy zatem od temperatury (wzrasta wraz z nią).

Jeśli jedna z okładek rozpatrywanego tutaj kondensatora jest metalem normalnym (np. nadprzewodnikiem powyżej temperatury krytycznej T_k, tzn. że nie ma przerwy energetycznej czyli $\Delta_2 = 0$), to — jak łatwo można sprawdzić na odpowiednim schemacie poziomów energetycznych — prąd tunelowy popłynie, gdy $U = \Delta_1/e$, a charakterystyka $I = f(U)$ nie będzie mieć garbu. Gładką charakterystykę otrzymuje się również wtedy, gdy metale po obu stronach bariery są tymi samymi nadprzewodnikami (prąd tunelowy popłynie, gdy napięcie osiągnie wartość $U = 2\Delta/e$, ponieważ teraz $\Delta_1 = \Delta_2$).

zgodność z teorią BCS

Dane otrzymane z charakterystyk złącza tunelowego i dotyczące szerokości przerwy energetycznej w nadprzewodnikach oraz jej zależność od temperatury potwierdzają przewidywania teorii Bardeena, Coopera i Schrieffera, dokładniej omówione w haśle nadprzewodnictwo. Ale nie tylko te dane można uzyskać w wyniku analizy pomiarów prądu tunelowego. Pierwsze i drugie pochodne charakterystyk napięciowo-prądowych, czyli dI/dU i d^2I/dU^2, mogą dać informacje o gęstości stanów elektronowych i o rozkładzie widmowym fononów w nadprzewodnikach.

Zjawiska Josephsona — tunelowanie par Coopera

Jeśli warstwa rozdzielająca dwa nadprzewodniki jest wystarczająco cienka, nie grubsza niż odległość, z jakiej oddziałują z sobą elektrony pary Coopera (ok. 1÷2 nm), oprócz tunelowania normalnych elektronów może wystąpić tunelowanie elektronów związanych w pary Coopera. Oczywiście, pary nie będą mogły tunelować, gdy któryś z nadprzewodników układu kanapkowego przejdzie w stan normalny.

Dla zrozumienia zjawisk Josephsona musimy posłużyć się pojęciem funkcji falowej opisującej stan

kwantowy cząstki. Próby wyjaśnienia tych procesów oparte na modelu klasycznym przeważnie zawodzą. Długość fali λ i częstość ν związane są z pędem p i energią ε cząstki zależnościami:

$$\lambda = h/p, \qquad \varepsilon = h\nu, \tag{1}$$

gdzie h — stała Plancka.

funkcja falowa pary Coopera

Stan kwantowy pary Coopera można opisać funkcją falową w postaci:

$$\psi = \psi_0 \sin 2\pi \left(\frac{x}{\lambda} - \nu t\right), \tag{2}$$

albo, po uwzględnieniu zależności (1):

$$\psi = \psi_0 \sin \frac{2\pi}{h}(px - \varepsilon t). \tag{3}$$

Dla uproszczenia posłużyliśmy się tu falą płaską rozchodzącą się w kierunku osi x. Litera t oznacza czas. Argument sinusa: $\varphi = 2\pi(px - \varepsilon t)/h$ nazywa się fazą funkcji falowej.

Sens fizyczny fazy funkcji falowej jest dość trudny do uchwycenia. Znacznie łatwiej trafia do wyobraźni fizyczna interpretacja amplitudy ψ_0 funkcji falowej: jej kwadrat oznacza prawdopodobieństwo znalezienia cząstki w danym miejscu. Jeśli są to cząstki naładowane, jak pary Coopera, to ψ_0^2 jest proporcjonalne do gęstości ładunku.

Funkcje falowe wszystkich par Coopera w jednospójnej bryle nadprzewodnika (tzn. w bryle pozbawionej otworów, w których mógłby zostać uwięziony strumień magnetyczny) mają jednakową długość. Wynika to stąd, że wszystkie pary mają jednakowy pęd. Pędy wszystkich par są równe zeru, gdy przez nadprzewodnik nie płynie prąd, i różne od zera, ale jednakowe, gdy nadprzewodnik jest włączony do obwodu prądu.

Funkcje falowe par elektronów mają także zgodną fazę — są spójne (koherentne), przy czym spójność rozciąga się na cały obszar nadprzewodnika. Jeżeli jest to nawet przewód nadprzewodzący w uzwojeniu elektromagnesu osiągający długość do kilku kilometrów, to funkcje falowe par elektronów są spójne na całej długości tego nadprzewodnika.

Gdy przez nadprzewodnik nie płynie prąd, pęd par Coopera równa się zeru, a długość fali jest nieskończona (pomijamy tu kwestię zaniku funkcji falowej na granicy nadprzewodnika w obszarze o grubości ξ). Oznacza to, że faza funkcji falowej w całym nadprzewodniku jest taka sama. Gdy prąd płynie, pęd par jest różny od zera i długość fali ma wartość skończoną, jednakową dla wszystkich par. W poszczególnych punktach nadprzewodnika faza funkcji falowej ma wtedy różną wartość. Różnica faz w dwu punktach nadprzewodnika oznacza więc, że między tymi punktami przepływa prąd par Coopera. Różnica faz funkcji falowych występuje również po obu stronach bariery w złączu Josephsona. W zjawiskach Josephsona odgrywa rolę właśnie ta różnica faz, a nie bezwzględne wartości samych faz.

różnica faz funkcji falowych

Bezwzględna wartość fazy funkcji falowej jest wielkością bez znaczenia dla zjawisk fizycznych. Wybór tej wielkości jest zresztą zupełnie dowolny, gdyż ani punkt początkowy (zero na osi x), ani chwila, od której rozpoczynamy liczenie czasu, nie są ustalone z góry. Pod tym względem faza funkcji falowej przypomina potencjał pola elektrycznego: zero potencjału jest także umowne. W zjawiskach fizycznych gra rolę różnica potencjałów (napięcie), a nie sam potencjał. Faza funkcji falowej ma jeszcze tę własność, znaną dobrze z ruchu falowego, że może być określona (już po wyborze zera) z dokładnością do 2π. Fazy różniące się między sobą o całkowitą wielokrotność liczby 2π (czyli o $2n\pi$, gdzie $n = 1, 2, 3, ...$) są identyczne. Bezwzględna wartość różnicy faz funkcji falowych może się więc zmieniać od 0 do 2π.

stały prąd par Coopera

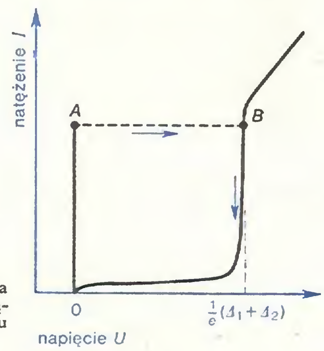
Jeśli na barierze w złączu Josephsona wytworzy się stała w czasie różnica faz funkcji falowych (które nie muszą być jednakowe po obu stronach bariery), to przez barierę płynie stały prąd par Coopera. Obliczenia, których tu przytaczać nie będziemy, wykazują, że gęstość prądu tunelowania par j jest uzależniona od różnicy faz $\varphi_1 - \varphi_2 = \Theta$ w następujący sposób:

$$j = j_k \sin \Theta \tag{4}$$

i może przybierać wartości od zera do j_k. Maksymalna wartość gęstości prądu (j_k) wynosi około 10^{-2} A/cm², a więc jest tego samego rzędu, co gęstość prądu tunelowania elektronów normalnych.

Bariera w złączu Josephsona zachowuje się jak nadprzewodnik, gdyż prąd par Coopera, przepływający przez nią, nie powoduje spadku potencjału elektrycznego. Gdyby jednak (np. na skutek zwiększenia siły elektromotorycznej w obwodzie źródła zasilania) gęstość prądu j na barierze wzrosła powyżej wartości j_k, to na barierze pojawi się napięcie U odpowiadające przepływającemu prądowi i oporowi danej bariery dla prądu tunelowania elektronów normalnych. Prąd stały par Coopera przestanie wtedy płynąć, a przez barierę popłynie prąd tunelowy elektronów normalnych, czyli na charakterystyce $I = f(U)$ nastąpi przeskok od punktu A do punktu B (wzdłuż linii przerywanej na rys. 7). Zmniejszenie siły elektromotorycznej w obwodzie umożliwia przywrócenie stałego prądu Josephsona, przy czym przejście to następuje wzdłuż

prąd tunelowy elektronów normalnych



Rys. 7. Charakterystyka prądowo-napięciowa złącza Josephsona (dla prądu stałego)

dolnej linii ciągłej. j_k jest krytyczną gęstością prądu stałego dla złącza Josephsona, tj. maksymalną gęstością prądu j, jaki może płynąć przez dane złącze nie powodując na nim spadku napięcia. Po pojawieniu się napięcia przez złącze może płynąć prąd stały elektronów normalnych i zmienny prąd Josephsona (szersze omówienie tego zjawiska zawiera rozdział Zjawiska niestacjonarne).

Kwanty strumienia magnetycznego

pary Coopera w nadprzewodniku z otworem

Funkcja falowa par Coopera nie zawsze przedstawia falę płaską poruszającą się w jednym tylko kierunku. Np. w nadprzewodniku niejednospójnym, czyli z otworem, w którym może być uwięziony strumień magnetyczny, fala związana z parą Coopera może otaczać ten otwór. Oznacza to, że dokoła otworu może płynąć prąd par Coopera. Aby w dowolnie obranym punkcie w nadprzewodniku faza funkcji falowej zachowała swą jednoznaczności, wzdłuż krzywej zamkniętej, otaczającej uwięziony strumień, musi się ułożyć całkowita liczba długość fali. Innymi słowy, po jednokrotnym okrążeniu uwięzionego strumienia faza funkcji falowej może się zmienić tylko o całkowitą krotność liczby 2π, czyli o $2\pi n$, gdzie n jest liczbą całkowitą.

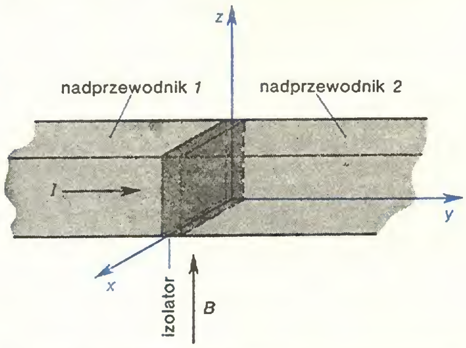
Prąd par Coopera dokoła otworu nie może się więc zmieniać dowolnie, lecz tylko porcjami takimi, aby każdej porcji odpowiadała zmiana fazy funkcji falowej o 2π. Zatem i strumień magnetyczny w otworze musi się zmieniać porcjami, odpowiadającymi zmianom prądu par Coopera. Porcje (kwanty) strumienia

magnetycznego to znane nam już fluksony (→ Nadprzewodnictwo), równe $2{,}07 \cdot 10^{-15}$ Wb. Zmiana strumienia o jeden kwant odpowiada zmianie fazy o 2π. Liczba n jest równa liczbie kwantów strumienia uwięzionego w otworze. Jeżeli nadprzewodniki II rodzaju znajdują się w stanie mieszanym, dotyczy to również strumienia magnetycznego zawartego w pojedynczym fluksoidzie. Doszliśmy więc do wniosku, że kwantowanie strumienia magnetycznego jest również zjawiskiem, w którym, oprócz zjawisk Josephsona, odgrywa rolę zmiana fazy funkcji falowej nośników prądu nadprzewodzącego.

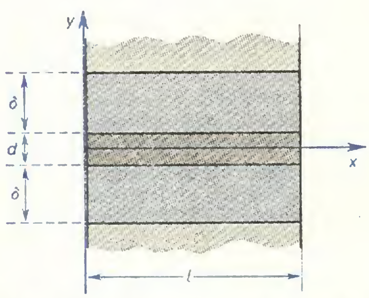
Wpływ pola magnetycznego na prąd Josephsona

Rozpatrzmy przypadek, kiedy złącze Josephsona znajduje się w zewnętrznym polu magnetycznym, którego wektor indukcji $\vec{B}$ jest skierowany wzdłuż płaszczyzny warstwy izolującej. Obierzmy układ współrzędnych jak na rys 8. Wtedy wektor indukcji ma kierunek osi z.

złącze Josephsona



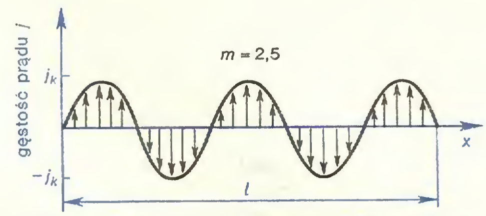
Rys. 8. Złącze Josephsona w zewnętrznym polu magnetycznym (z układem współrzędnych)



Rys. 9. Obszar wnikania pola magnetycznego w złącze Josephsona, gdzie wektor $\vec{B}$ jest prostopadły do płaszczyzny rysunku, δ głębokość wnikania w nadprzewodnik, d grubość izolatora

Rys. 9 przedstawia przekrój złącza w płaszczyźnie $x-y$. Wektor indukcji $\vec{B}$ pola magnetycznego wnika w cały obszar warstwy izolującej i częściowo w nadprzewodnik na głębokość δ. Strumień magnetyczny przenikający złącze wynosi zatem $\Phi = B(2\delta + d)l$, gdzie δ jest głębokością wnikania pola w nadprzewodnik, d — grubością warstwy izolującej, l — długością warstwy w kierunku osi x. Jeśli pole w złączu jest jednorodne, to różnica faz funkcji falowych będzie się równomiernie zmieniać z odległością x oscylując między wartościami 0 i 2π. Liczba oscylacji jest równa liczbie m kwantów strumienia magnetycznego φ_0 mieszczących się w złączu, czyli $m = \Phi/\varphi_0$. Liczba m nie musi tutaj być liczbą całkowitą, ponieważ złącze nie ze wszystkich stron jest otoczone nadprzewodnikiem i strumień może znajdować się także poza złączem. Rys. 10 przedstawia rozkład gęstości prądu Josephsona $j = j_k \sin\Theta$ wzdłuż płaszczyzny złącza dla $m = 2{,}5$. Wartość j_k, podobnie jak wówczas, gdy nie istnieje pole zewnętrzne, jest maksymalną wartością gęstości prądu, który może przepływać przez dane złącze nie powodując pojawienia się na nim napięcia.

rozkład gęstości prądu Josephsona



Rys. 10. Zależność gęstości prądu j od współrzędnej wzdłuż płaszczyzny złącza (oś x jest prostopadła do wektora $\vec{B}$ pola magnetycznego)

Maksymalne natężenie prądu par Coopera I_k, który może płynąć przez złącze w polu o indukcji B, otrzymamy całkując wyrażenie $j = j_k \sin\Theta = f(x)$ po całej czynnej powierzchni złącza. W rozpatrywanym przypadku sprowadza się to do obliczenia pola zawartego między sinusoidą $j = f(x)$ i osią x na odcinku osi od 0 do l (rys. 10) i pomnożeniu wyniku przez długość złącza wzdłuż osi z. Przy sumowaniu pól garbów utworzonych przez sinusoidę $j = f(x)$ należy pamiętać, aby garbom leżącym pod osią x, czyli odpowiadającym ujemnej wartości j, przyporządkować znak minus.

Łatwo sprawdzić na rys. 10, że wynik całkowania (dodawania pól garbów) będzie różny od zera tylko wtedy, gdy na długości l ułoży się niejednakowa liczba garbów nad osią x i pod osią x. Maksymalne natężenie prądu I_k będzie tym większe, im większe pole między sinusoidą i osią x pozostanie nieskompensowane, tzn. pozbawione swego partnera z przeciwnym znakiem. Łatwo też zauważyć, że im większa liczba fluksonów znajdzie się w obrębie złącza, tym mniejsze pole na wykresie $j = f(x)$ może pozostać nieskompensowane i tym słabszy prąd Josephsona może płynąć przez złącze. Oczywiście, prąd może płynąć tylko wtedy, gdy złącze zostanie włączone w obwód źródła prądu.

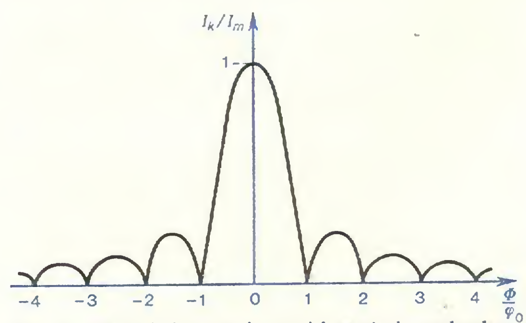
Obliczenia wykazują, że maksymalne natężenie prądu Josephsona płynącego przez złącze umieszczone w polu magnetycznym można przedstawić następującą zależnością:

maksymalne natężenie prądu Josephsona

$$I_k = I_m \frac{\sin\pi(\Phi/\varphi_0)}{(\Phi/\varphi_0)}, \qquad (5)$$

gdzie Φ jest strumieniem zawartym w złączu, φ_0 — kwantem strumienia. Wykres tej zależności przedstawia rys. 11. Przekroczenie natężenia I_k spowoduje pojawienie się napięcia na złączu i zanik stałego prądu Josephsona. I_m jest maksymalnym natężeniem prądu Josephsona, który może płynąć przez złącze w nieobecności pola magnetycznego.

Na wielkość strumienia magnetycznego w złączu możemy wpływać zmieniając natężenie zewnętrznego



Rys. 11. Zależność krytycznej wartości natężenia prądu złącza Josephsona od wielkości strumienia magnetycznego w złączu

pola magnetycznego. Z powyżej przytoczonych rozważań i z wzoru (5) wynika, że za każdym razem, gdy strumień magnetyczny Φ w złączu osiąga wartość równą całkowitej wielokrotności kwantu strumienia φ_0, przez złącze nie może płynąć stały prąd par Coopera (może natomiast płynąć normalny prąd tunelowy). Kwant strumienia jest więc okresem oscylacji maksymalnego natężenia I_k stałego prądu tunelowego par Coopera przez złącze. Dla typowego złącza ($l \approx 2 \cdot 10^{-2}$ cm, $\delta \approx 5 \cdot 10^{-5}$ cm, $d \ll \delta$) okres ten w jednostkach indukcji magnetycznej jest rzędu 10 μT (0,1 Gs).

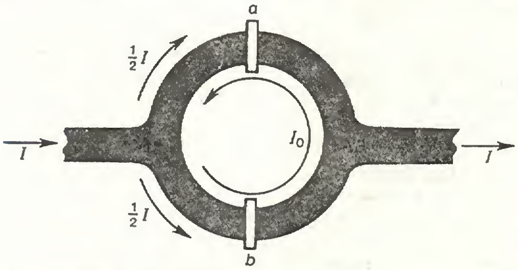
dyfrakcja na złączu Josephsona

Krzywa $I_k/I_m = f(\Phi/\varphi_0)$ na rys. 11 przypomina krzywą rozkładu natężenia światła w obrazie dyfrakcyjnym otrzymanym od jednej szczeliny. Na złączu Josephsona występuje więc dyfrakcja funkcji falowych par Coopera. Możemy również obserwować zjawiska interferencyjne w dwu połączonych złączach Josephsona.

Interferencja kwantowa

nadprzewodnikowy interferometr kwantowy

Zjawisko interferencji fal par elektronowych wystąpi, jeśli w obwód prądu włączymy dwa złącza Josephsona połączone z sobą równolegle nadprzewodzącymi rozgałęzieniami. Schemat takiego urządzenia, zwanego nadprzewodnikowym interferometrem kwantowym, przedstawia rys. 12. Przyjmijmy dla uproszczenia, że



Rys. 12. Pierścień nadprzewodzący z dwoma złączami Josephsona (nadprzewodnikowy interferometr kwantowy)

oba złącza są identyczne, a pętla utworzona z doprowadzeń nadprzewodzących ma oś symetrii. Pomińmy na razie modulację faz funkcji falowych w samych złączach i rozpatrzmy, jak strumień magnetyczny zawarty w pętli wpływa na maksymalną wartość natężenia prądu Josephsona płynącego przez złącza. Oznaczmy ten strumień przez Φ_T; nie musi on być całkowitą wielokrotnością fluksonu φ_0, ponieważ nie jest to strumień uwięziony w nadprzewodniku (pętla ma dwa złącza Josephsona, czyli dwie szczeliny w nadprzewodzącym materiale, przez które może przenikać strumień magnetyczny). Ze strumieniem magnetycznym Φ_T związany jest prąd elektryczny o natężeniu $I_0 = \Phi_T/L$, krążący w pętli (L — współczynnik samoindukcji pętli). Prąd I_0 będzie prądem nadprzewodzącym, dopóki na barierach nie pojawi się napięcie. Zmiana fazy funkcji falowej w nadprzewodzącej części pętli, spowodowana przepływem tego prądu, po jednokrotnym okrążeniu wynosi $2\pi m$, gdzie m jest liczbą fluksonów w strumieniu magnetycznym Φ_T, czyli $m = \Phi_T/\varphi_0$. Jak już wspominaliśmy, dzięki istnieniu szczelin (złącz) w pętli, liczba m nie musi być liczbą całkowitą.

Oprócz cyrkulującego prądu I_0 może jeszcze przez pętlę płynąć prąd I ze źródła zasilania obwodu. Ze względu na symetrię pętli, prąd ten rozdziela się jednakowo na obie jej części (na rys. 12 górna i dolna) tak, że na każdą część pętli przypada $\frac{1}{2}I$. Prąd I w pętli jest prądem nadprzewodzącym, jeśli jego natężenie nie przekroczy pewnej wartości dopuszczalnej. Przez pętlę popłyną więc następujące prądy: prąd o natężeniu $\frac{1}{2}I - I_0$ przez górną część pętli i prąd o natężeniu $\frac{1}{2}I + I_0$ przez dolną część. Zmiany faz na barierach a i b, odpowiadające tym prądom, wynoszą Θ_a i Θ_b. Wzory wiążące zmiany faz z natężeniami prądów Josephsona mają postać zgodną z zależnością (4):

$$\begin{aligned} \tfrac{1}{2}I - I_0 &= I_a \sin\Theta_a, \\ \tfrac{1}{2}I + I_0 &= I_b \sin\Theta_b, \end{aligned} \tag{6}$$

gdzie I_a i I_b są prądami krytycznymi dla barier a i b. Ponieważ z założenia bariery te są jednakowe, to $I_a = I_b = I_k$.

Na całkowitą zmianę fazy funkcji falowej w górnej części pętli, na drodze między punktami A i B, składa się zmiana fazy Θ_a na barierze a oraz zmiana fazy w samym nadprzewodniku $(\Delta\varphi)_a$. Analogicznie w dolnej części faza zmieni się o $\Theta_b + (\Delta\varphi)_b$. Ze względu na konieczność zachowania jednoznaczności faz w punktach A i B, zmiany w obu częściach pętli mogą się różnić tylko o $2\pi n$, gdzie n jest liczbą całkowitą lub zerem. Mamy więc:

$$\Theta_a + (\Delta\varphi)_a = \Theta_b + (\Delta\varphi)_b + 2\pi n. \tag{7}$$

Mimo symetrii geometrycznej względem prostej AB, zarówno zmiany faz na barierach, jak i zmiany faz w nadprzewodzących rozwidleniach pętli mogą być różne po obu stronach prostej AB, gdyż w obecności strumienia Φ_T w obu częściach pętli płyną prądy o różnych natężeniach; w jednej części pętli prąd I_0 płynie zgodnie z prądem $\frac{1}{2}I$, a w drugiej części — przeciwko prądowi $\frac{1}{2}I$. Na różne zmiany faz w obu częściach symetrycznej pętli wpływa więc wartość strumienia Φ_T wywołującego przepływ prądu I_0. Wiemy już, że w nadprzewodzących częściach pętli przepływ tego prądu spowoduje zmianę fazy o $2\pi m$, czyli $(\Delta\varphi)_a - (\Delta\varphi)_b = 2\pi m = 2\pi(\Phi_T/\varphi_0)$.
Równanie (7) możemy więc zapisać:

$$\Theta_a - \Theta_b = 2\pi\left(\frac{\Phi_T}{\varphi_0} - n\right), \tag{8}$$

$$\text{albo:} \quad \Theta_a - \pi\left(\frac{\Phi_T}{\varphi_0} - n\right) = \Theta_b + \pi\left(\frac{\Phi_T}{\varphi_0} - n\right). \tag{9}$$

Oznaczając literą ϑ wartość wyrażenia stojącego po jednej (którejkolwiek) stronie powyższego równania otrzymamy:

$$\Theta_a = \vartheta + \pi\left(\frac{\Phi_T}{\varphi_0} - n\right) \quad \text{oraz} \quad \Theta_b = \vartheta - \pi\left(\frac{\Phi_T}{\varphi_0} - n\right).$$

Po podstawieniu tych zależności do równania (6) dodajemy je stronami i po odpowiednich przekształceniach równanie ma postać:

$$I = 2I_k \sin\vartheta \cdot \cos\left[\pi\left(n - \frac{\Phi_T}{\varphi_0}\right)\right].$$

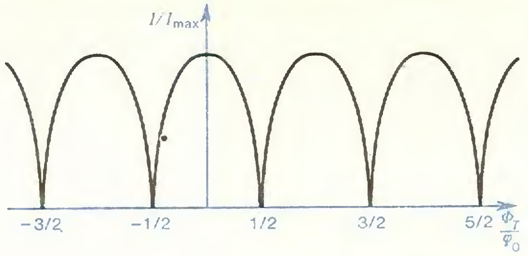
Ponieważ zarówno I jak i I_k muszą mieć ten sam znak, to występujące tu funkcje trygonometryczne muszą być albo obie dodatnie, albo obie ujemne. Można więc napisać: $I = 2I_k|\sin\vartheta| \cdot |\cos[\pi(n - \Phi_T/\varphi_0)]|$, albo, po wprowadzeniu oznaczenia $2I_k|\sin\vartheta| = I_{\max}$:

$$\frac{I}{I_{\max}} = \left|\cos\frac{\pi\Phi_T}{\varphi_0}\right|. \tag{10}$$

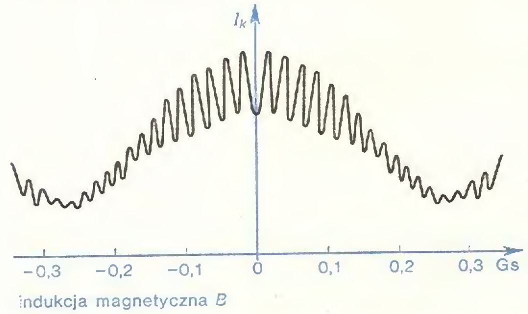
Wykresem tej funkcji jest krzywa przedstawiona na rys. 13. Widzimy, że prąd o maksymalnej wartości natężenia, nie powodujący jeszcze spadku potencjału (napięcia) na złączach, można przepuścić przez pętlę wtedy, gdy strumień magnetyczny w pętli jest całkowitą wielokrotnością fluksonu. Gdy zaś $\Phi_T = (n + \frac{1}{2})\varphi_0$, nawet bardzo słaby prąd spowoduje pojawienie się napięcia.

Okresem zmian krytycznego natężenia prądu Josephsona dla pętli, podobnie jak dla pojedynczego złącza, jest flukson. Jednak z powodu dużej powierzchni pętli w porównaniu z powierzchnią przekroju poprzecznego złącza, dla pętli okres ten w jednostkach

indukcji jest znacznie mniejszy. Jeśli np. pętla ma powierzchnię czynną równą 1 cm^2, to zmianie strumienia o 1 flukson odpowiada zmiana indukcji mag-



Rys. 13. Zależność maksymalnej wartości natężenia prądu pierścienia z dwoma złączami Josephsona od wielkości strumienia magnetycznego zawartego w pierścieniu



Rys. 14. Zależność prądu krytycznego nadprzewodzącego interferometru kwantowego od zewnętrznego pola magnetycznego

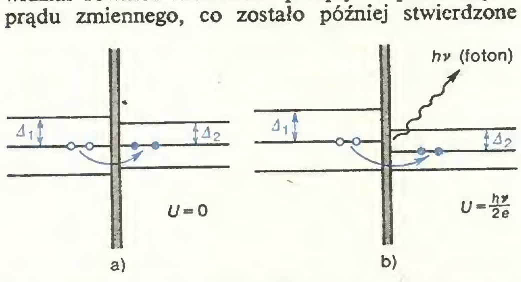
netycznej o $2 \cdot 10^{-11}$ T. Rys. 14 przedstawia wynik nałożenia się oscylacji krytycznego prądu Josephsona, spowodowanych zmianami strumienia magnetycznego w pętli, na oscylacje wywołane zmianami strumienia w samych złączach.

zastosowanie interferometru kwantowego

Na prąd Josephsona w pętli wpływają zauważalnie zmiany indukcji magnetycznej nawet sto razy mniejsze od okresu tych zmian. Za pomocą interferometru kwantowego można więc mierzyć zmiany indukcji wynoszące 10^{-13} T. Takiej czułości nie osiąga żaden konwencjonalny miernik. Dzięki swej wysokiej czułości na zmiany pola magnetycznego, nadprzewodzący interferometr kwantowy służy również do pomiarów takich wielkości jak napięcie lub natężenie prądu elektrycznego, przy czym jego czułość jest o kilka rzędów lepsza niż przyrządów klasycznych.

Zjawiska niestacjonarne

Przepływ stałego prądu Josephsona przez złącze w obecności pola magnetycznego i związane z tym zjawiska interferencyjne są zjawiskami ustalonymi w czasie (stacjonarnymi). Prócz tych zjawisk Josephson przewidział również możliwość przepływu przez złącze prądu zmiennego, co zostało później stwierdzone



Rys. 15. Schemat przejść tunelowych Josephsona: a) bez napięcia na złączu (zjawisko stacjonarne), b) z emisją fotonu pod napięciem na złączu (zjawisko niestacjonarne). Pochłonięcie fotonu spowoduje przejście w odwrotnym kierunku (wbrew polarności źródła napięcia)

doświadczalnie i nazwane zmiennoprądowym albo niestacjonarnym zjawiskiem Josephsona. W zjawiskach stacjonarnych przepływ stałego prądu nadprzewodzącego przez złącze nie powoduje na nim spadku potencjału elektrycznego (napięcia). Przepływ zmiennego prądu Josephsona związany jest z pojawieniem się napięcia na złączu (rys. 15).

zmiennoprądowe zjawisko Josephsona

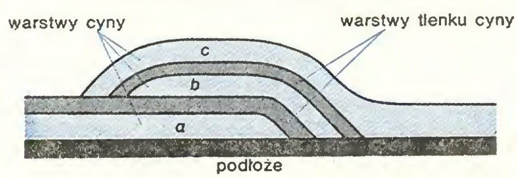
Jeśli na złączu ustali się stałe w czasie napięcie U_0, to energia par Coopera przechodzących przez złącze zmieni się o wartość $\varepsilon_1 - \varepsilon_2 = qU_0$, gdzie q jest ładunkiem pary. Zmiana energii powoduje zmianę fazy $\varphi_1 - \varphi_2 = 2\pi/h(\varepsilon_1 - \varepsilon_2)t = qU_0 t/\hbar$ (wynika to z zależności (3)). Różnica faz po obu stronach złącza zmienia się więc w czasie, a ponieważ prąd Josephsona zależy od różnicy faz, będzie to prąd zmienny. Wyrażenie dla gęstości prądu Josephsona otrzymujemy podstawiając w równaniu (4) różnicę faz $\Theta = qU_0 t/\hbar$:

$$j = j_k \sin\left(\frac{qU_0}{\hbar} \cdot t\right). \tag{11}$$

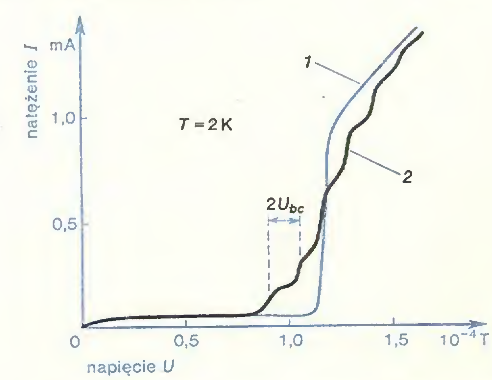
złącze Josephsona jako generator

Prąd Josephsona płynący przez łącze pod stałym napięciem U_0 zmienia się więc okresowo z częstością kołową $\omega = qU_0/\hbar$. Stałe napięcie na złączu Josephsona może zatem generować prąd zmienny. Moc takiego generatora jest jednak bardzo mała, gdyż najczęściej wynosi około 10^{-11} W.

Między napięciem na złączu a częstością oscylacji prądu $\nu = \omega/2\pi$ istnieje następująca zależność liczbowa: $\frac{\nu}{U_0} = \frac{q}{h} = 483{,}6$ MHz/μV ($q = 2e$ oraz h są stałymi uniwersalnymi, a stosunek $h/q = \varphi_0$ jest równy kwantowi strumienia magnetycznego). Napięcie U_0 powinno być mniejsze od $U = (\Delta_1 + \Delta_2)/e$, gdyż przy wyższych napięciach prąd Josephsona szybko zanika, wzrasta natomiast prąd tunelowania elektronów normalnych. Stosowana na ogół wartość napięcia U_0



Rys. 16. Złącze tunelowe Giaevera



Rys. 17. Charakterystyka prądowo-napięciowa złącza przedstawionego na rys. 16. Krzywa 1 — charakterystyka $I = f(U)$ złącza $a-b$ przy napięciu na złączu $b-c$ równym zeru ($U_{bc} = 0$); krzywa 2 — charakterystyka $I = f(U)$ złącza $a-b$ przy napięciu $U_{bc} \neq 0$

leży w przedziale od kilku μV do kilku mV, a odpowiadająca tym napięciom częstość może dochodzić nawet do tysiąca GHz. Wykrycie prądów tak wielkiej częstości jest możliwe dzięki istnieniu pola elektromagnetycznego związanego z tym prądem. Doświadczenia wykazujące bezpośrednio istnienie takiego pola wykonał J. Giaever w 1964 r., używając do tego celu dwu złącz połączonych z sobą wspólną okładziną (rys. 16). Złącza wykonano w sposób następujący.

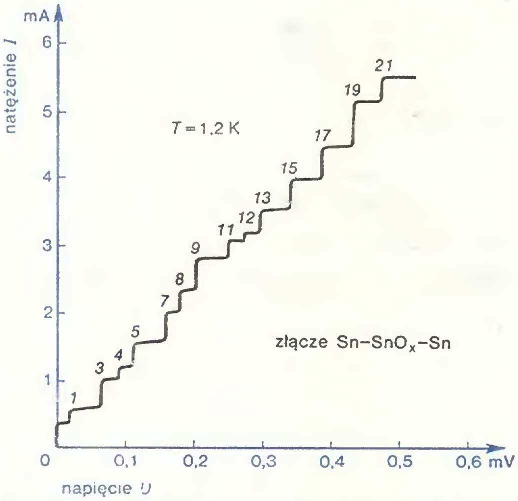
doświadczenie Giaevera

Na podłoże naniesiono warstwę cyny (a), poddano ją długotrwałemu utlenianiu na powierzchni, aby wytworzyć grubą warstwę tlenku, na nią naparowano następną warstwę cyny (b) z cienką warstwą tlenku na powierzchni i w końcu na wierzch naniesiono trzecią warstwę cyny (c). Warstwa izolująca między a i b była wystarczająco gruba, aby tłumić prąd Josephsona, mógł jednak przez nią przepływać prąd tunelowy elektronów normalnych wtedy, gdy do a i b przyłożono napięcie U_{ab} (rys. 17, krzywa *1*). Charakterystyka $I = f(U)$ dla tego złącza zmieniła się po przepuszczeniu pod napięciem U_{bc} zmiennego prądu Josephsona przez warstwę izolującą między b i c (krzywa *2*). Na krzywej pojawiły się schodki odpowiadające wartościom napięcia $U_{ab} = (2\Delta \pm nqU_{bc})/e$, gdzie 2Δ jest szerokością przerwy energetycznej dla cyny. Szerokość poszczególnych schodków $2U_{bc}$ odpowiada energii $2eU_{bc} = \hbar\omega$ fotonów emitowanych przez złącze Josephsona (bc) i pochłanianych w złączu normalnym (ab).

Skok normalnego prądu tunelowego na wyższy schodek następuje za każdym razem, gdy energia fotonów wzrasta o nową porcję, równą qU_{bc}.

złącze Josephsona jako odbiornik

Złącze Josephsona może być również odbiornikiem energii promieniowania elektromagnetycznego, wytworzonego przez inne źródło, np. generator mikrofalowy. Mikrofalowe promieniowanie doprowadza się do złącza falowodem. Okazuje się, że wtedy oprócz zmiennego prądu Josephsona przez złącze może płynąć także stały prąd Coopera, a charakterystyka napięciowo-prądowa napromieniowanego złącza jest schodkowa, jak na rys. 18.



Rys. 18. Schodkowa charakterystyka $I = f(U)$ złącza Josephsona w zmiennym polu elektromagnetycznym

Doprowadzenie do złącza promieniowania mikrofalowego o częstości $f = \Omega/2\pi$ można potraktować jako przyłożenie dodatkowego napięcia przemiennego $u = u_m \cos \Omega t$ (oprócz napięcia stałego U_0). Ogólne napięcie wyniesie więc $U(t) = U_0 + u_m \cos \Omega t$, a ponieważ zmiana w czasie różnicy faz Θ związana jest z napięciem $U(t)$ zależnością $d\Theta/dt = qU(t)/\hbar$, to różnica faz będzie równa:

$$\Theta = \frac{q}{\hbar} \int U(t)\,dt = \frac{q}{\hbar} U_0 t + \frac{qu_m}{\hbar\Omega} \sin \Omega t + \beta, \qquad (12)$$

gdzie β jest stałą całkowania (różnica faz dla $t = 0$).

gęstość prądu Josephsona w polu zmiennym

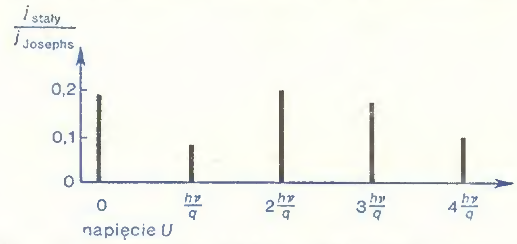
Po podstawieniu wyrażenia (12) do wzoru (4) otrzymamy równanie dla gęstości prądu Josephsona w zmiennym polu elektromagnetycznym:

$$j = j_k \sin\left(\frac{qU_0}{\hbar} t + \frac{qu_m}{\hbar\Omega} \sin \Omega t + \beta\right). \qquad (13)$$

Zmienny prąd Josephsona, o częstości własnej $\omega = qU_0/\hbar$, jest więc modulowany częstością Ω obcego pola elektromagnetycznego, czyli sama częstość Josephsona ω zmienia się w czasie. Dzięki tej modulacji prąd Josephsona zawiera wiele składowych z różnymi częstościami. Wśród tych składowych mogą się znaleźć również składowe z częstością równą zeru, co odpowiada stałej w czasie składowej prądu.

stałe składowe prądu Josephsona

Stałe składowe prądu Josephsona można obliczyć po rozwinięciu równania (13) w szereg Fouriera–Bessela. Okazuje się wtedy, że prąd stały płynie każdorazowo,



Rys. 19. Stałoprądowe składowe charakterystyki $I = f(U)$ w niestacjonarnym zjawisku Josephsona, obliczone z równania (13)

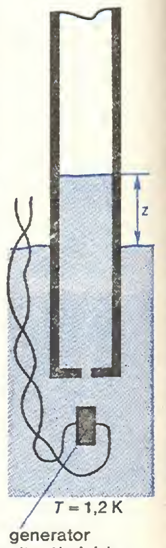
gdy częstość josephsonowska $\omega = qU_0/\hbar$ zrówna się z całkowitą krotnością częstości Ω pola elektromagnetycznego, czyli $qU_0/\hbar = n\Omega$. Charakterystyka napięciowo-prądowa powinna się więc składać z pionowych kresek, występujących w punktach odpowiadającym napięciom na złączu, równym 0, $\hbar\Omega/q$, $2(\hbar\Omega/q)$, $3(\hbar\Omega/q)$ itd. (rys. 19). Różnica między charakterystyką $I = f(U)$ obliczoną z równania (13) i wyznaczoną doświadczalnie (rys. 18) jest spowodowana tym, że w doświadczeniu obserwujemy stały prąd Josephsona na tle prądu tunelowego elektronów normalnych, które także wzbudzają się pod wpływem promieniowania elektromagnetycznego.

Dzięki opisanym powyżej własnościom złącze Josephsona może służyć zarówno jako generator jak i detektor promieniowania elektromagnetycznego w zakresie mikrofalowym i w dalekiej podczerwieni. Innym sposobem wyzyskania złącza Josephsona jest określenie absolutnej jednostki napięcia opartej na równaniu $U = \hbar\Omega/q$ na podstawie badania charakterystyki prądu Josephsona; dokładność jej określenia zależy od dokładności, z jaką znamy stałe fizyczne e i h, oraz od dokładności, z jaką możemy mierzyć częstość $\Omega/2\pi$.

słabe nadprzewodnictwo

Powyżej przedstawiono tylko niektóre podstawowe zjawiska związane z tunelowaniem par Coopera. Z braku miejsca pominięto np. rezonansowe wzbudzenie stałej składowej nadprzewodzącego prądu własną falą elektromagnetyczną, powstałą w złączu w wyniku oddziaływania słabego zewnętrznego pola magnetycznego (ok. 10^{-4} T) na zmienny prąd Josephsona, lub zjawiska zachodzące w pierścieniu nadprzewodzącym z jednym złączem Josephsona. Zwłaszcza te ostatnie zjawiska mają ważne znaczenie ze względu na zastosowanie praktyczne.

Prąd Josephsona może płynąć nie tylko przez złącze z barierą w postaci cienkiej warstwy izolatora. Barierę może stanowić również warstwa półprzewodnika albo normalnego metalu (np. miedzi); może być nią również przewężenie w samej warstwie nadprzewodzącej (mikromostek Dayema) albo kontakt punktowy nadprzewodników. Takie miejsca w nadprzewodniku, w których występują opisane zjawiska, nazywają się słabymi złączami, a zespół tych zjawisk nosi nazwę słabego nadprzewodnictwa.

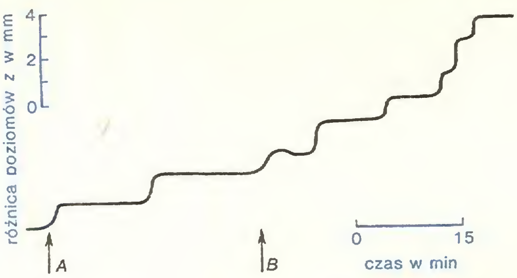


Rys. 20. Schemat aparatury do badania zjawiska Josephsona w ciekłym helu

Zjawisko Josephsona występuje także w nadpłynnym helu. Słabym złączem łączącym dwa naczynia z ciekłym helem II jest otwór o średnicy ok. 10 μm w cienkim (0,1 mm grubości) metalowym dnie rury wstawionej do kriostatu z helem II (rys. 20). W normalnych warunkach poziomy cieczy w obu naczyniach (tj. w rurze i w kriostacie) wyrównują się. Jeśli jednak pod otworem umieścimy kwarcowy generator

zjawisko Josephsona w nadpłynnym helu

ultradźwięków, to poziom helu II w rurze podniesie się na wysokość z spełniającą zależność: $mgz = nh\nu$, gdzie m — masa atomu helu, g — przyśpieszenie ziemskie, ν — częstość drgań kwarcu, n — liczba całkowita lub stosunek dwu niewielkich liczb całkowitych.



Rys. 21. Zjawisko Josephsona w nadpłynnym helu. Zależność różnicy poziomów od czasu trwania doświadczenia. Strzałka A wskazuje moment włączenia oscylatora, strzałka B — chwilę, w której zwiększono moc drgań

Rys. 21 przedstawia zależność różnicy poziomów z od czasu trwania eksperymentu. Różnica poziomów utrzymuje się bez zmian dopóki nie wzrośnie moc oscylatora kwarcowego; wtedy różnica poziomów zwiększa się. Wykres przypomina „schodkową" charakterystykę $I = f(U)$ dla złącza Josephsona w nadprzewodnikach pod napięciem i w zewnętrznym polu elektromagnetycznym. W helu II rolę napięcia odgrywa różnica potencjałów grawitacyjnych gz, rolę ładunku — masa atomu helu, rolę częstości pola elektromagnetycznego — częstość ultradźwięku. Można oczekiwać również innych analogii między zjawiskami Josephsona w nadprzewodnikach i w nadpłynnym helu, dotychczas jednak, ze względu na trudności eksperymentalne, nie były one obserwowane.

R. P. Feynman i in. *Feynmana wykłady z fizyki*, t. 3, Warszawa 1974; J. Rauluszkiewicz *Zjawiska tunelowe w nadprzewodnikach*, Post. Fiz. 23, 181 (1972); A. C. Rose-Innes, E. H. Rhoderick *Nadprzewodnictwo*, Warszawa 1973.

Zastosowanie nadprzewodnictwa

Eugeniusz Trojnar

Mimo iż nadprzewodnictwo zostało odkryte w 1911 r., na szerszą skalę zaczęto je stosować dopiero od 1961 r., tj. od czasu, kiedy nauczono się wytwarzać stopy i związki nadprzewodzące o wysokich parametrach krytycznych. Wyniki badań nad zjawiskami Josephsona także rozszerzyły zakres zastosowań nadprzewodnictwa, głównie w elektronice i technice pomiarowej.

Wszystkie podstawowe własności nadprzewodników, a więc brak oporu elektrycznego, doskonały diamagnetyzm oraz zjawiska tunelowe i kwantowanie strumienia magnetycznego mogą mieć zastosowanie praktyczne.

Najbardziej oczywistą korzyścią ze stosowania nadprzewodników jest możliwość uniknięcia lub zmniejszenia strat energetycznych na ciepło Joule'a w urządzeniach elektrycznych. Oczywiście, w ogólnym rozrachunku ekonomicznym należy uwzględnić koszty utrzymania tych urządzeń w odpowiednio niskiej temperaturze.

Niekiedy o celowości zastosowania nadprzewodników w urządzeniach energetycznych decyduje nie tylko zysk ekonomiczny. Ważnym czynnikiem może być również możliwość znacznego zmniejszenia rozmiarów tych urządzeń przy zachowaniu tej samej mocy nominalnej. Ma to szczególne znaczenie tam, gdzie rozmiary i ciężar urządzeń odgrywają decydującą rolę.

W technice pomiarowej dzięki stosowaniu urządzeń nadprzewodnikowych udało się polepszyć czułość przyrządów pomiarowych o kilka rzędów, a więc osiągnąć wyniki, które innymi sposobami były nieosiągalne.

Omówimy w skrócie najważniejsze urządzenia, w których stosuje się lub można zastosować nadprzewodniki. Zacznijmy od elektromagnesów nadprzewodnikowych, gdyż w tej dziedzinie nadprzewodniki znajdują obecnie najszersze zastosowanie.

Elektromagnesy nadprzewodnikowe

elektromagnesy zwykłe

W zwykłych elektromagnesach z uzwojeniem miedzianym dla wzmocnienia pola stosuje się zazwyczaj rdzenie żelazne. Ten sposób jest skuteczny tylko wtedy, gdy pole wytworzone przez uzwojenie jest słabsze od pola nasycenia rdzenia. W silniejszych polach wkład od rdzenia jest stosunkowo niewielki, dlatego stosowanie rdzeni staje się nieopłacalne. Najlepsze elektromagnesy rdzeniowe mogą wytwarzać pola o indukcji do 6 T w wąskiej szczelinie między nadbiegunnikami (tj. w przestrzeni o objętości rzędu kilku cm³), a ciężar tych elektromagnesów sięga kilku ton.

Pola silniejsze niż 6 T wytwarzają bezrdzeniowe solenoidy z uzwojeniem miedzianym, chłodzone wodą. Pobierają one moc rzędu kilku MW, przy czym ta moc wydziela się w uzwojeniu i musi być odprowadzona na zewnątrz za pomocą odpowiedniego systemu chłodzenia. Ciężar miedzianych uzwojeń w takich elektromagnesach wynosi kilkaset kilogramów.

solenoidy nadprzewodnikowe

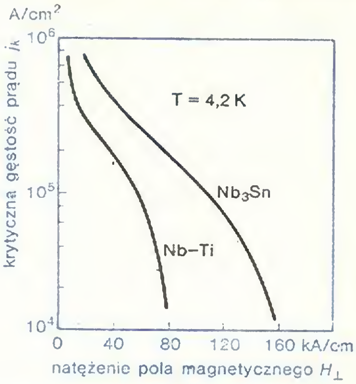
Stosując nadprzewodnikowe uzwojenia w solenoidach (il. 58, tabl. 16) można uzyskać pola o indukcji do 17 T w objętości roboczej rzędu kilkudziesięciu lub nawet kilkuset cm³, lub pola 4 T w objętości rzędu kilku m³. Elektromagnes z uzwojeniem nadprzewodzącym nie rozprasza mocy, koszty jego eksploatacji są więc znacznie niższe niż elektromagnesu z uzwojeniem normalnym, zwłaszcza przy długich okresach pracy. Po uruchomieniu elektromagnesu nadprzewodnikowego można końce jego uzwojeń zewrzeć krótkim przewodem nadprzewodzącym i trwały prąd w uzwojeniach może krążyć już po odłączeniu źródła zasilania. Ciężar elektromagnesu nadprzewodnikowego zależy od wielkości przestrzeni, w której wytworzone jest pole. Jeśli ta objętość wynosi ok. 200 cm³, to na uzwojenia wystarczy kilka kilogramów materiału nadprzewodzącego. Duże elektromagnesy nadprzewodnikowe do celów specjalnych mogą jednak ważyć kilkaset lub nawet kilka tysięcy kilogramów.

Materiały na uzwojenia elektromagnesów nadprzewodnikowych

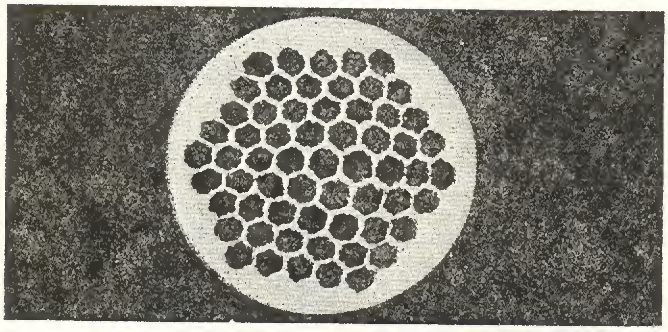
Od nadprzewodników, z których wykonuje się uzwojenia elektromagnesów, wymaga się spełnienia następujących warunków: 1) muszą one mieć odpowiednio wysokie wartości drugiego krytycznego natężenia pola magnetycznego H_{k_2} w temperaturze pracy, tj. przeważnie w temperaturze wrzenia helu pod normalnym ciśnieniem (4,2 K) lub w nieco wyższej temperaturze w wypadku chłodzenia parami helu. 2) muszą przenosić duże gęstości prądu, rzędu kilkuset A/mm², w polu o natężeniu zbliżonym do H_{k_2}.

Do produkcji drutów nadprzewodzących używa się

stop Nb-Ti najczęściej stopu Nb-Ti, dla którego H_{k_2} (4,2 K) = = 90 kA/cm, T_k = 10 K i gęstość prądu nasycenia j_k (4,2 K) = 0,6·10⁵ A/cm² w polu H = 64 kA/cm (rys. 1). Przewody na uzwojenia elektromagnesów przeważnie wykonane są z wiązki wielu cienkich drutów Nb-Ti, od kilku do kilkudziesięciu µm średnicy, skręconych dookoła osi wiązki (skok skrętu ok. 5 mm)



Rys. 1. Zależność krytycznej gęstości prądów j_k od natężenia zewnętrznego pola magnetycznego $H_\perp$ dla Nb-Ti i Nb_3Sn w T = 4,2 K



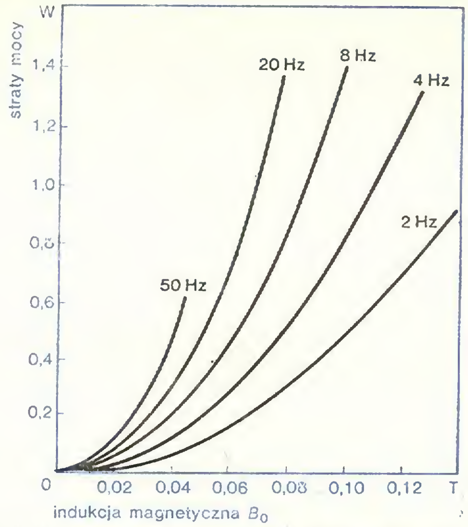
Rys. 2. Zdjęcie szlifu poprzecznego przewodu nadprzewodnikowego

i zaprasowanych w otoczkę (matrycę) z normalnego metalu, najczęściej miedzianą (rys. 2). Mała grubość i skręcenie drutów nadprzewodzących zapobiegają rozprzestrzeniającym się lawinowo w drucie skokom strumienia magnetycznego. Takie nagłe zmiany strumienia powodują wydzielanie się ciepła, mogłyby więc doprowadzić do lokalnego przegrzania przewodu i do utraty jego nadprzewodnictwa. Normalny metal matrycy odgrywa rolę termicznego i elektrycznego stabilizatora chroniącego przewód przed skutkami lokalnego przegrzania. Materiał matrycy ma znacznie lepsze przewodnictwo cieplne i elektryczne niż stop nadprzewodnikowy w stanie normalnym, lepiej odprowadza ciepło do kąpieli helowej, a także hamuje gwałtowne przeskoki strumienia magnetycznego.

Ze stopu Nb-Ti wykonuje się uzwojenia dla elektromagnesów wytwarzających pola o indukcji do 8 T. Do wytwarzania silniejszych pól (do 17 T) sporządza się elektromagnesy nawinięte z taśmy lub z drutu, w skład których wchodzi związek międzymetaliczny Nb_3Sn wykazujący własności nadprzewodzące. Ma on następujące parametry krytyczne: H_{k_2} (4,2 K) = = 160 kA/cm, T_k = 18 K, j_k (4,2 K) = 10⁵ A/cm² w polu ok. 110 kA/cm. Związek międzymetaliczny Nb_3Sn jest kruchy i nie daje się walcować ani wyciągać w druty. Przewody z Nb_3Sn wytwarza się metodą dyfuzji termicznej cyny w odpowiednio przygotowane druty lub taśmy niobowe.

związek Nb_3Sn

Podobne własności nadprzewodzące i mechaniczne jak Nb_3Sn wykazują związki V_3Ga i V_3Si. Rozwój technologii wytwarzania stopów nadprzewodzących może przynieść nowe materiały na uzwojenia: szczególnie obiecującym materiałem wydaje się być stop potrójny $Nb_{0,79}(Al_{0,73}Ge_{0,27})_{0,21}$ o parametrach krytycznych H_{k_2} (4,2 K) = 220 kA/cm, T_k = 23 K.



Rys. 3. Zależność strat energetycznych w nadprzewodzącej cewce z Nb-Ti od amplitudy B_0 przemiennego pola magnetycznego o różnej częstości. Objętość uzwojeń 85 cm³

W projektach zastosowań twardych nadprzewodników w urządzeniach pracujących w zmiennych polach magnetycznych należy liczyć się ze stratami energetycznymi w tych nadprzewodnikach. Straty te są wynikiem ruchu strumienia magnetycznego i nieodwracalności procesu namagnesowania (histereza magnetyczna) nadprzewodnika (rys. 3).

Laboratoryjne i przemysłowe zastosowanie elektromagnesów nadprzewodnikowych

aparatura do rezonansu jądrowego

Elektromagnesy nadprzewodnikowe stanowią obecnie powszechne wyposażenie laboratoriów naukowych, w których bada się własności substancji w silnych polach magnetycznych. Elektromagnesy o dużej jednorodności pola stanowią istotną część aparatury do rezonansu jądrowego zastępując stare, ciężkie elektromagnesy rdzeniowe, wymagające w dodatku bardzo stabilnych, a więc drogich źródeł zasilania (stabilność pola w elektromagnesach nadprzewodnikowych uzyskuje się w krótkozwartym reżimie pracy, a wysoką jednorodność pola — przez specjalne nawinięcie).

mikroskop elektronowy

W mikroskopach elektronowych można stosować soczewki magnetyczne z uzwojeniem nadprzewodzącym, co pozwala uzyskać większą zdolność rozdzielczą niż przy użyciu zwykłych soczewek magnetycznych. Elektromagnesy nadprzewodnikowe wielkich rozmiarów są wykorzystane w fizyce jądrowej do komór pęcherzykowych. Np. magnes wodorowej komory pęcherzykowej w Argonne (USA) ma 4,8 m średnicy wewnętrznej, wytwarza pole o indukcji 1,8 T i zawiera 45 t uzwojeń nadprzewodzących. W laboratoriach fizyki wysokich energii buduje się także akceleratory cząstek (np. synchrotrony) z dużymi elektromagnesami nadprzewodnikowymi.

badanie plazmy

Elektromagnesy nadprzewodnikowe używane są również do badań plazmy i, być może, znajdą szersze zastosowanie w generatorach magnetohydrodynamicznych, jeśli rozpowszechnią się one w energetyce. W energetyce elektromagnesy nadprzewodnikowe mogą także służyć do akumulacji energii (w postaci energii pola magnetycznego), która byłaby wyzyskana w godzinach szczytowego zapotrzebowania. Solenoid nadprzewodzący o współczynniku samoindukcji L może zmagazynować w swej objętości energię magnetyczną równą $\frac{1}{2}LI^2$, gdzie I jest natężeniem prądu

akumulacja energii

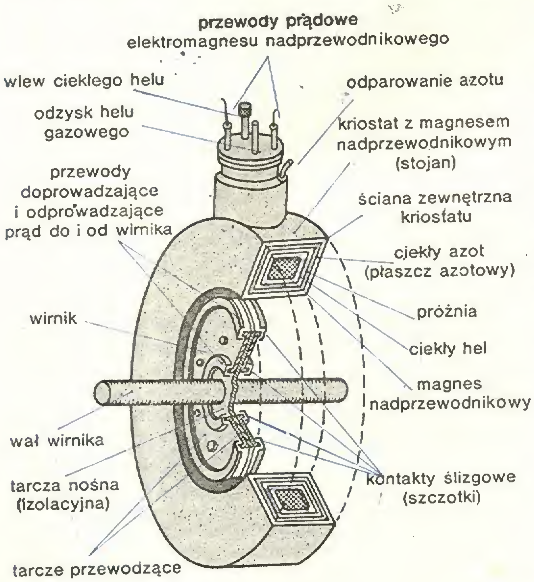
krążącego bez strat w uzwojeniach krótkozwartego solenoidu. Zmagazynowaną energię pola magnetycznego można by w godzinach szczytu przetworzyć na energię prądu elektrycznego.

Obecnie jednak powszechną uwagę skupiają dwie inne możliwości szerokiego, przemysłowego zastosowania elektromagnesów nadprzewodnikowych: w maszynach elektrycznych (silnikach i prądnicach) i w lewitujących (unoszących się) pociągach.

Maszyny elektryczne

maszyny na prąd stały

Pierwszymi maszynami elektrycznymi, w których zastosowano uzwojenia nadprzewodnikowe, były silniki jednobiegunowe, lub jednakobiegunowe, zwane też unipolarnymi lub homopolarnymi (rys. 4). Ru-



Rys. 4. Silnik unipolarny

chomą częścią takiego silnika jest okrągła tarcza metalowa (albo bęben) wirująca w stałym polu magnetycznym wytwarzanym przez elektromagnes nadprzewodnikowy. Siła napędzająca wirnik jest siłą oddziaływania pola magnetycznego na stały prąd elektryczny płynący przez tarczę wirnika w kierunku radialnym (albo w bębnie po tworzącej). Prąd do wirnika doprowadzają kontakty ślizgowe.

Silniki jednobiegunowe z uzwojeniem nadprzewodnikowym odznaczają się prostotą konstrukcji i małymi rozmiarami przy stosunkowo dużej mocy. Straty energetyczne są mniejsze niż w silnikach konwencjonalnych. Trudnym zagadnieniem technicznym jest tu jednak problem kontaktów ślizgowych, doprowadzających prąd o dużym natężeniu, gdyż są to silniki niskonapięciowe. Dla zwiększenia napięcia (i zmniejszenia natężenia prądu) stosuje się układ wielu tarcz połączonych w szereg. Każdy silnik może być także prądnicą, jeśli z kontaktów będziemy pobierać prąd w zamian za energię mechaniczną poruszającą wirnik.

Nadprzewodnikowe maszyny jednobiegunowe znajdują zastosowanie w przemyśle chemicznym do elektrolizy, w układach napędzających pompy wodne, walcarki lub młyny w przemyśle hutniczym, górniczym, papierniczym; dzięki dobremu stosunkowi mocy do ciężaru mogą także służyć do napędu statków i okrętów.

maszyny na prąd przemienny

Nadprzewodzące uzwojenia można także stosować w magneśnicach maszyn prądu przemiennego. Wirnik synchronicznej prądnicy z uzwojeniem nadprzewodzącym, zasilany prądem stałym, wytwarza stałe pole magnetyczne, wirujące wewnątrz nieruchomego twornika (stojana) i wzbudza w jego uzwojeniach prąd przemienny. Przy tej samej prędkości wirowania moc prądnicy jest proporcjonalna do indukcji magnetycznej w szczelinie między wirnikiem i stojanem. W konwencjonalnych prądnicach indukcja sięga 1 T; dalsze jej zwiększanie jest utrudnione z powodu nasycenia żelaza. Nadprzewodzące uzwojenia wirnika umożliwiają zwiększenie indukcji do 4 T bez zastosowania żelaznego rdzenia, a więc przy zmniejszonych rozmiarach i masie wirnika. Uzwojenia stojana takiej prądnicy wykonuje się jednak z normalnego metalu, tj. z miedzi, gdyż zastosowanie nadprzewodzących uzwojeń pracujących w silnym zmiennym polu magnetycznym byłoby związane ze znacznymi stratami energetycznymi w tych uzwojeniach, porównywalnymi ze stratami w uzwojeniach miedzianych.

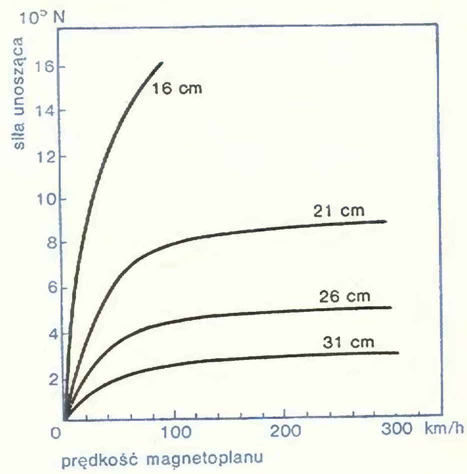
Unoszone magnetycznie pociągi (magnetoplany)

Intensywne prace badawcze dotyczące pociągów unoszonych magnetycznie prowadzą kraje, w których sprawa szybkiego transportu osób jest palącym zagadnieniem (Japonia, USA, RFN). Prędkość, jaką osiągają najszybsze zwyczajne pociągi, nie przekracza na ogół 250 km/h.

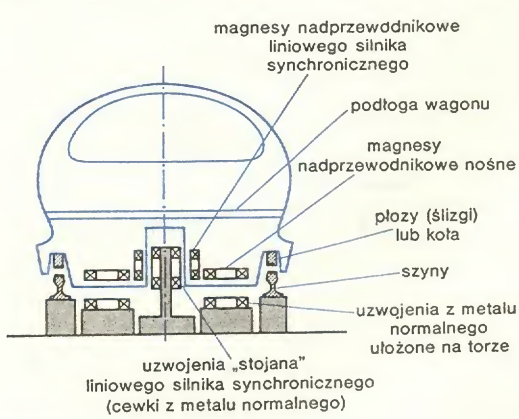
Czynnikami ograniczającymi szybkość pociągu są wibracje i słaba przyczepność kół do szyn. Tor, po którym jedzie szybki pociąg, musi spełniać bardzo wysokie wymagania dotyczące jakości. Pociąg unoszący się nad torem pozwala ominąć te trudności. Obecnie przeprowadzane są próby z doświadczalnymi pojazdami, unoszonymi magnetycznie.

magnetoplan

Przy małych prędkościach pociąg jedzie po torze, ale już przy prędkości 90 km/h unosi się nad torem



Rys. 5. Zależność siły unoszenia od prędkości magnetoplanu



Rys. 6. Przekrój poprzeczny magnetoplanu

na wysokości około 30 cm (rys. 5). W podwoziu pociągu umieszczone są elektromagnesy nadprzewodnikowe, których pole magnetyczne, poruszając się wraz z pociągiem, wzbudza prądy wirowe w metalowych płytach (lub cewkach) ułożonych na torze. Pole magnetyczne tych prądów jest lustrzanym odbiciem pola elektromagnesów. Siła odpychania między tymi polami jest siłą unoszącą pojazd (rys. 6).

silnik liniowy synchroniczny

Siłę napędową, poruszającą pociąg do przodu, dostarcza liniowy synchroniczny silnik elektryczny. Silnik ten jest właściwie zwykłym silnikiem synchronicznym ze stojanem rozwiniętym w linię. Rolę jego wirnika spełniają dodatkowe boczne elektromagnesy nadprzewodnikowe, umocowane sztywno w pojeździe. Stojan silnika to uzwojenia miedziane lub aluminiowe, ułożone wzdłuż toru pojazdu. Prąd elektryczny zasilający silnik jest włączany synchronicznie z ruchem pojazdu w obwód coraz to nowych segmentów stojana znajdujących się nieco przed pojazdem i wyłączanych przy mijaniu danego segmentu przez pojazd. Biegnące pole magnetyczne wytwarzane przez prąd w tych segmentach pociąga za sobą elektromagnesy boczne pojazdu, czyli wirnik silnika. W ten sposób pojazd jest prowadzony przez biegnącą falę magnetyczną. Elektromagnesy silnika, oprócz funkcji napędowej, pełnią jeszcze funkcję stabilizatorów, tj. utrzymują pojazd nad środkiem toru.

Magneśnicę silnika i elektromagnesy nośne pojazdu zasila prądnica umieszczona w pojeździe i napędzana np. turbiną spalinową. Doświadczalny pojazd osiąga prędkość 500 km/h, zabiera 100 pasażerów z bagażem (ok. 10 t) i waży 35 t z pełnym wyposażeniem i zapasem paliwa na dwie godziny jazdy. Przy maksymalnej prędkości pojazd pobiera moc rzędu kilku MW; jest ona w głównej mierze zużywana na pokonanie oporów aerodynamicznych (il. 62, tabl. 16).

Nadprzewodnikowe linie przesyłowe (kable)

Kable nadprzewodnikowe mogą zastąpić napowietrzne linie przesyłowe. Przesyłanie wielkich mocy normalnymi liniami napowietrznymi wymaga stosowania wysokich napięć, lecz zwiększanie napięcia wiąże się ze wzrostem strat energetycznych na upływ i ulot.

Kable nadprzewodnikowe stwarzają możliwość przysyłania bez strat prądu stałego lub przemiennego o dużym natężeniu. Należy tylko liczyć się z wydatkowaniem energii na utrzymywanie kabla w niskiej temperaturze. Straty na odbieranie doprowadzonego z zewnątrz ciepła mogą wynosić ok. 10 W na 1 km długości kabla.

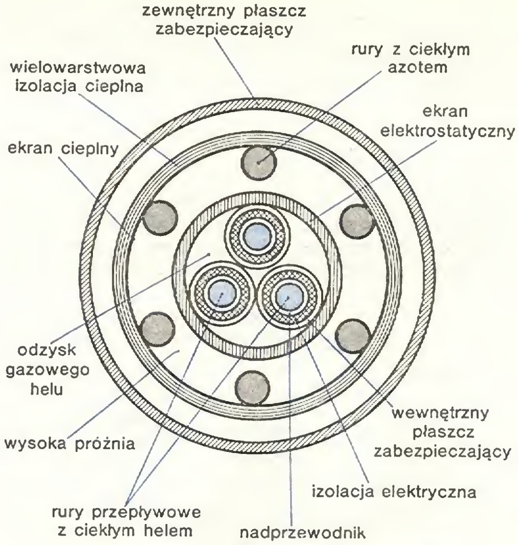
kable prądu stałego

Kable prądu stałego można wykonać z twardego nadprzewodnika II typu. Ze względu na koszt chłodzenia kabla, jego zastosowanie może być opłacalne dopiero przy przesyłaniu bardzo dużych mocy (ponad 100 MW). Przy odpowiednim doborze nadprzewodnika, obciążenie kabla prądem stałym mogłoby wynosić ok. 10^5 A/cm^2.

Nadprzewodnikowy kabel prądu stałego włączony w istniejący system energetyczny prądu przemiennego wymagałby zastosowania przetworników na obu swych końcach.

kable prądu przemiennego

Do przesyłania bez strat prądu przemiennego można użyć kabla z nadprzewodnika I typu lub doskonałego nadprzewodnika II typu pod warunkiem, że amplituda natężenia pola wytworzonego przez prąd w kablu nie przekroczy wartości H_{k_1} (rys. 7). Ponieważ kabel musi pracować w temperaturze nie niższej niż 4,2 K (dalsze obniżanie temperatury wiąże się z dużymi kłopotami technicznymi), mogą być użyte tylko dwa metale: ołów i niob (stopy na ogół mają małe wartości H_{k_1} i trudno z nich wykonać doskonały nadprzewodnik). Prąd przemienny w kablu z nadprzewodnika I typu popłynie tylko po jego powierzchni, dlatego w takich kablach najkorzystniejsze byłoby zastosowanie cienkiej warstwy nadprzewodnika naniesionej na nienadprzewodzące podłoże.



Rys. 7. Przekrój poprzeczny kabla nadprzewodnikowego

Ostatnio jednak pojawiają się projekty zastosowania w kablach na trójfazowy prąd przemienny nadprzewodników twardych, np. Nb-Ti lub Nb_3Sn, gdyż kable z nich wykonane bardziej nadają się do przesyłania dużych mocy, a straty nie powinny przekraczać 0,5% (kable pracują w niezbyt silnym polu własnym). Można by więc osiągnąć sprawność równą 99,5%, czyli znacznie większą niż w normalnych liniach przesyłowych (95%).

chłodzenie kabli

Chłodzenie kabli nadprzewodnikowych o dużej długości jest jednak trudnym problemem technicznym, dlatego wprowadzenie ich do systemów energetycznych nie będzie chyba sprawą najbliższej przyszłości. Obecnie stosuje się je raczej w zamkniętych układach energetycznych, np. na statkach.

Zastosowanie nadprzewodników w technice wielkich częstości

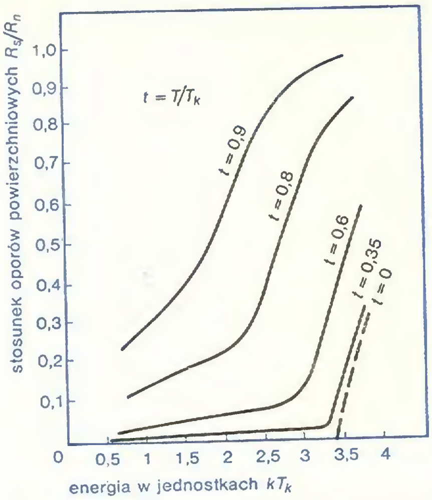
Opór powierzchniowy nadprzewodników

W stanie Meissnera (tzn. w stanie czysto nadprzewodzącym) nadprzewodnik w szybkozmiennym polu elektromagnetycznym wykazuje różny od zera opór powierzchniowy, gdyż w $T > 0$ oprócz par Coopera istnieją w nadprzewodniku elektrony normalne, które rozpraszają energię, podobnie jak w metalach normalnych. Jednak w nadprzewodnikach głębokość wnikania δ pola elektromagnetycznego daleko od T_k jest dużo mniejsza niż grubość warstwy naskórkowości δ_{skin} w normalnym metalu przy wysokich częstościach, poza tym udział elektronów normalnych szybko maleje ze spadkiem temperatury poniżej T_k; oba te czynniki powodują, że opór powierzchniowy nadprzewodników jest znacznie mniejszy niż opór powierzchniowy normalnych metali.

Przy bardzo wielkich częstościach, gdy energia fotonów pola elektromagnetycznego przekracza szerokość przerwy energetycznej ($h\nu \geqslant 2\Delta$), opór powierzchniowy zwiększa się w wyniku pochłaniania energii pola na rozrywanie par Coopera i wzbudzenie elektronów ponad przerwę energetyczną. Rys. 8 przedstawia zależność oporu powierzchniowego nad-

zależność oporu od częstości i temperatury

przewodnika od częstości pola elektromagnetycznego ν i od zredukowanej temperatury $t = T/T_k$. Ponieważ w temperaturze $T = 0$ szerokość przerwy $2\Delta\ (0) = 3{,}5kT_k$, opór powierzchniowy nadprze-



Rys. 8. Opór powierzchniowy nadprzewodnika I typu (Al) w zmiennym polu elektromagnetycznym; R_s oznacza opór powierzchniowy w stanie nadprzewodzącym, R_n — w stanie normalnym

wodnika z $T_k \approx 10$ K będzie różny od zera dopiero przy $\nu \geqslant 10^{12}$ Hz.

Nadprzewodniki II typu w stanie mieszanym albo nadprzewodniki niedoskonałe, w które wnika pole magnetyczne począwszy od niewielkich wartości natężeń, wykazują w zmiennym polu straty na histerezę i na prądy wirowe. Aby uniknąć tych strat, w technice wysokich częstości stosuje się tylko czyste nadprzewodniki o możliwie wysokiej wartości temperatury krytycznej T_k. Najczęściej są to: niob $T_k = 9$ K i ołów $T_k = 7{,}2$ K.

Wnęki rezonansowe wysokiej dobroci

Dobrocią wnęki rezonansowej nazywamy stosunek zmagazynowanej w niej energii pola elektromagnetycznego w ciągu jednego cyklu do energii rozproszonej na ścianach wnęki. Wnęka miedziana przy częstościach rzędu 1 GHz może w normalnej temperaturze mieć dobroć $Q = 5 \cdot 10^4$. Oziębienie wnęki do temperatury ciekłego helu, chociaż zmniejsza opór właściwy czystej miedzi kilkaset razy, jej dobroć zwiększy tylko kilkakrotnie z powodu anomalnego zjawiska naskórkowości. Zjawisko to występuje wtedy, gdy średnia długość l drogi swobodnej elektronu, od której zależy przewodnictwo metalu, staje się większa od grubości warstwy naskórkowości δ_{skin} (jest to warstwa, w którą wnika zmienne pole elektromagnetyczne). Wtedy elektrony poruszające się w głąb metalu nie mogą być przez pole przyśpieszone na całej swej drodze między kolejnymi zderzeniami i dla przewodnictwa są właściwie stracone. Anomalne zjawisko naskórkowości ogranicza więc zwiększanie się przewodnictwa powierzchniowego metali przy oziębianiu.

anomalne zjawisko naskórkowości

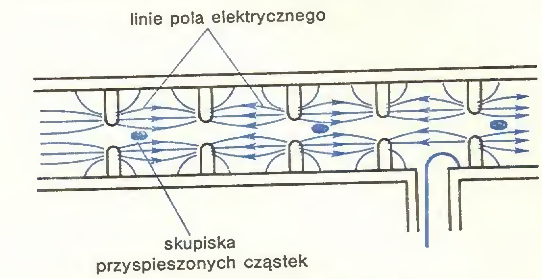
Inaczej sprawa przedstawia się w przypadku ścianek z nadprzewodnika. Opór powierzchniowy, chociaż różny od zera przy wysokich częstościach, jest jednak znacznie mniejszy niż w normalnych metalach, straty energetyczne są nieznaczne i dobroć wnęki może sięgać $5 \cdot 10^9$ przy tej samej częstości 1 GHz. Dobroć wnęki jest funkcją częstości; dla metalu normalnego $Q \sim \sqrt{\nu}$, dla nadprzewodnika $Q \sim 1/\nu$. Jako materiału na ściany wnęki używa się najczęściej ołowiu lub niobu, na ogół w postaci cienkiej warstwy naniesionej na miedziane podłoże.

zależność dobroci od częstości

Przyspieszanie liniowe

Nadprzewodnikowe wnęki rezonansowe znajdują obecnie zastosowanie w liniowych akceleratorach naładowanych cząstek (np. elektronów). Akcelerator liniowy składa się z wielu komór rezonansowych połączonych w szereg (rys. 9). Pole elektromagnetyczne w każdej z wnęk zmienia się okresowo, a faza drgań pola w sąsiednich wnękach różni się o $\pi/2$. Naładowana cząstka, jeśli dostanie się do kanału akceleratora, będzie poddawana w każdej wnęce działaniu pola elektrycznego. Aby to pole mogło zwiększać energię cząstki, musi ona trafiać do wnęki w odpowiedniej

akcelerator nadprzewodnikowy



Rys. 9. Schemat nadprzewodnikowego liniowego akceleratora

fazie drgań pola, czyli czas przelotu cząstki od jednej wnęki do drugiej musi się równać $^1/_2$ okresu drgań.

Ponieważ trudno jest zwiększać stopniowo częstości drgań pola ze wzrostem szybkości cząstki, działaniu pola poddawane są cząstki już przyśpieszone do prędkości bliskiej prędkości światła. Działanie akceleratora sprowadza się wtedy do zwiększania energii cząstki przez zwiększenie masy relatywistycznej.

Zastosowanie doskonałego diamagnetyzmu

Łożyska beztarciowe

Nadprzewodnikowe łożyska beztarciowe działają na zasadzie unoszenia nadprzewodnika przez pole magnetyczne. Siła działająca na jednostkę powierzchni nadprzewodnika ze strony pola magnetycznego, czyli ciśnienie magnetyczne, wynosi $P = {}^1/_2\mu_0 H^2$, gdzie H — natężenie pola przy powierzchni nadprzewodnika. Natężenie to nie może nigdzie przekraczać wartości H_{k1}. Stąd maksymalna wartość P w wypadku nadprzewodnika wykonanego np. z niobu może wynosić 15 000 N/m² w $T = 0$ lub 11 000 N/m² w $T = 4{,}2$ K.

Beztarciowe łożyska można by zastosować w nadprzewodnikowych maszynach elektrycznych, np. w generatorach prądu przemiennego. Dotychczas jednak łożyska te znalazły zastosowanie jedynie w nadprzewodnikowym żyroskopie. Żyroskop taki jest to kula z nadprzewodzącego niobu podtrzymywana w próżni przez pole magnetyczne prądów niezanikających, krążących w dwóch nadprzewodzących pierścieniach. Kula wiruje z prędkością kątową kilkuset obrotów na minutę, praktycznie bez strat energii. Oś obrotu kuli można obserwować przy pomocy przyrządów optycznych. W ruch obrotowy nadprzewodzącą kulę wprowadza się stycznymi do jej powierzchni strumieniami chłodnego gazowego helu. Po nadaniu kuli odpowiedniej prędkości kątowej przestrzeń, gdzie wiruje kula, odpompowuje się do wysokiej próżni.

nadprzewodnikowy żyroskop

Ekrany magnetyczne

Ponieważ ani stałe, ani zmienne pole magnetyczne nie wnika w nadprzewodnik (z wyjątkiem cienkiej

warstwy powierzchniowej), można w nadprzewodzącej osłonie utrzymać przestrzeń wolną od pól magnetycznych. Najkorzystniejszym kształtem dla takiej osłony (ekranu) jest kształt elipsoidalny, gdyż wtedy linie sił pola łagodnie opływają przeszkodę. Jeśli ekran nadprzewodnikowy ma osłaniać przestrzeń roboczą od stałego pola magnetycznego, przed jego oziębieniem do $T < T_k$ należy w miejscu, gdzie następuje oziębianie, skompensować możliwie najdokładniej zewnętrzne pole magnetyczne, aby nie spowodować jego zamrożenia wewnątrz ekranu.

Nadprzewodnikowe ekrany stosuje się najczęściej do osłaniania niskotemperaturowych części czułych przewodów pomiarowych od wpływu zewnętrznych pól elektromagnetycznych.

Zastosowania w elektronice, technice pomiarowej i obliczeniowej

Galwanometry i woltomierze

Pomiar słabych napięć stałych konwencjonalnymi przyrządami sprawia wiele kłopotów z powodu sił termoelektrycznych, powstających w obwodzie pomiarowym, oraz szumów cieplnych w przyrządach pomiarowych, ograniczających ich czułość. Zastosowanie nadprzewodników pozwala na uniknięcie tych przeszkód (siły termoelektryczne w stanie nadprzewodzącym są równe zeru, a szumy cieplne w temperaturze ciekłego helu są bardzo małe). W przypadku zmiennych napięć, osłonięcie niskotemperaturowej części urządzenia pomiarowego ekranem nadprzewodnikowym i zastosowanie nadprzewodnikowych transformatorów wzmacniających sygnał umożliwia zmniejszenie szumów do poziomu nieosiągalnego tradycyjnymi metodami.

galwanometr z nadprzewodzącą cewką

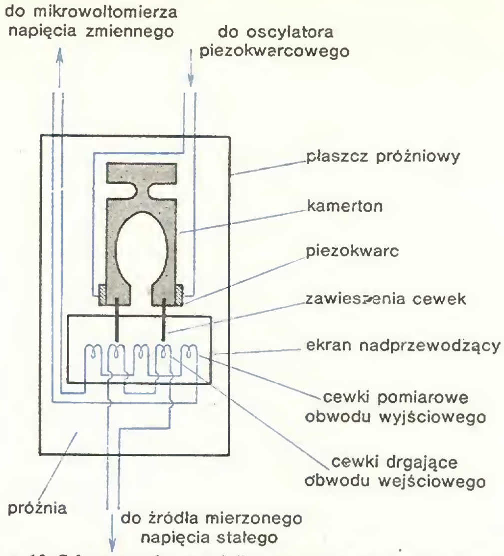
Pierwszą próbą zastosowania nadprzewodników w przyrządach pomiarowych było skonstruowanie w 1952 r. galwanometru z nadprzewodnikową cewką. Dzięki temu można było zmniejszyć opór całego obwodu do $10^{-7}\,\Omega$. Aby przy tak małym oporze R stała czasowa τ przyrządu nie przekraczała sensownej wartości ($\tau = L_{ef}/R$), indukcyjność efektywna (L_{ef}) obwodu powinna być również bardzo mała. Cewka miała więc tylko jeden zwój, a indukcja stałego pola galwanometru wynosiła zaledwie 10^{-6} T. Galwanometr miał stałą czasową $\tau = 15$ s, a czułość prądowa wynosiła 10^{-5} A, co odpowiada czułości napięciowej $10^{-7}\,\Omega \cdot 15^{-5}\,\text{A} = 10^{-12}$ V. Dla porównania przytoczymy, że najczulsze galwanometry konwencjonalne przy takiej samej stałej czasowej mają czułość napięciową rzędu 10^{-8} V.

nadprzewodnikowy przerywacz

Inny typ nadprzewodnikowego urządzenia do pomiaru bardzo słabych napięć stałych skonstruowano w 1955 r. Jest to nadprzewodnikowy przerywacz (modulator) w niskoomowym obwodzie mierzonego napięcia, działający podobnie, jak mechaniczny wibrator (czoper) w konwencjonalnych miernikach napięcia stałego. Przerywaczem jest cienki drucik nadprzewodzący, przechodzący okresowo w stan normalny (czyli przerywający obwód) pod wpływem pola magnetycznego, zmieniającego się sinusoidalnie z częstością 800 Hz. Działanie przerywacza sprowadza się do zamiany mierzonego napięcia stałego na napięcie pulsujące, które po wzmocnieniu nadprzewodnikowymi transformatorami można mierzyć normalnymi przyrządami. Poziom szumów własnych przyrządu udało się zmniejszyć do 10^{-11} V. Stała przyrządu jest rzędu 1 s. Przyrząd nadaje się do pomiaru słabych sił elektromagnetycznych ze źródeł z małym oporem wewnętrznym lub do pomiaru małych oporów (opór reszty obwodu, oprócz przerywacza, musi być znacznie mniejszy od oporu przerywacza w stanie normalnym).

pikowoltomierz z wibrującymi cewkami

Stosunkowo niedawno (1967 r.) zbudowano przyrząd z wibrującymi cewkami do pomiaru napięć stałych (rys. 10) o czułości do 10^{-13} V. Seryjna wersja



Rys. 10. Schemat nadprzewodnikowego pikowoltomierza na napięcie stałe, z wibrującymi cewkami

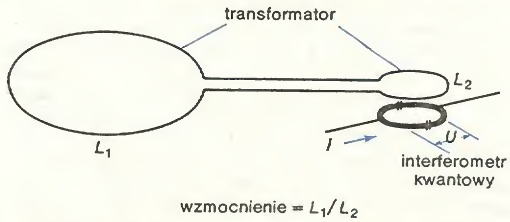
tego przyrządu (dostępna w handlu) ma czułość 10^{-12} V i stałą czasową rzędu kilku s. Para cewek nadprzewodzących połączona ze źródłem mierzonego napięcia stałego jest pobudzona do drgań przez generatory piezokwarcowe, oscyluje z częstością 1 kHz między nieruchomymi nadprzewodzącymi cewkami pomiarowymi i wzbudza w nich napięcie przemienne, a to zaś po wzmocnieniu może być mierzone konwencjonalnymi metodami.

Obecnie bardzo często do pomiarów bardzo słabych napięć używa się tzw. nadprzewodnikowych interferometrów kwantowych, których działanie opiera się na zjawiskach Josephsona. Urządzenia te wykorzystuje się również przy pomiarach słabych zmian pola magnetycznego, dlatego woltomierze działające na tej zasadzie będą omówione łącznie z magnetometrami.

Magnetometry

nadprzewodnikowy interferometr kwantowy z dwoma złączami

Do pomiaru bardzo małych zmian pola magnetycznego można wyzyskać nadprzewodnikowy interferometr kwantowy w kształcie pierścienia z dwoma złączami Josephsona. Zmiany pola powodują zmiany prądu krytycznego w pierścieniu, a więc i zmiany napięcia na nim, te zaś łatwo można zmierzyć. Taki przyrząd umożliwia wykrycie zmian indukcji magnetycznej rzędu 10^{-13} T. Czułość przyrządu można polepszyć przez zastosowanie transformatora nadprze-



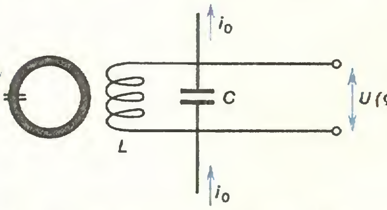
Rys. 11. Schemat nadprzewodnikowego interferometru kwantowego z transformatorem wzmacniającym

wodnikowego na wejściu (rys. 11). Jeśli zmiany pola są większe niż okres zmian krytycznego natężenia prądu dla interferometru kwantowego (→ Zjawiska

tunelowe w nadprzewodnikach), to pomiar tych zmian sprowadza się do liczenia okresów, czyli kwantów strumienia magnetycznego wchodzących do pierścienia lub wychodzących z niego. Niektórym badaczom udało się takim przyrządem naliczyć aż 2000 oscylacji krytycznego natężenia prądu.

nadprzewodnikowy interferometr kwantowy z jednym złączem

Opracowano również nadprzewodnikowy interferometr kwantowy z jednym złączem, który pracuje na prądzie zmiennym o częstości radiowej. Jego czułość jest tego samego rzędu, co interferometru stałoprądowego z podwójnym złączem, ale umożliwia on wzmocnienie sygnału wyjściowego w temperaturze ciekłego helu, a więc bez zwiększania poziomu szumów. Wzmocnienie następuje przez sprzężenie interferometru z obwodem *LC*, zanurzonym w kąpieli helowej (rys. 12). W obwodzie tym płynie prąd i_0

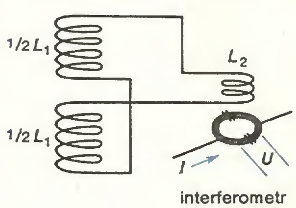


Rys. 12. Schemat zmiennoprądowego interferometru kwantowego

o częstości rezonansowej. Zmiany strumienia magnetycznego w obrębie pierścienia interferometru powodują zmiany krytycznego natężenia prądu pierścienia i modulują amplitudę napięcia w obwodzie *LC* z okresem modulacji równym zmianie strumienia o 1 kwant. Dla wzmocnienia sygnału można też zastosować transformator nadprzewodnikowy na wejściu.

gradiometr

Odmianą transformatora nadprzewodnikowego jest gradiometr (rys. 13), służący do pomiaru bardzo małych gradientów pola magnetycznego. Gradiometr



Rys. 13. Schemat gradiometru

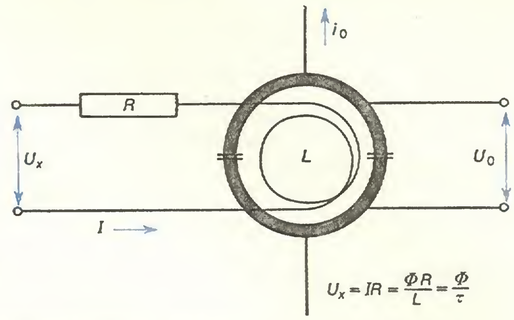
składa się z dwu jednakowych cewek nadprzewodzących, nawiniętych w przeciwnych kierunkach i połączonych nadprzewodzącymi przewodami. Zmiany w czasie jednorodnego pola magnetycznego nie wzbudzają w gradiometrze prądu, ponieważ prądy z obu cewek kompensują się do zera. Różny od zera prąd w obwodzie cewek popłynie tylko wówczas, gdy zajdą przestrzenne zmiany pola magnetycznego. Sygnał z cewek przekazuje się do magnetometru przez sprzężenie indukcyjne.

zastosowanie magnetometru

Czułe magnetometry służą do pomiarów bardzo słabych zmian wielkości magnetycznych. Mierzy się nimi podatność magnetyczną ciał w słabych polach magnetycznych albo fluktuacje podatności magnetycznej nadprzewodników w pobliżu temperatury przemiany T_k. Magnetometry takie można stosować do pomiaru małych przesunięć w ciele magnetycznym przy wykrywaniu fal grawitacyjnych. Podejmowano również z ich pomocą próby wykrycia kwarków. Tego typu magnetometr może być także używany w medycynie i biologii, np. do wykonania magnetokardiogramu (tzn. do rejestrowania słabych zmian pola magnetycznego wywołanych pracą serca) lub do badania bioprądów w żywych organizmach za pośrednictwem wytwarzanych przez nie pól magnetycznych.

Magnetometr można zamienić na woltomierz sprzęgając go indukcyjnie z obwodem mierzonego napięcia (rys. 14). Napięcie U_x wywoła w obwodzie z oporem *R* prąd o natężeniu *I*, który wytworzy w pętli o indukcyjności *L* strumień magnetyczny Φ. Strumień Φ przenika pierścień magnetometru i po-

woltomierz nadprzewodnikowy



Rys. 14. Schemat woltomierza nadprzewodnikowego

woduje zmianę krytycznego natężenia prądu pierścienia (i_k), a więc zmianę napięcia na pierścieniu U_0. Aby stała czasowa przyrządu była możliwie niewielka, pętla powinna mieć małą indukcyjność (mało zwojów). Jeśli jest tylko jeden zwój, a czułość magnetometru jest równa dziesiątej części fluksonu (10^{-16} Wb), to przy stałej czasowej 1 s czułość napięciowa wyniesie 10^{-16} V. Zwykle nie udaje się osiągnąć tej czułości, gdyż szum napięciowy na oporze $R = L/\tau \approx 10^{-8}\ \Omega$ w temperaturze $T = 4{,}2$ K wynosi ok. 10^{-15} V.

Generatory i detektory

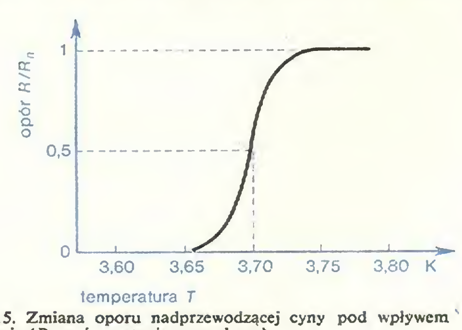
generatory

Złącze Josephsona pod stałym napięciem *U* generuje monochromatyczne i spójne promieniowanie elektromagnetyczne w zakresie mikrofalowym i w dalekiej podczerwieni, tj. w przedziale częstości od 5 do 1000 GHz. Częstość promieniowania ν zależy od napięcia ($\nu = 2eU/h$). Największa dotychczas osiągnięta moc promieniowania wyniosła 10^{-9} W. Przypuszcza się jednak, że będzie można uzyskać 10^{-7} W, a może nawet 10^{-6} W. Szerokość linii może być mniejsza niż 1 kHz dla częstości 10 GHz, co stanowi mniej niż 10^{-7}.

detektory

Złącze Josephsona może być również detektorem promieniowania. Pochłaniane promieniowanie prowadzi do wystąpienia schodków na charakterystyce prądowo-napięciowej złącza. Taki detektor jest równocześnie częstościomierzem, ponieważ napięcia, przy których pojawiają się schodki, wynoszą $U = (nh\nu/2e)$, gdzie *n* jest liczbą całkowitą. Czułość złącza dochodzi do 10^{-10} W.

Większą czułość (10^{-13} W dla $\nu = 70$ GHz) detektora uzyskuje się nie przy wyznaczaniu charakterystyki schodkowej, lecz przy pomiarze zmiany maksymalnego natężenia prądu Josephsona pod wpływem



Rys. 15. Zmiana oporu nadprzewodzącej cyny pod wpływem nagrzania (R_n opór w stanie normalnym)

promieniowania. Promieniowanie periodycznie przerywa się wirującą przesłoną i mierzy się zmiany prądu miernikiem czułym na fazę.

bolometr nadprzewodnikowy

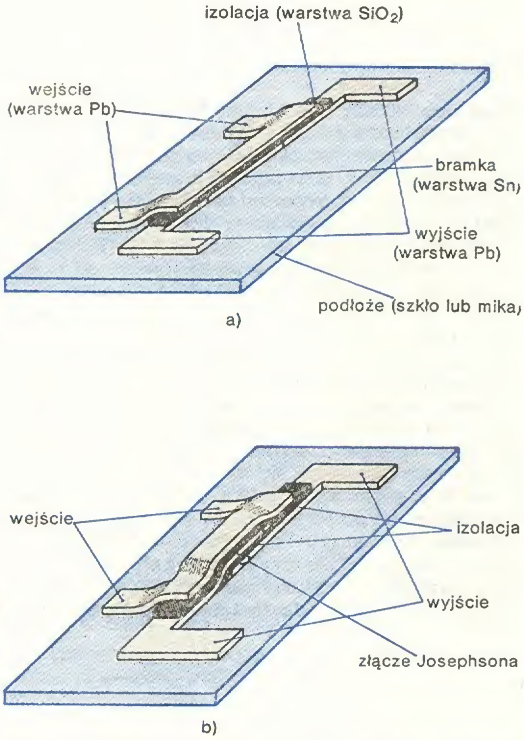
Do detekcji promieniowania może służyć również zwykłe złącze tunelowe (→ Zjawiska tunelowe w nadprzewodnikach — zjawiska niestacjonarne), a także zwykły nadprzewodnikowy bolometr w postaci tasiemki z nadprzewodzącego materiału. Tasiemkę należy utrzymywać w temperaturze odpowiadającej połowie przedziału przejścia w stan oporowy (rys. 15). Pochłaniane promieniowanie powoduje wzrost temperatury bolometru, a zatem i wzrost jego oporu. Czułość nadprzewodnikowego bolometru może sięgać 10^{-12} W. Posługiwanie się nim jest jednak utrudnione z powodu konieczności utrzymywania stałej temperatury w kriostacie w bardzo wąskim przedziale ($\Delta T \approx 0{,}001$ K).

Termometr szumowy

Szumy napięciowe na oporze omowym są funkcją temperatury, można je więc wyzyskać do celów termometrycznych. Ponieważ częstość promieniowania Josephsona jest związana z napięciem na złączu zależnością $h\nu = 2eU$, fluktuacje napięciowe będą modulować częstość Josephsona i wpływać na szerokość pasma promieniowania. Złącze Josephsona podłącza się do źródła słabego napięcia U w szereg z oporem R. Fluktuacje napięciowe na oporze R w temperaturze T dają rozmycie częstości promieniowania Josephsona: szerokość pasma promieniowania emitowanego

kriotron zwykły

kriotron ze złączem Josephsona



Rys. 16. Szkic kriotronu: a) zwykłego, b) ze złączem Josephsona

przez złącze jest więc proporcjonalna do temperatury w skali bezwzględnej: $\Delta\nu = 4\pi kTR/\varphi_0^2$, gdzie φ_0 jest kwantem strumienia magnetycznego. Termometrem szumowym wyznaczono już temperaturę do 20 mK, a przypuszcza się, że można będzie rozszerzyć zakres do 1 mK.

Wzorzec jednostki napięcia

Schodkowa charakterystyka prądowo-napięciowa złącza Josephsona pobudzonego promieniowaniem elektromagnetycznym albo stałym polem magnetycznym może być wykorzystana do ustalenia wzorca jednostki napięcia, opartej na stałych uniwersalnych: ładunku elementarnym e i stałej Plancka h. Stosunek tych stałych, e/h, z innych pomiarów, nie związanych ze zjawiskiem Josephsona, jest znany z dokładnością 10^{-6}, zaś częstość promieniowania można zmierzyć z dokładnością do ok. 10^{-8}; to określa nam dokładność, z jaką możemy ustalić jednostkę napięcia. Doświadczalnie stwierdzono, że stosunek $U/\nu = h/(2e)$ nie zależy od: materiału, typu złącza (tunelowe, punktowe lub inne), temperatury, natężenia pola magnetycznego lub mocy promieniowania, numeru schodka (n), ani innych ubocznych czynników. Ta właśnie niezależność od jakichkolwiek warunków doświadczenia predysponuje złącze Josephsona do oparcia na nim międzynarodowego wzorca jednostki napięcia. Takie wzorce napięcia są już stosowane w wielu krajach; w Polsce opracowuje się taki wzorzec na zlecenie Głównego Urzędu Jakości i Miar.

Elementy maszyn cyfrowych

Na możliwość wyzyskania nadprzewodnictwa w technice obliczeniowej zwrócono uwagę jeszcze w 1956 r. po skonstruowaniu urządzenia zwanego kriotronem (rys. 16). Prąd sterujący w obwodzie wejściowym kriotronu może — za pośrednictwem swego pola magnetycznego — wprowadzać obwód wejściowy kriotronu w jeden z dwu stabilnych stanów: w stan bezoporowy (nadprzewodzący) i w stan oporowy. Dzięki temu kriotron działa jako przerzutnik i można go zastosować w układach logicznych i komórkach pamięci maszyn cyfrowych. Bramkę w obwodzie wyjściowym kriotronu, na której pojawia się lub znika napięcie, może być cienka warstewka nadprzewodnika albo — w czulszych kriotronach — złącze Josephsona.

kriotron

Opracowano także inne nadprzewodnikowe elementy pamięci, jak persistor czy komórka Crove'a. Elementy te wyróżniają się małymi rozmiarami i małym poborem mocy. Mimo tych zalet, nadprzewodnikowe elementy nie znalazły dotychczas tak szerokiego zastosowania w maszynach cyfrowych, jak spodziewano się początkowo. Być może przeszkodą jest niewygoda spowodowana koniecznością pracy w kąpieli helowej. Z oszacowań ekonomicznych wynika, że stosowanie nadprzewodników jest korzystne dopiero w dużych układach pamięci, powyżej stu milionów bitów.

B. B. Goodman *Zastosowania nadprzewodnictwa*, Post. Fiz. 24, 371 (1973).

KRYSZTAŁY

Budowa kryształów

Zygmunt Trzaska Durski

Od czasów najdawniejszych człowiek znajduje w ziemi dziwne naturalne twory o licznych płaskich ścianach, różnorodnych prawidłowych kształtach, pięknych barwach i wspaniałym połysku. Już starożytni nazwali te twory kryształami. Później, w miarę doskonalenia narzędzi badawczych, gdy wynaleziono soczewkę i mikroskop, a następnie, gdy odkryto zjawisko dyfrakcji promieni rentgenowskich na kryształach, oraz gdy zbudowano mikroskop elektronowy — okazało się, że większość naturalnych i otrzymywanych sztucznie substancji w stałym stanie skupienia zbudowana jest z kryształów lub krystalitów nieraz bardzo małych. Wyrażając się ściślej mówimy dziś, że te wszystkie substancje są ciałami krystalicznymi.

występowanie kryształów

I tak np. agregatami ziarn minerałów, najczęściej krystalicznych są skały, a w glebie znajdują się bardzo małe kryształy minerałów ilastych. Ciałami krystalicznymi są śnieg, lód, ziarna piasku oraz prawie wszystkie rudy metali. Z ciał krystalicznych składają się naturalne i sztuczne materiały budowlane (np. granit, wapień, cement). Ciałami krystalicznymi są także produkty hutnicze (metale i ich stopy), większość wytworów przemysłu chemicznego (np. soda, mocznik, saletra, naftalen) i większość środków farmakologicznych (np. aspiryna, witaminy, penicylina) oraz niektóre artykuły spożywcze (cukier, sól kuchenna).

właściwośc. ciał krystalicznych

Ciała krystaliczne mają wiele ciekawych właściwości optycznych, mechanicznych, elektrycznych, magnetycznych, dzięki którym wyróżniają się spośród innych ciał w stałym stanie skupienia (tzw. ciał bezpostaciowych).

Jedną z podstawowych cech kryształów jest ich anizotropia, tzn. zależność właściwości optycznych, mechanicznych czy elektrycznych od kierunku w krysztale. Wszystkie kryształy wykazują anizotropię twardości (tj. mają różną twardość w różnych kierunkach), rozszerzalności cieplnej, przewodnictwa cieplnego, przewodnictwa elektrycznego. We wszystkich też kryształach — z wyjątkiem tych, które należą do układu regularnego — występuje zjawisko podwójnego załamania światła. Wiele kryształów ma zdolność skręcania płaszczyzny polaryzacji światła. Kryształy niektórych substancji wykazują piezoelektryczność i piroelektryczność. Ciała krystaliczne mają ściśle określoną temperaturę topnienia przy określonym ciśnieniu (gdy jednocześnie nie zachodzi chemiczny rozkład substancji).

zastosowania ciał krystalicznych

Wymienione wyżej właściwości ciał krystalicznych wykorzystuje się szeroko w nauce i technice. Kryształy „wkroczyły" również do naszego życia codziennego: bez specjalnych kryształów nie działałyby m.in. gramofony elektryczne, radia tranzystorowe, magnetofony, telewizory czy nawet zegarki (te „na kamieniach"). Z niektórych kryształów (kwarc, sól kamienna) przezroczystych dla promieni nadfioletowych wykonuje się pryzmaty stosowane w spektrografach optycznych, a np. z kryształu fluorytu przezroczystego w podczerwieni — soczewki do obiektywów noktowizorów. Podwójne załamanie światła w kryształach kalcytu wykorzystuje się w przyrządach polaryzacyjnych (np. pryzmaty Nicola). Niektóre ciała krystaliczne wykazujące silną anizotropię absorpcji światła służą do sporządzania błon polaryzujących światło (polaroidów). Luminescencję niektórych kryształów wykorzystano w licznikach cząstek jonizujących (liczniki scyntylacyjne). Monokryształy rubinu zastosowano do budowy maserów i laserów. Kryształy piezoelektryczne stosuje się m.in. do wytwarzania ultradźwięków, do pomiarów ciśnień (np. w cylindrach silników, w górotworze) oraz jako podstawowe części przetworników elektroakustycznych. Kryształy półprzewodnikowe (np. krzem, german) znalazły zastosowanie w elektronice. Kryształy o wysokiej twardości mają również różnorodne zastosowanie, np. diamenty służą m.in. do wyrobu narzędzi wiertniczych i tarcz szlifierskich oraz do cięcia szkła, a z rubinów wykonuje się łożyska (w precyzyjnych przyrządach pomiarowych, np. w busolach i zegarkach). Kryształy mające piękne barwy i połysk służą do wyrobu klejnotów i ozdób (np. diamenty, rubiny, szafiry, szmaragdy, ametysty).

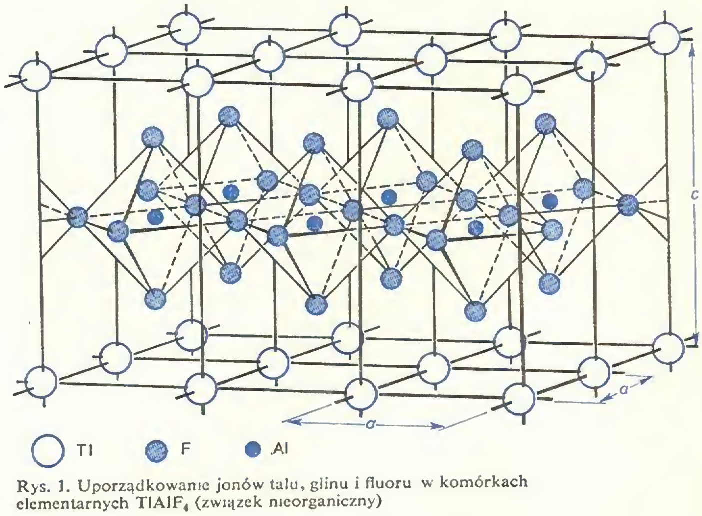
Współczesna nauka i technika potrzebują ciągle nowych, o coraz to innych właściwościach, ciał krystalicznych. Na potrzeby techniki nie wystarczają już kryształy naturalne, które często są albo zbyt zanieczyszczone różnymi domieszkami albo są mało doskonałe ze względu na występujące w nich defekty budowy. Stąd konieczność sztucznego otrzymywania, jak mówimy — „hodowania" — różnych monokryształów (→ Otrzymywanie monokryształów), czasami o niemal idealnej budowie, często o nowych, nie znanych dotąd a nieraz wprost zaskakujących właściwościach.

ciała krystaliczne

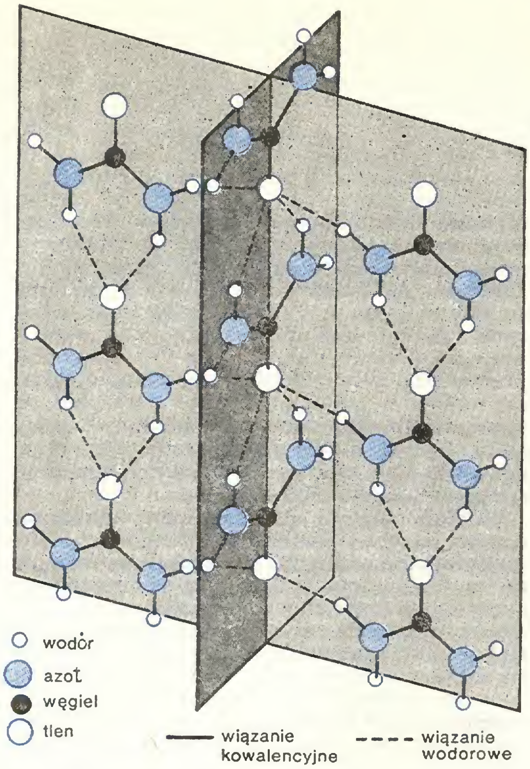
Cóż to jest więc ciało krystaliczne? Jak możemy je zdefiniować? Czym różni się od innych tworów naturalnych, jak zwierzęta i rośliny, czy też od ciał bezpostaciowych, ciekłych lub gazowych? Wyżej użyto takich terminów, jak ciało krystaliczne, kryształ, krystalit, monokryształ. Czym te ciała różnią się między sobą?

Wszystkie właściwości kryształów (a ogólniej ciał krystalicznych), łącznie z ich postacią zewnętrzną i symetrią, są wynikiem — jak to dzisiaj wiemy — specyficznej budowy wewnętrznej kryształów, mianowicie nieskończonego trójwymiarowego uporządkowania atomów lub jonów czy cząsteczek. Przykłady takiego uporządkowania (w związku nieorganicznym i organicznym) pokazano na rys. 1 i 2.

Ciekawe, że już pierwsi badacze kryształów, wiedzeni genialną intuicją, starali się wytłumaczyć powstawanie wysoko symetrycznych postaci kryształów i tworzenie się płaskich ścian — specyficzną budową wewnętrzną. Potrzebne im jednak były do tego celu pewne podstawowe cząstki — elementy, z których



Rys. 1. Uporządkowanie jonów talu, glinu i fluoru w komórkach elementarnych $TlAlF_4$ (związek nieorganiczny)



Rys. 2. Uporządkowanie cząsteczek $CO(NH_2)_2$ w krysztale mocznika (związek organiczny)

hipotezy wewnętrznej budowy kryształów

próbowali odtworzyć kryształy. I tak np. J. Kepler (1611) przypuszczał, że sześciokątne płatki śniegu zbudowane są z kulistych cząstek wody, ściśle stykających się z sobą, a R. Hooke (1665) odtworzył zewnętrzne postacie kryształów ałunu i soli kamiennej układając warstwami jednakowe kulki (należy zauważyć, że dzieje się to ok. 150 lat przed powstaniem teorii atomistycznej J. Daltona — 1803). Ch. Huygens (1690) przypisywał najmniejszym cząstkom kryształów kalcytu kształt elipsoidy obrotowej. Również i M.W. Łomonosow (1749) próbował wytłumaczyć wielościanową postać kryształów prawidłowym ułożeniem w przestrzeni kulistych cząstek elementarnych (korpuskuł). D. Guglielmini (1688) i R. J. Haüy (1782) na podstawie obserwacji łupliwości kryształów wysunęli przypuszczenie, że najmniejsze cząstki, z których zbudowane są kryształy, mają postać wielościanów.

Dziś wiemy, że podstawowe elementy, z których zbudowane są kryształy (i in. ciała krystaliczne), mają postać równoległościanów (udowodnił to teoretycznie w 1850 r. A. Bravais, a doświadczalnie potwierdził M. Laue w 1912 r.), a one z kolei utworzone są z atomów (jonów, cząsteczek). Obecnie ciałem krystalicznym nazywa się każde ciało w stałym stanie skupienia mające uporządkowaną, prawidłową (sieciową) budowę wewnętrzną. Zależnie od warunków krystalizacji ciała krystaliczne mogą tworzyć monokryształy lub ciała polikrystaliczne. Ciało polikrystaliczne (polikryształ) składa się z licznych, mikroskopowej wielkości kryształów lub krystalitów. Monokryształem jest każdy pojedynczy, duży lub nawet bardzo mały kryształ lub krystalit, nie wykazujący zrostów i pęknięć oraz nie posiadający wrostków innych substancji.

polikryształ, monokryształ

Jak już wspomniano, naturalne kryształy, np. rubinu, mają płaskie ściany, natomiast otrzymane w piecu Verneuila monokryształy rubinu mają kształt gruszki. Otrzymane sztucznie „gruszki rubinowe” są ciałami krystalicznymi o takim samym składzie chemicznym jak naturalny kryształ rubinu. Wspólną cechą łączącą naturalny kryształ rubinu i syntetyczną „gruszkę rubinową” — krystalit — jest więc nie ich wygląd zewnętrzny, lecz ich jednakowa — przy jednakowym składzie chemicznym — budowa wewnętrzna. Tak więc kryształem nazywa się ciało w stałym stanie skupienia, o prawidłowej (sieciowej) budowie wewnętrznej mające naturalną postać wielościanową; natomiast krystalit, mający taką samą budowę wewnętrzną jak kryształ, ograniczony jest dowolną powierzchnią.

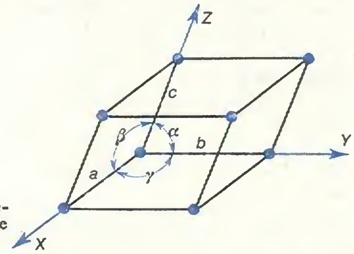
kryształ a krystalit

Wewnętrzna budowa ciał krystalicznych

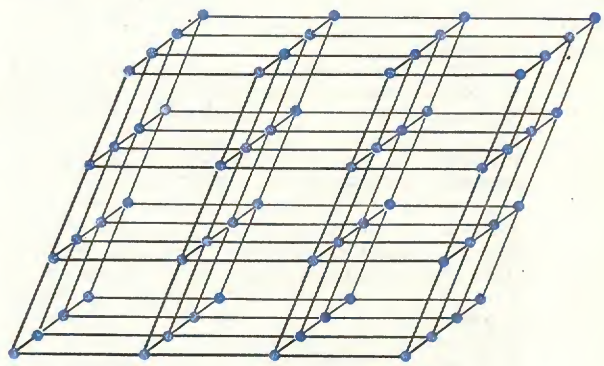
Sieć przestrzenna

Podstawowe jednostki, z których zbudowane są ciała krystaliczne, mają kształt równoległościanów zw. komórkami elementarnymi (rys. 3). W danym ciele krystalicznym wszystkie komórki elementarne są oczywiście jednakowe. Nieskończony zbiór jednakowych komórek elementarnych, ułożonych względem siebie równolegle i ściśle wypełniających przestrzeń stanowi sieć przestrzenną (rys. 4). Mówimy, że sieć

komórka elementarna



Rys. 3. Komórka elementarna i jej stałe sieciowe



Rys. 4. Sieć przestrzenna

przestrzenna jest geometrycznym, trójwymiarowym schematem wewnętrznej budowy kryształów (ciał krystalicznych). Według tego schematu ułożone są w krysztale wszystkie atomy, jony lub cząsteczki. Oczywiście, sieć przestrzenna jako pewien schemat jest pojęciem abstrakcyjnym i żadnej „sieci” — jako takiej — w rzeczywistym krysztale nie ma.

sieć przestrzenna

Rozmieszczenie atomów (jonów, cząsteczek) w pojedynczej komórce elementarnej nazywamy strukturą kryształu (strukturą ciała krystalicznego, strukturą krystaliczną). Natomiast powtarzające się periodycznie w trzech wymiarach przestrzeni (tj. wg schematu sieci przestrzennej), wypełnione atomami (jonami, cząsteczkami) komórki elementarne tworzą sieć krystaliczną. Mówiąc inaczej: sieć przestrzenna jest „szkieletem”, który po wypełnieniu w określony sposób atomami (jonami, cząsteczkami) staje się siecią krystaliczną.

struktura kryształu

sieć krystaliczna

W sieci przestrzennej krawędzie komórek elementarnych przecinają się w punktach zw. węzłami sieciowymi. Węzłami nazywa się też punkty na prostych

węzły sieciowe

sieciowych nie będących krawędziami komórek elementarnych (np. na środkach ścian komórki elementarnej, rys. 14). Sieć przestrzenną można traktować nie tylko jako nieskończony zbiór komórek elementarnych, ale i jako nieskończony uporządkowany zbiór punktów — węzłów.

prosta sieciowa

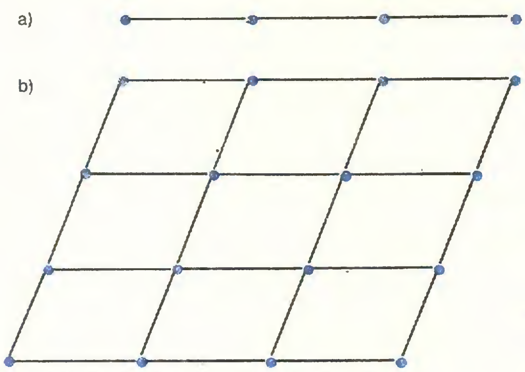
Prosta przeprowadzona przez dwa dowolne węzły jest prostą sieciową (rys. 5a). Na prostej sieciowej znajduje się nieskończona liczba węzłów w jednakowych od siebie odległościach.

periody identyczności

Odległości te nazywamy periodami identyczności (odległościami translacyjnymi).

płaszczyzna sieciowa

Przez trzy, nie leżące na jednej prostej, węzły przechodzi płaszczyzna sieciowa (rys. 5b). Takich płaszczyzn i prostych można wybrać



Rys. 5. Prosta sieciowa (a) i płaszczyzna sieciowa (b)

w sieci przestrzennej nieskończenie wiele.

stałe sieciowe

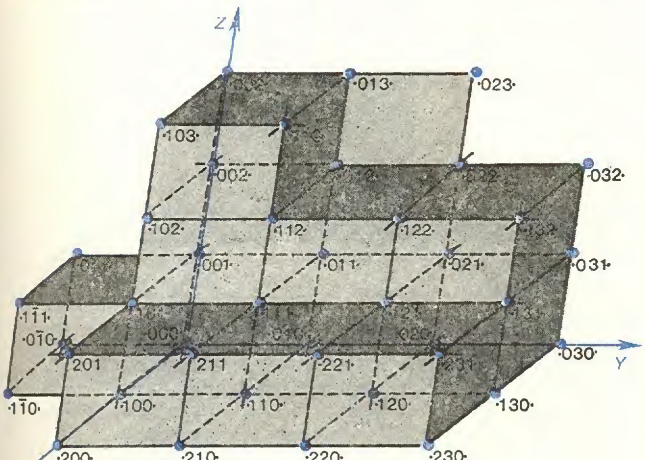
Kształt i rozmiary komórki elementarnej określane są przez tzw. stałe sieciowe (parametry sieciowe), którymi są długości krawędzi komórki a, b, c oraz kąty α, β, γ między tymi krawędziami (rys. 3). Sieć przestrzenna jest jednoznacznie określona przez jej komórkę elementarną. Oznacza to, że wystarczy znać stałe sieciowe, by móc odtworzyć sieć przestrzenną.

współrzędne węzła sieci

W sieci przestrzennej mamy do czynienia z nieskończoną liczbą węzłów. Położenie (pozycję) wybranego węzła określa się względem układu osi krystalograficznych X, Y, Z za pomocą współrzędnych $\cdot xyz \cdot$ (rys. 6).

osie krystalograficzne

Osie krystalograficzne pokrywają się z trzema nierównoległymi do siebie krawędziami komórki elementarnej.



Rys. 6. Współrzędne węzłów prymitywnej sieci przestrzennej

Sieci Bravais'go

układy krystalograficzne

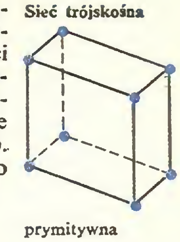
Biorąc pod uwagę różne możliwe wartości stałych sieciowych oraz symetrię sieci przestrzennych można sieci przestrzenne podzielić na sześć głównych typów, zwanych układami krystalograficznymi (tab. str. 447).

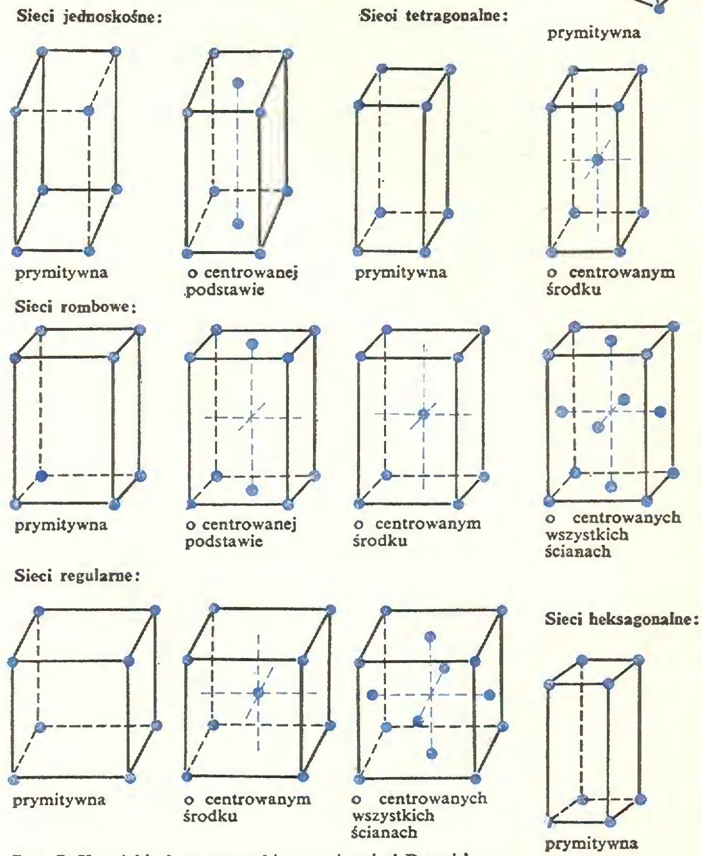
Prócz sieci przestrzennych mających węzły tylko w narożach komórek elementarnych istnieją jeszcze sieci — również spełniające warunki wymienione w tabeli, w których węzły występują także na środkach ścian lub w środku geometrycznym komórki elementarnej. Pierwsze z wymienionych wyżej sieci nazywają się sieciami prymitywnymi, a pozostałe — sieciami centrowanymi.

sieci Bravais'go

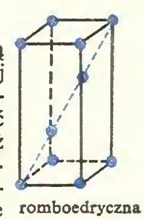
Istnieje 14 różnych typów sieci prymitywnych i centrowanych, nazywają się one sieciami Bravais'go (lub sieciami translacyjnymi). Podobnie jak sieci, również ich komórki elementarne nazywa się prymitywnymi lub centrowanymi (rys. 7). Często zamiast mówić o 14 typach sieci Bravais'go

Sieć trójskośna





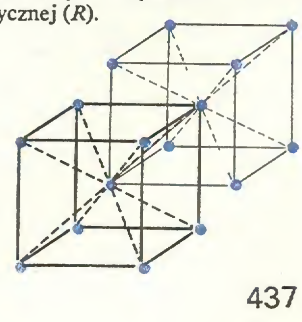
Rys. 7. Komórki elementarne 14-tu typów sieci Bravais'go



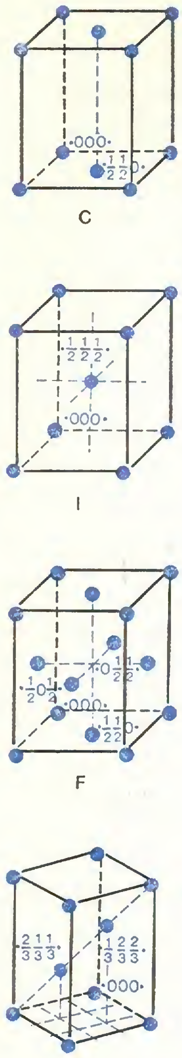
mówi się o 14 typach komórek elementarnych. Cechą charakterystyczną sieci Bravais'go jest to, że w takiej sieci każda prosta sieciowa jest równomiernie obsadzona węzłami. Można ten warunek wypowiedzieć też inaczej, mianowicie: przesunięcie całej sieci wzdłuż dowolnej prostej sieciowej na odległość równą periodowi identyczności tej prostej (lub jego wielokrotności) prowadzi zawsze do pokrycia się sieci samej ze sobą (do nałożenia się na siebie wszystkich węzłów).

symbole sieci

Warunek charakteryzujący sieć Bravais'go spełniany jest w sieciach prymitywnych (symbol P) oraz w sieciach o centrowanych podstawach (symbol A, B lub C — zależnie od pary centrowanych ścian komórki elementarnej), w sieciach o centrowanym środku (I), w sieciach o centrowanych wszystkich ścianach (F) i w sieci romboedrycznej (R).



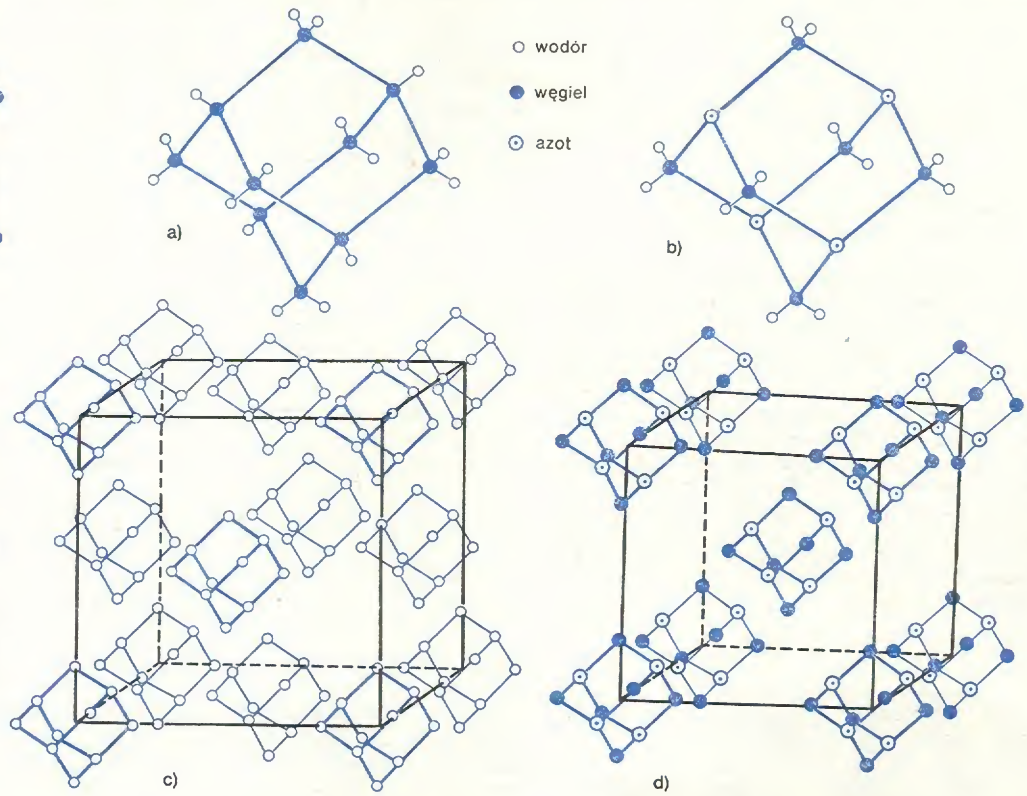
Rys. 8. Sieć Bravais'go I jako zespół wstawionych w siebie dwóch sieci prymitywnych



Rys. 9. Współrzędne węzłów w centrowanych sieciach Bravais'go *C, I, F R*

Każdą sieć centrowaną można zawsze rozpatrywać jako sieć złożoną z dwu, trzech lub czterech wstawionych w siebie i odpowiednio przesuniętych identycznych sieci prymitywnych (rys. 8). W komórce o centrowanych podstawach mamy dwie sieci prymitywne wstawione w siebie, przy czym węzeł początkowy jednej z nich ma współrzędne ·000·, a drugiej $\cdot\frac{1}{2}\frac{1}{2}0\cdot$. W sieci typu *I* współrzędne węzłów początkowych sieci wstawionych są ·000· i $\cdot\frac{1}{2}\frac{1}{2}\frac{1}{2}\cdot$ (dwie sieci wstawione w siebie); w sieci typu *F*— ·000·, $\cdot\frac{1}{2}\frac{1}{2}0\cdot$, $\frac{1}{2}0\frac{1}{2}\cdot$ i $\cdot 0\frac{1}{2}\frac{1}{2}\cdot$ (cztery sieci wstawione w siebie); w sieci typu *R*— ·000·, $\cdot\frac{2}{3}\frac{1}{3}\frac{1}{3}\cdot$, $\cdot\frac{1}{3}\frac{2}{3}\frac{2}{3}\cdot$ (trzy sieci wstawione w siebie) (rys. 9).

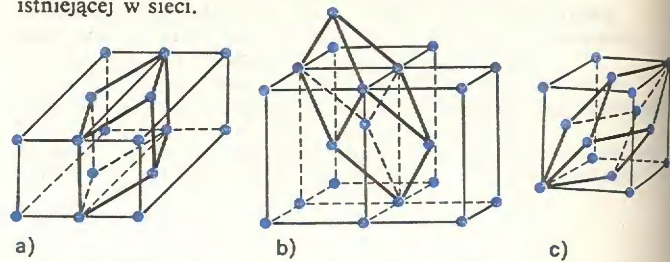
Znajomość typu sieci Bravais'go w danym krysztale daje pewną informację o sposobie rozmieszczenia atomów w komórce elementarnej. Jeżeli np. w krysztale, mającym sieć przestrzenną typu *F*, ugrupowanie *n* atomów znajduje się wokół węzła ·000·, to takie same grupy *n* atomów i w takim samym położeniu (orientacji) muszą znajdować się wokół węzłów $\cdot\frac{1}{2}0\frac{1}{2}\cdot$, $\cdot 0\frac{1}{2}\frac{1}{2}\cdot$, $\cdot\frac{1}{2}\frac{1}{2}0\cdot$, a w krysztale o sieci typu *I* takie same ugrupowania atomów i w takiej samej orientacji znajdują się wokół węzłów ·000· i $\cdot\frac{1}{2}\frac{1}{2}\frac{1}{2}\cdot$. Ilustrują to dobrze struktury dwóch organicznych związków: adamantanu $C_{10}H_{16}$ i urotropiny $(CH_2)_6N_4$ (kształt ich cząsteczek oraz struktury utworzonych z nich kryształów przedstawia rys. 10). Kształt cząsteczek obu związków jest taki sam, ale sieć przestrzenna kryształu adamantanu jest siecią regularną typu *F*, a kryształu urotropiny — regularną typu *I*.



Rys. 10. Kształt cząsteczek adamantanu (a) i urotropiny (b) oraz struktury kryształów (atomy wodoru na rysunkach pominięto) adamantanu (c) i urotropiny (d). Cząsteczka adamantanu zbudowana jest z 6 grup CH_2 i 4 grup CH, w cząsteczce urotropiny 4 grupy CH zastąpione są atomami azotu N

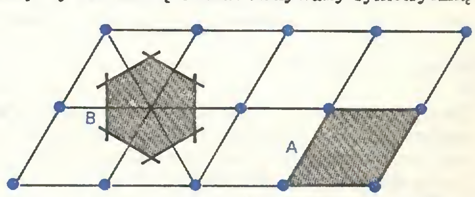
komórka Wignera-Seitza

W sieci przestrzennej kryształu istnieje możliwość wyboru równoległościennej komórki elementarnej w różny sposób. Zazwyczaj wybiera się ją tak, aby kształt jej ścian odzwierciedlał symetrię sieci, oraz aby jej objętość była minimalna, liczba kątów prostych między krawędziami była maksymalna, a węzły znajdowały się w położeniach zgodnych z jedną z 14 komórek brawesowskich. Dla każdej sieci o centrowanej brawesowskiej komórce elementarnej można jednak zawsze utworzyć komórkę prymitywną, niecentrowaną. Komórki prymitywne utworzone dla sieci centrowanych (przykłady takich komórek pokazuje rys. 11) zwykle nie oddają rzeczywistej symetrii istniejącej w sieci.



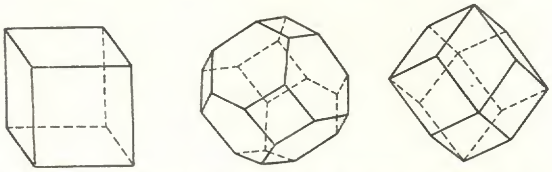
Rys. 11. Przekształcanie komórek centrowanych w prymitywne: a) komórki o centrowanych podstawach (typ *C*); b) o centrowanym środku (typ *I*); c) o centrowanych wszystkich ścianach (typ *F*)

Metodę wydzielania symetrycznych komórek prymitywnych we wszystkich sieciach przestrzennych podali Wigner i Seitz. W celu utworzenia takiej komórki dany węzeł sieci przestrzennej łączy się odcinkami prostymi z najbliższymi węzłami. Następnie przez środki tych odcinków prowadzi się płaszczyzny do nich prostopadłe (rys. 12). Przecinając się ze sobą płaszczyzny te tworzą obszar nazywany symetryczną



Rys. 12. Komórka elementarna Bravais'go (*A*) i komórka prymitywna Wignera-Seitza (*B*) dwuwymiarowej sieci heksagonalnej

komórką prymitywną lub komórką Wignera–Seitza (przykłady — rys. 13). Komórki te w sieci przestrzennej są wielościanami, którymi można ściśle, bez przerw, wypełnić przestrzeń. Komórki Wignera–Seitza dla regularnych sieci typu *P*, *I* oraz *F* przedstawiono na rys. 13. Komórki Wignera–Seitza znajdują zastosowanie np. w teorii stref Brillouina (→ Struktura elektronowa ciał stałych), w której pierwszą strefą Brillouina nazywa się komórkę Wignera–Seitza utworzoną w sieci odwrotnej (→ Krystalografia rentgenowska).



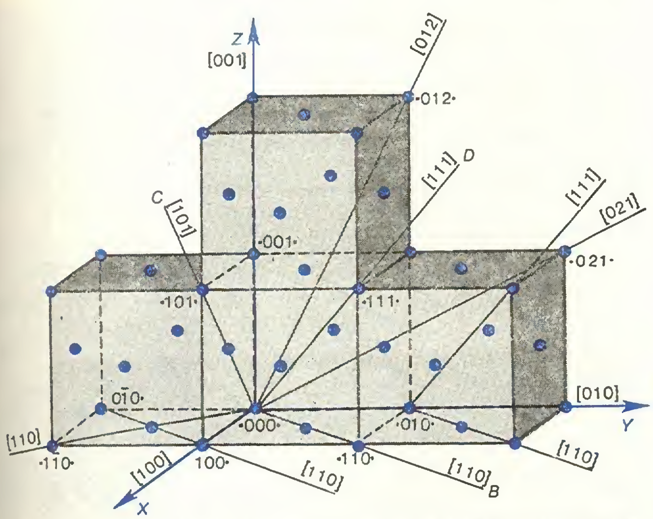
Rys. 13. Komórki Wignera–Seitza regularnych sieci typu *P*, *I* oraz *F*

Symbole prostych i płaszczyzn sieciowych

Przez nieskończoną liczbę węzłów znajdujących się w sieci przestrzennej możemy prowadzić nieskończoną liczbę prostych sieciowych i płaszczyzn sieciowych. Płaszczyzny sieciowe w sieci przestrzennej nachylone różnie w stosunku do osi krystalograficznych różnią się między sobą zasadniczo sposobem rozmieszczenia węzłów i odległościami od sąsiednich płaszczyzn sieciowych. Również i proste sieciowe poprowadzone z początku układu osi współrzędnych (osi krystalograficznych) są różne, np. węzły znajdują się na nich w różnych odległościach, a i odległości między sąsiednimi równoległymi prostymi są różne.

Weźmy np. sieć krystaliczną złota (układ regularny) i wybierzmy w tej sieci komórkę elementarną w taki sposób, aby atomy złota znajdowały się w narożach komórki i na środkach ścian (rys. 14, 16). Taki wybór komórki elementarnej nie jest koniecznością, gdyż

komórka elementarna złota

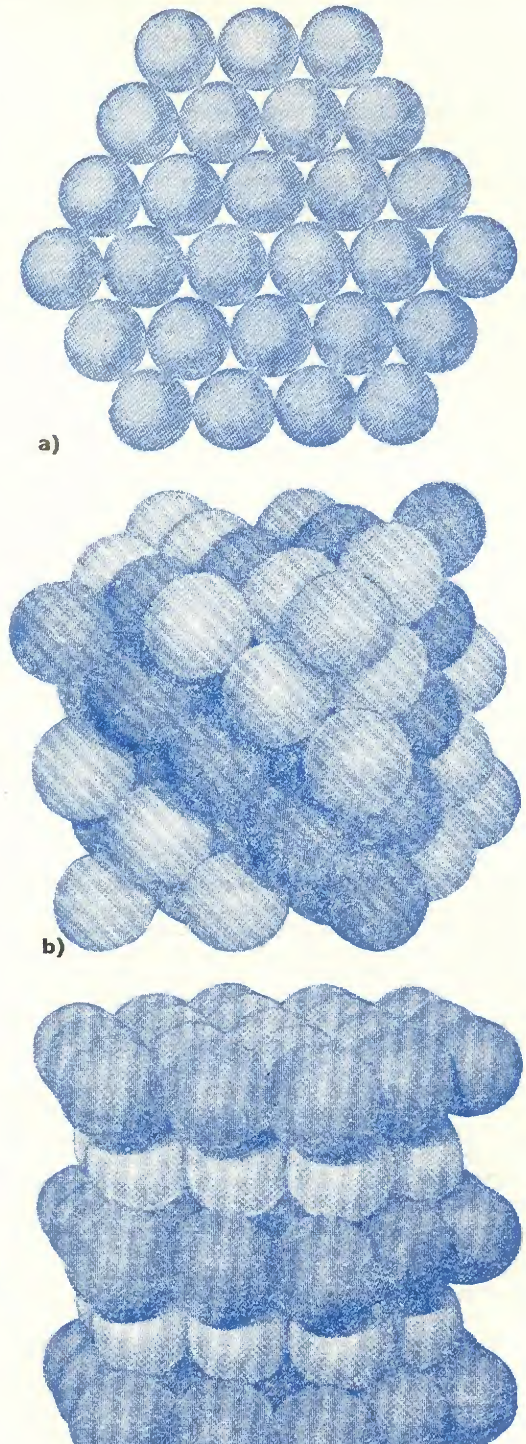


Rys. 14. Przykłady prostych sieciowych w sieci krystalicznej złota

można ją wybrać i tak, aby w jej narożach nie znajdował się żaden atom (np. jak na rys. 10). Taki sam typ sieci (*F*) jak złoto mają też aluminium, miedź, nikiel i in.

wskaźniki prostej sieciowej

W komórce elementarnej złota atomy leżące na prostej *X* znajdują się w odległości *a* od siebie; ale na prostej sieciowej *B* atomy leżą w odległościach $a\sqrt{2}/2$, a na prostej *D* w odległości $a\sqrt{3}$. Proste te są więc różne; oznaczamy je tzw. wskaźnikami, tj. 3 liczbami całkowitymi *m*, *n*, *p* względem siebie pierwszymi, ujętymi w nawias kwadratowy: [*mnp*]. Liczby *m*, *n*, *p* jednoznacznie określają kierunek prostej sieciowej. Wskaźniki prostej sieciowej przechodzącej przez węzeł ·000· i jakiś drugi są takie same jak wskaź-

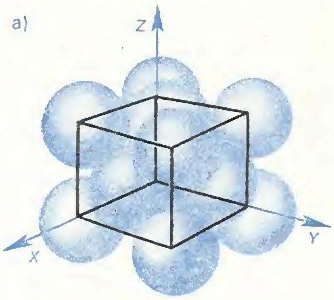


Rys. 15. Najgęstsze ułożenia jednakowych kul w przestrzeni, a) najgęściej upakowana warstwa heksagonalna, b) najgęstsze ułożenie regularne (heksagonalne warstwy najgęściej upakowane są równoległe do warstwy kul zaciemnionych), c) najgęstsze ułożenie heksagonalne (warstwy najgęstszego upakowania położone są poziomo)

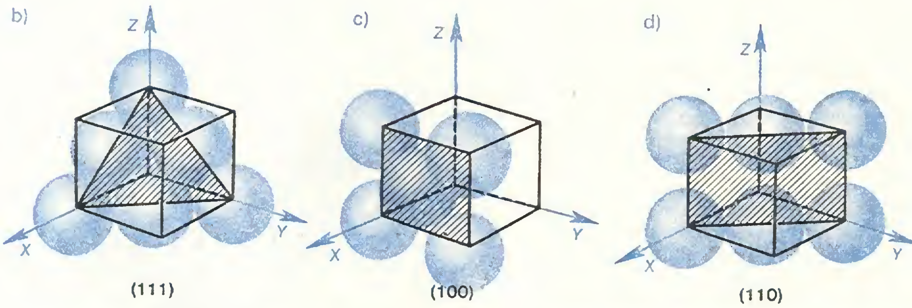
niki tego drugiego węzła sprowadzone do postaci liczb całkowitych. Na przykład prosta X przechodząca przez węzeł $\cdot 000 \cdot$ i $\cdot 100 \cdot$ ma symbol [100], prosta B przechodząca przez $\cdot 000 \cdot$ i $\cdot \frac{1}{2}\frac{1}{2}0 \cdot$ ma symbol [110], prosta C — symbol [101], a prosta D — [111]. W sieci przestrzennej wszystkie proste sieciowe równoległe do siebie mają jednakowe wskaźniki.

struktury najgęstszego ułożenia

Przypatrzmy się teraz płaszczyznom sieciowym wyznaczonym w strukturze kryształu złota przez różne ugrupowania atomów. W krysztale złota (miedzi, niklu) wszystkie atomy mają jednakowy kształt — przyjmujemy, że kulisty — i tworzą tzw. strukturę najgęstszego ułożenia (upakowania; rys. 15). W strukturach tego typu, jednakowych rozmiarów atomy, jony a nawet cząsteczki ułożone są w sposób zapewniający możliwie największe wypełnienie przestrzeni



Rys. 16. Kryształ złota: a) struktura najgęstszego upakowania; b), c) i d) obsadzenie atomami płaszczyzn sieciowych (111), (100) i (110)



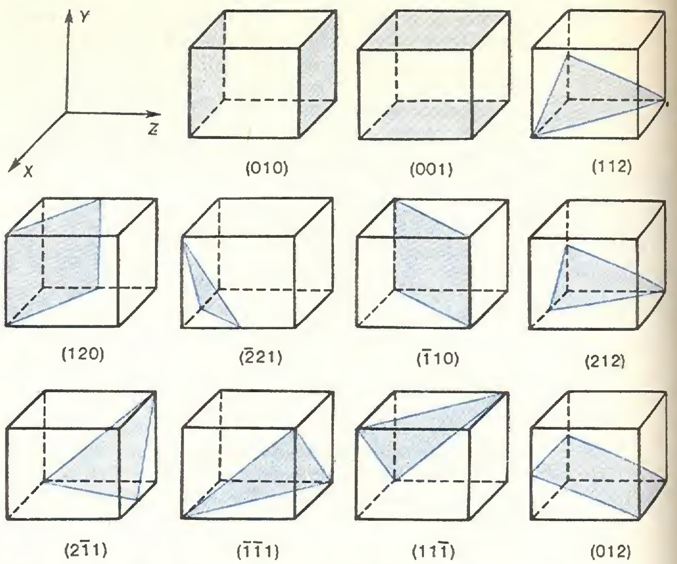
(ok. 74%). Struktury te powstają w wyniku nakładania jedna na drugą najgęściej upakowanych warstw heksagonalnych. Największe znaczenie w teorii struktur kryształów mają dwa najgęstsze ułożenia — regularne i heksagonalne. Najgęstsze ułożenie regularne ma sieć Bravais'go typu F, a heksagonalne — typu P, przy czym w komórce elementarnej tej ostatniej sieci znajdują się dwa węzły w pozycjach $\cdot 000 \cdot$ i $\cdot \frac{2}{3}\frac{1}{3}\frac{1}{2} \cdot$.

Na rysunku 16b, c, d pokazane są trzy płaszczyzny sieciowe przechodzące przez komórkę elementarną kryształu złota. Na tych płaszczyznach wyraźnie widać różne ich obsadzenie przez atomy złota. W płaszczyźnie sieciowej pokazanej na rys. 16b atomy złota tworzą najgęściej upakowaną warstwę; atomy w tej płaszczyźnie ściśle przylegają do siebie. Na następnych dwóch rysunkach (16c, d) wyraźnie widać, że rozmieszczenie atomów w zaznaczonych płaszczyznach jest inne, a przede wszystkim znacznie luźniejsze niż w pierwszym wypadku.

Zauważmy też, że każda z zaznaczonych płaszczyzn jest inaczej położona w stosunku do osi krystalograficznych. Aby te płaszczyzny móc od siebie odróżniać i jakoś „nazywać" zaopatrujemy je w pewne symbole. Oznaczamy je symbolami Millera (hkl) złożonymi z trzech liczb całkowitych h, k, l (wskaźniki Millera) względem siebie pierwszych.

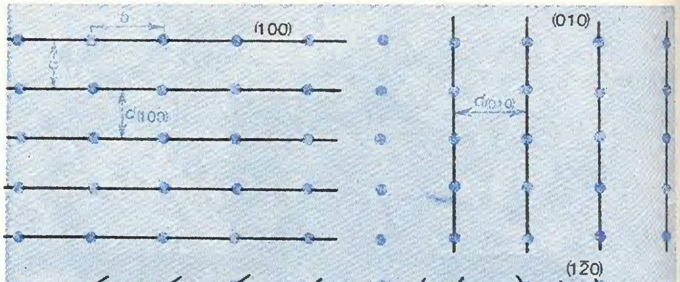
symbole i wskaźniki Millera

Wskaźniki h, k, l pokazują, ile razy odcinki odcięte na osiach krystalograficznych przez daną płaszczyznę są mniejsze od periodów identyczności wzdłuż odpowiednich osi. Zgodnie z tą definicją płaszczyzny sieciowe pokazane na rys. 16 mają symbole (111),



Rys. 17. Położenie różnych płaszczyzn sieciowych w komórce elementarnej

(100) i (110). Położenie kilku innych płaszczyzn sieciowych przedstawione jest na rys. 17.



Rys. 18. Rzut sieci przestrzennej na płaszczyznę XY (oś Z jest prostopadła do płaszczyzny rysunku) z zaznaczonymi śladami niektórych płaszczyzn sieciowych ($hk0$)

Wskaźnik 0 w symbolu płaszczyzny sieciowej oznacza, że płaszczyzna ta jest równoległa do związanej z tym wskaźnikiem osi krystalograficznej (np. wskaźnik h związany jest z osią X itd.). Jeżeli płaszczyzna sieciowa przecina ujemny zwrot danej osi krystalograficznej, to ujemny jest również odpowiedni wskaźnik; znak „minus" umieszcza się nad wskaźnikiem, np. $(\bar{1}\bar{1}1)$.

Wszystkie płaszczyzny sieciowe równoległe do siebie mają takie same wskaźniki Millera. Dla zespołu równoległych płaszczyzn sieciowych określa się odległość międzypłaszczyznową $d_{(hkl)}$, tj. odległość między dwoma sąsiednimi płaszczyznami, mierzoną wzdłuż prostej prostopadłej do tych płaszczyzn (rys. 18). Na rysunku łatwo można zauważyć, że im wyższe są wskaźniki, tym sąsiednie płaszczyzny znajdują się bliżej siebie.

odległość międzypłaszczyznowa

Symetria kryształów

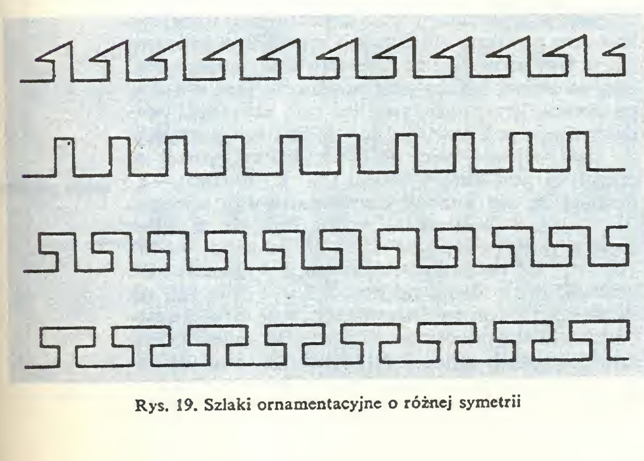
Symetria

Patrząc na szlaki ornamentacyjne pokazane na rys. 19, intuicyjnie wyczuwamy, że mają one jakąś symetrię, że są symetryczne. Od razu rzuca się w oczy pewne uporządkowanie i powtarzanie się podstawowych „wzorów" w tych ornamentach.

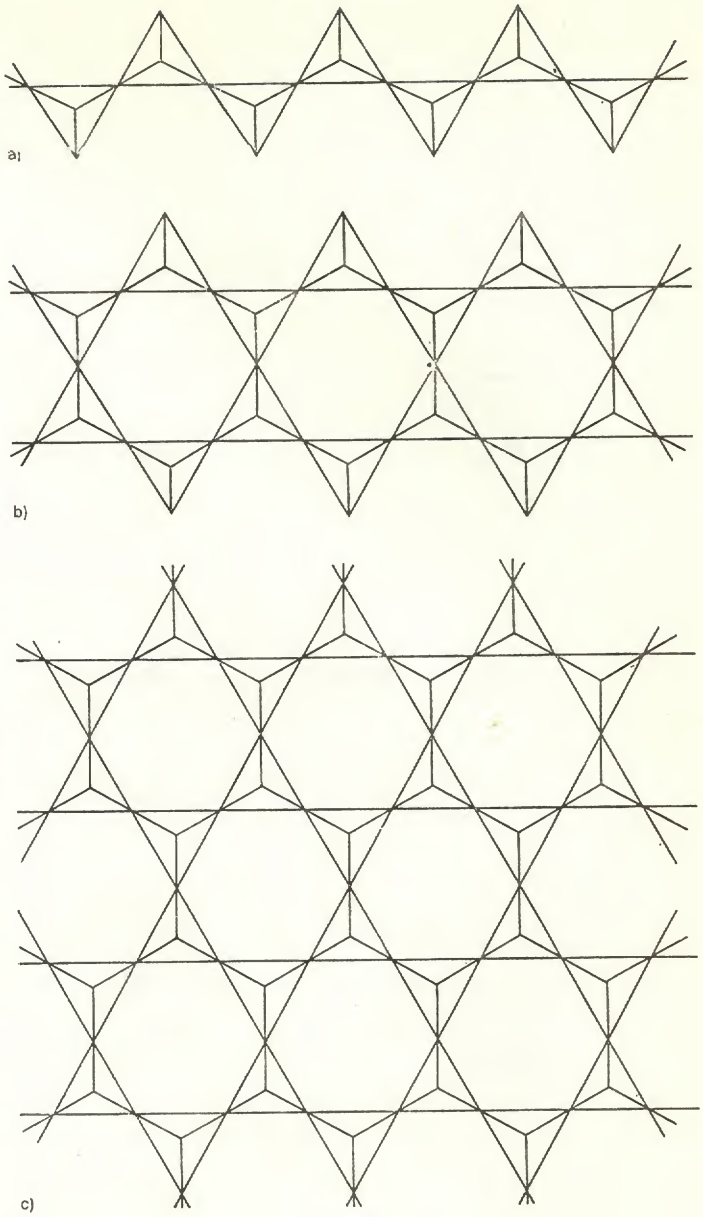
Przyjrzyjmy się teraz np. uporządkowaniu atomów krzemu i tlenu (rys. 20) w strukturach kryształów niektórych krzemianów. Atomy krzemu i tlenu układają się w tych kryształach w pewne wzory, powielane w nieskończoność. Wzory te wykazują również pewną symetrię. Zauważyć można łatwo wyraźne podobieństwo uporządkowania z rys. 20a do uporządkowania w jednowymiarowym ornamencie z rys. 19. Symetryczne wzory tworzone przez atomy (jony, cząsteczki) w kryształach bywają różne, nieraz bardzo skomplikowane i bardzo oryginalne. Przykład takiego, wprawdzie o nieskomplikowanej symetrii, ale oryginalnego wzoru pokazano na rys. 21, przedstawiającym symetrię ułożenia atomów w różnych warstwach struktury kryształu minerału — apofyllitu.

Bardzo często symetrię wykazują cząsteczki związków chemicznych. Niekiedy jest to symetria stosunkowo prosta, jak np. w cząsteczce mocznika (rys. 22), a niekiedy złożona i skomplikowana, jak np. w cząsteczkach adamantanu i urotropiny (rys. 10a, b).

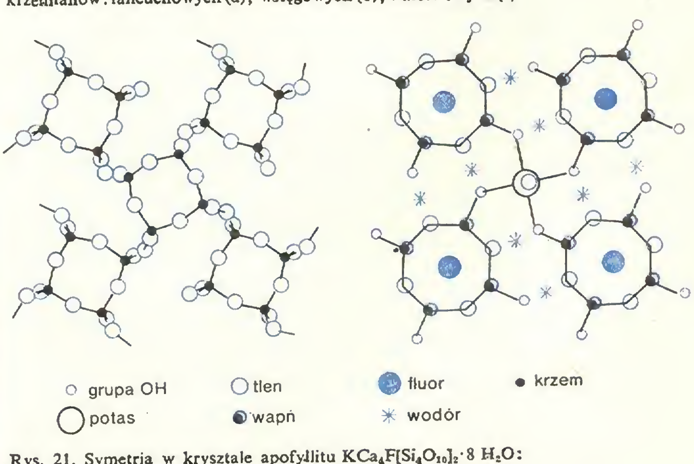
Symetria występuje nie tylko w różnych uporządkowaniach atomów (jonów, cząsteczek) w ciałach krystalicznych i w samych cząsteczkach, nie tylko w różnych ornamentach, w budowlach, dziełach sztuki, sprzętach codziennego użytku (człowiek — sam na zewnątrz ukształtowany symetrycznie — lubi wytwarzać przedmioty symetryczne, ponieważ symetria kojarzy mu się z pięknem i harmonią kształtów i barw), ale i w przyrodzie ożywionej: symetrycznie — w każdym razie na zewnątrz — zbudowane są



Rys. 19. Szlaki ornamentacyjne o różnej symetrii

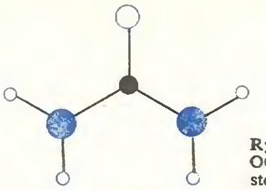


Rys. 20. Uporządkowanie atomów krzemu i tlenu w strukturach krzemianów: łańcuchowych (a), wstęgowych (b), warstwowych (c)



Rys. 21. Symetria w krysztale apofyllitu $KCa_4F[Si_4O_{10}]_2 \cdot 8\ H_2O$: a) warstwa utworzona przez grupy krzemotlenowe, b) warstwa utworzona przez jony Ca, K, F, O i OH

przeważnie zwierzęta i rośliny. W tych ostatnich przypadkach czasem jest to taka symetria, jaka występuje w kryształach, a czasem inna (il. 63, 65 i 66, tabl. 17).

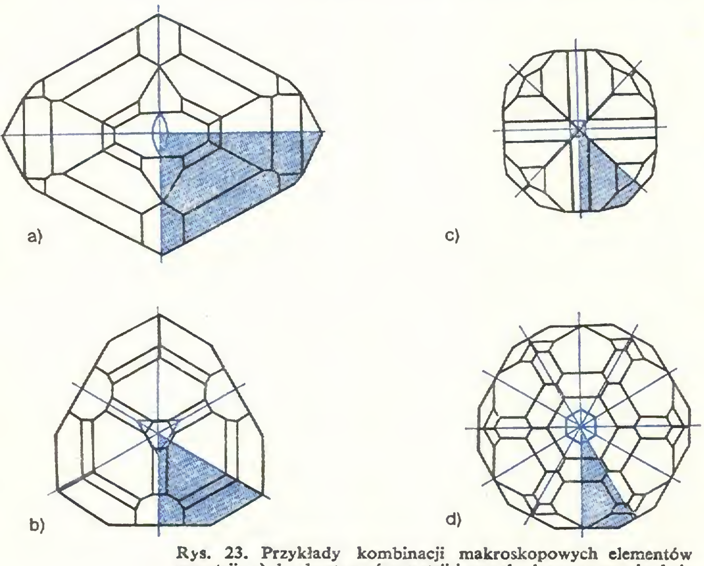


Rys. 22. Cząsteczka mocznika $OC(NH_2)_2$. Symetria własna cząsteczki: *mm*

Naukę o symetrii, jej rodzajach i sposobach określania na najwyższym chyba poziomie postawili krystalografowie, wykrywając i badając symetrię najpierw zewnętrznych postaci kryształów, a później i struktur kryształów. Symetria jest właściwością figury polegającą na tym, że przy określonych zmianach położenia tej figury nowe położenia pokrywają się z położeniem pierwotnym (tj. figura pokrywa się sama ze sobą). O figurze geometrycznej wtedy mówimy, że jest symetryczna, gdy składa się z jednakowych, prawidłowo rozmieszczonych części.

symetria kryształów

Jak w świetle tych sformułowań można zdefiniować symetrię kryształów? Jak już wiemy, kryształy podczas swobodnego wzrostu otaczają się płaskimi ścianami. Naturalne ściany kryształu są zawsze równoległe do określonych płaszczyzn sieciowych sieci krystalicznej. Ściany kryształów przecinają się wzdłuż linii prostych — krawędzi, którym w sieciach krystalicznych odpowiadają określone proste sieciowe. Trzy lub więcej ścian przecinających się w jednym punkcie tworzy naroże kryształu. Patrząc na rzuty ortogonalne kryształów przedstawione w tabeli Makroskopowe elementy symetrii i na rys. 23 symetrię zewnętrznych postaci kryształów można zdefiniować jako: prawidłowe powtarzanie się w przestrzeni jednakowych pod względem geometrycznym i fizycznym ścian, krawędzi i naroży. Należy tu jednak podkreślić, że symetria zewnętrznych postaci kryształów jest zjawiskiem wtórnym, wynikającym z symetrii atomowej budowy danego kryształu, tzn. z symetrii sieci przestrzennej (sieci krystalicznej).



Rys. 23. Przykłady kombinacji makroskopowych elementów symetrii: a) dwukrotna oś symetrii i przechodzące przez nią dwie płaszczyzny symetrii (kryształ: topaz), b) trójkrotna oś symetrii i trzy płaszczyzny symetrii (kryształ: turmalin), c) czterokrotna oś symetrii i cztery płaszczyzny symetrii (kryształ: rutyl), d) sześciokrotna oś symetrii i sześć płaszczyzn symetrii (kryształ: beryl); powtarzany elementami symetrii fragment kryształu jest zacieniowany; osie i płaszczyzny symetrii są prostopadłe do płaszczyzny rysunku

Podobnie jak symetrię kryształu, symetrię sieci przestrzennej (czy sieci krystalicznej) można określić jako prawidłowe powtarzanie się w przestrzeni węzłów (atomów, cząsteczek), prostych sieciowych, płaszczyzn sieciowych, a także komórek elementarnych.

elementy symetrii

Aby zrealizować dowolne przekształcenie symetryczne, posługujemy się pewnymi pomocniczymi punktami, liniami lub płaszczyznami zw. elementami symetrii. Elementy symetrii służące do opisu przekształceń symetrycznych, którym podlegają fragmenty zewnętrznych postaci kryształów, nazywają się makroskopowymi elementami symetrii. Makroskopowe elementy symetrii — do których należą osie symetrii zwykłe, inwersyjne i przemienne, środek symetrii i płaszczyzna symetrii — odznaczają się tym, że wykonanie za ich pomocą odpowiednich przekształceń symetrycznych doprowadza figurę z powrotem do położenia wyjściowego. W związku z tym mówimy o nich, że są elementami symetrii figur skończonych. Figurami skończonymi są na przykład krystalograficzne wielościany, tworzące zewnętrzne postacie kryształów.

makroskopowe elementy symetrii

W sieciach przestrzennych (krystalicznych), prócz makroskopowych elementów symetrii, występują strukturalne elementy symetrii, którymi są: oś translacji, śrubowe osie symetrii i płaszczyzny poślizgu. Wykonanie odpowiednich przekształceń symetrycznych za pomocą strukturalnych elementów symetrii nie doprowadza przekształcanej figury do położenia wyjściowego. Strukturalne elementy symetrii są elementami symetrii figur nieskończonych. Figurami nieskończonymi są np. ornamenty z rys. 19, a także sieci przestrzenne czy sieci krystaliczne.

strukturalne elementy symetrii

Rezultaty działania makroskopowych elementów symetrii można obserwować na tzw. kryształach idealnych nawet nie uzbrojonym okiem, jeśli tylko kryształ jest dostatecznie duży. Strukturalne elementy symetrii ujawniają się natomiast prawie wyłącznie w zjawiskach dyfrakcyjnych.

Makroskopowe elementy symetrii

Jeżeli w krysztale (w jego sieci krystalicznej) istnieje taki punkt, że na dowolnej prostej przeprowadzonej przez ten punkt, w jednakowej od niego odległości znajdują się jednakowe fragmenty kryształu (sieci krystalicznej, sieci przestrzennej), to ten punkt jest środkiem symetrii (centrum symetrii) kryształu. Schemat działania środka symetrii oraz rezultaty przekształceń względem środka symetrii przedstawione zostały w tabeli Makroskopowe elementy symetrii. W tabeli tej pokazano również symbole środka symetrii stosowane w krystalografii.

środek symetrii

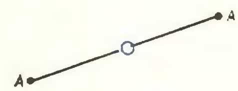
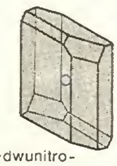
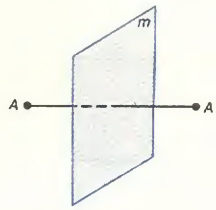
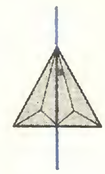
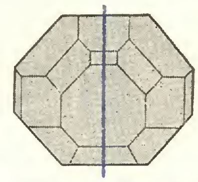
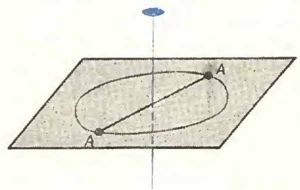
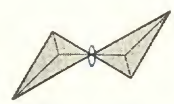
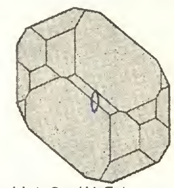
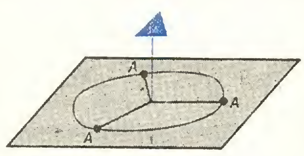
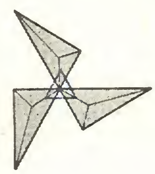
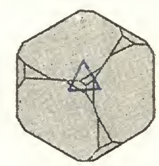
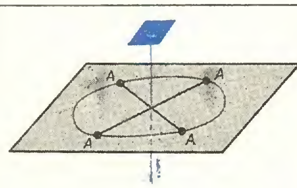
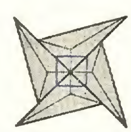
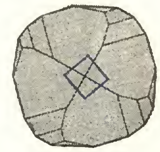
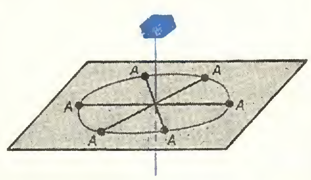
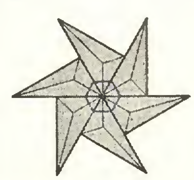
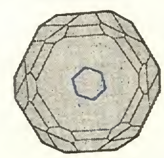
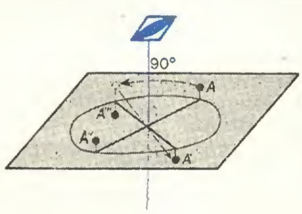
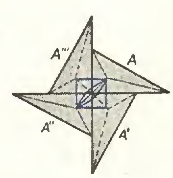
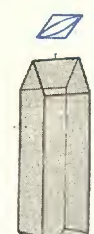
Gdy dwie części kryształu (sieci krystalicznej) pozostają do siebie w takim stosunku, jak przedmiot do swego obrazu w płaskim zwierciadle, to element symetrii łączący ze sobą te części nazywa się płaszczyzną symetrii.

płaszczyzna symetrii

Jeżeli w krysztale (w sieci krystalicznej) dany jego fragment powtarza się dwa lub więcej razy (mówimy też w tym wypadku, że fragment ten „nakłada się" sam na siebie, lub że przekształca się sam w sobie) po obrocie kryształu o stale ten sam kąt wokół pewnej prostej i gdy powtarzające się fragmenty znajdują się stale w jednakowej odległości od tej prostej, to prosta ta jest osią symetrii. W kryształach — ze względu na ich budowę sieciową — mogą występować tylko osie symetrii, wokół których ta sama część kryształu (sieci) powtarza się co kąt $\alpha = 180°$, 120°, 90°, 60° czyli odpowiednio: 2, 3, 4 i 6 razy. Tzw. krotność osi n równa się więc 2, 3, 4, 6 ($n = 360°/\alpha$). W związku z tym osie nazywamy dwu-, trój-, cztero- i sześciokrotnymi osiami symetrii. W tabeli przedstawione zostały schematy działania osi oraz działania osi na asymetryczny czworościan i na ściany kryształu.

osie symetrii

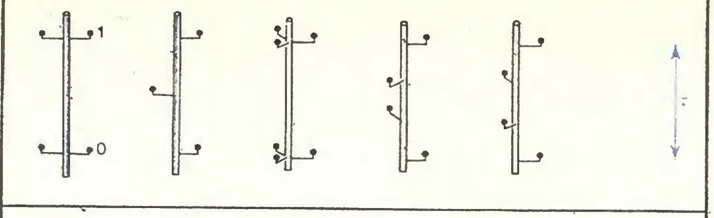
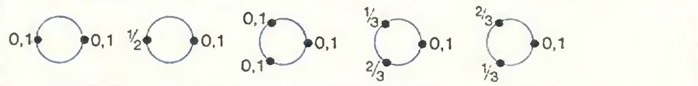
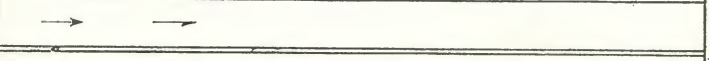
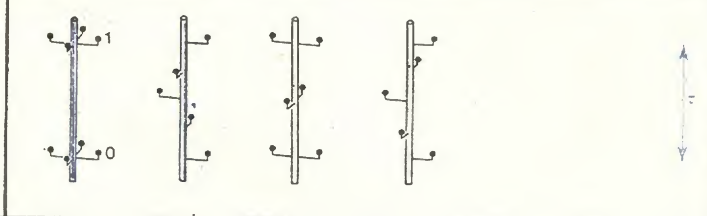
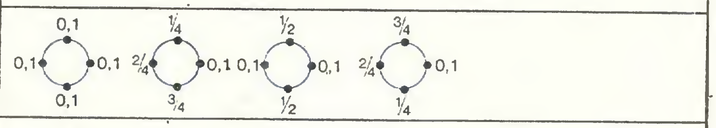
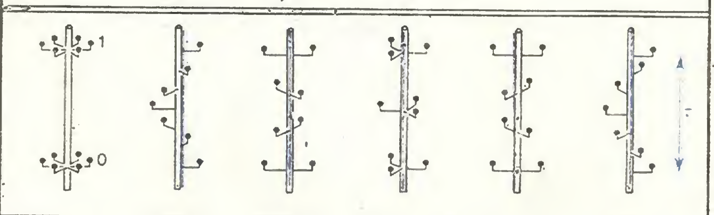
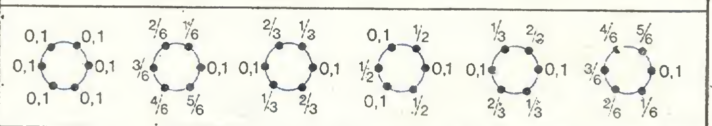
Makroskopowe elementy symetrii

element symetrii	symbole			schemat działania	bryła powstała w wyniku działania elementu symetrii na asymetryczny czworościan	przykład działania elementu symetrii na perspektywicznym lub ortogonalnym rzucie kryształu
	międzynarodowe a)		Kreutza b)			
	literowy (cyfrowy)	graficzny				
środek symetrii	$\bar{1}$	○	*C*			 1,4-dwunitro--2,5 dwubromobenzen
płaszczyzna symetrii	*m*	|	*P*			 mirabilit $Na_2SO_4 \cdot 10\,H_2O$
dwukrotna oś symetrii	2	⬮	L^2			 epidot $Ca_2(Al,Fe)$ $Al_2O(OH)[SiO_4][Si_2O_7]$
trójkrotna oś symetrii	3	▲	L^3			 fenakit Be_2SiO_4
czterokrotna oś symetrii	4	■	L^4			 szelit $CaWO_4$
sześciokrotna oś symetrii	6	⬢	L^6			 wanadynit $Pb_5[VO_4]_3Cl$
czterokrotna oś inwersyjna	$\bar{4}$	◪	A^4			 mocznik $OC(NH_2)_2$

a) według *International Tables for X-Ray Crystallography*, Birmingham 1952

b) Stefan Kreutz (1883-1941), polski mineralog i krystalograf.

Strukturalne elementy symetrii – osie śrubowe

schemat działania	w przestrzeni trójwymiarowej						
	w rzucie na płaszczyznę prostopadłą do osi						
symbole międzynarodowe	cyfrowy	2	2_1	3	3_1	3_2	
	graficzny osi prostop. do płaszcz. rys.						
	graficzny osi rownol. do płaszcz. rys.						
schemat działania	w przestrzeni trójwymiarowej						
	w rzucie na płaszczyznę prostopadłą do osi						
symbole międzynarodowe	cyfrowy	4	4_1	4_2	4_3		
	graficzny osi prostop. do płaszcz. rys.						
schemat działania	w przestrzeni trójwymiarowej						
	w rzucie na płaszczyznę prostopadłą do osi						
symbole międzynarodowe	cyfrowy	6	6_1	6_2	6_3	6_4	6_5
	graficzny osi prostop. do płaszcz. rys.						

czterokrotna oś inwersyjna

Oprócz tych zwykłych osi symetrii w kryształach występuje jeszcze czterokrotna oś inwersyjna. Jest ona złożonym elementem symetrii, co polega na tym, że dana część kryształu (sieci) powtarza się dopiero po wykonaniu dwóch przekształceń, mianowicie przekształcenia względem zwykłej czterokrotnej osi symetrii i przekształcenia względem środka symetrii. Takie złożenie dwóch przekształceń symetrycznych nazywa się ich iloczynem. Sposób dokonywania przekształceń względem osi inwersyjnej, schemat działania czterokrotnej osi inwersyjnej, działanie takiej osi na czworościan oraz działanie osi w krysztale przedstawione zostały w tabeli.

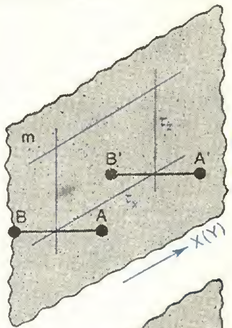
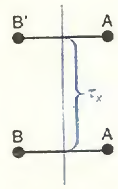
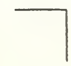
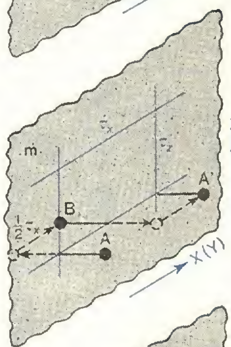
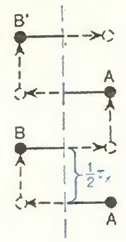
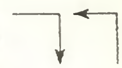
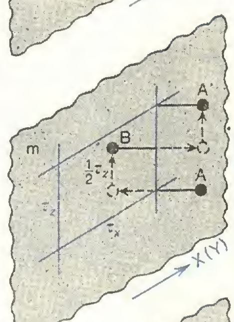
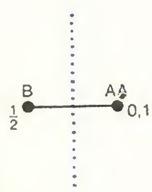
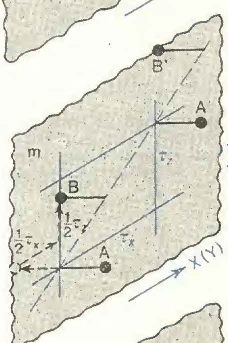
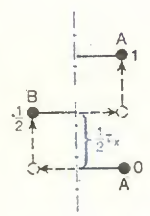
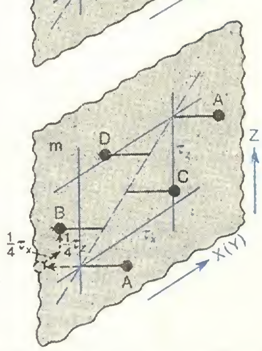
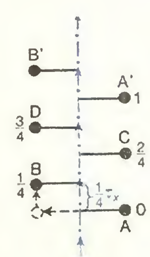
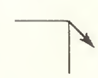
Inne osie, inwersyjne poza osią $\bar{4}$ — a także tzw. osie przemienne (będące wynikiem złożenia dwóch przekształceń: względem osi symetrii n-krotnej i prostopadłej do niej płaszczyzny symetrii) nie dają żadnych takich przekształceń, których nie można by uzyskać za pomocą wymienionych już prostych przekształceń.

Strukturalne elementy symetrii

oś translacji

Podstawowym elementem symetrii sieci przestrzennych jest oś translacji. Przekształcenie symetryczne względem osi translacji polega na równoległym przesuwaniu punktu (lub zbioru punktów) o stale ten sam, ściśle określony odcinek τ (rys. 24). Odcinek ten nazywa się odcinkiem translacji lub periodem identyczności. Sieć przestrzenną, ten nieskończony zbiór komórek elementarnych, można zbudować pod-

Strukturalne elementy symetrii – płaszczyzny poślizgu

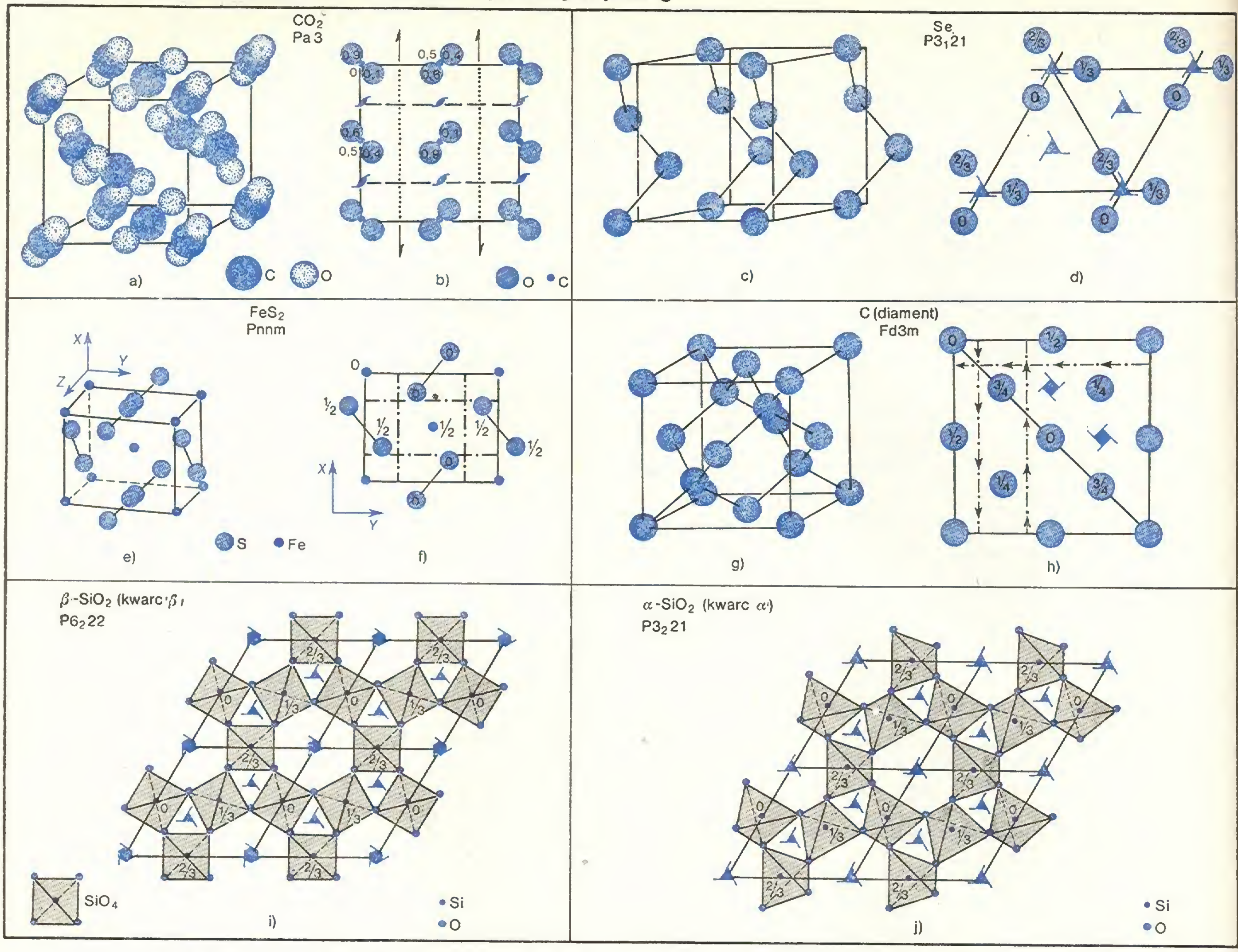
schemat działania		symbole międzynarodowe		
			graficzny	
w przestrzeni trójwymiarowej	w rzucie na płaszcz. prostop. do płaszczyzny poślizgu	literowy	płaszcz. prostopadłej do płaszczyzny rys.	płaszcz. równol. do płaszcz. rys.
		*m*	————	
		*a* (*b*)	— — — —	
		*c*	....	
		*n*	—·—·—·—	
		*d*	→·—·→·—·→	

dając jedną komórkę działaniu translacji w trzech kierunkach określonych osiami krystalograficznymi *X*, *Y*, *Z*. Złożonymi strukturalnymi elementami symetrii są śrubowe osie symetrii i płaszczyzny poślizgu.

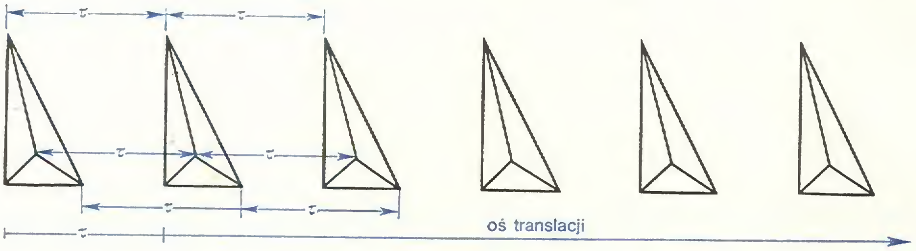
osie śrubowe

Oś śrubowa jest wynikiem wspólnego działania *n*-krotnej osi symetrii i translacji. Mówiąc inaczej, oś śrubowa jest iloczynem osi symetrii i translacji. Działanie takiej osi polega na tym, że po obrocie dowolnego fragmentu struktury kryształu wokół osi sy-

Przykłady działania niektórych osi śrubowych i płaszczyzn poślizg



a), c), e), g) ogólny wygląd struktury; b), d), f), h) rzut struktury na płaszczyznę *XY*; i), j) rzut na płaszczyznę *XY* struktury przedstawionej w postaci czworościanów SiO_4. Na rzutach zaznaczono położenia niektórych elementów symetrii: b) osi 2_1, płaszczyzn *b* i *c*; d) osi 3_1; f) płaszczyzn *n*; h) osi 4_1 i 4_3 oraz płaszczyzn *d*; i) osi 6_2 i 3_2; j) osi 3_2. Pod nazwą związku są podane symbole grupy przestrzennej struktury. Liczby oznaczają położenia atomów (czworościanów) nad płaszczyzną rysunku wzdłuż osi prostopadłej do niej wyrażone w ułamkach odpowiedniego periodu identyczności



Rys. 24. Translacja — przesunięcie równoległe wszystkich punktów o jednakowy odcinek

metrii o właściwy tej osi kąt α, fragment ten zostaje przesunięty w kierunku równoległym do osi symetrii o pewną, ściśle określoną część translacji (stale taką samą), równą $(p/n)\tau$, gdzie p i n są liczbami całkowitymi i $p < n$, np. w przypadku czterokrotnej osi śrubowej wykonuje się kolejne obroty o 90°, którym towarzyszą przesunięcia translacyjne bądź stale o $^1/_4\tau$, bądź stale o $^1/_2\tau$, albo też stale o $^3/_4\tau$ (tab. Strukturalne elementy symetrii — osie śrubowe). Tak więc istnieją trzy czterokrotne osie śrubowe, o trzech różnych składowych translacyjnych.

Osie śrubowe oznacza się symbolem n_p. Nazwa tych osi (śrubowe) pochodzi stąd, że pod działaniem takiej osi wszystkie punkty (fragmenty) sieci krystalicznej przesuwają się po liniach śrubowych (lewo- lub prawoskrętnych). W strukturach kryształów występuje 11 osi śrubowych: osie śrubowe prawoskrętne — 3_1, 4_1, 6_1, 6_2; osie śrubowe lewoskrętne — 3_2, 4_3, 6_4, 6_5 i osie śrubowe neutralne (tj. bez wyróżnionego kierunku skrętu) — 2_1, 4_2, 6_3. Schematy działania osi śrubowych pokazano w tabeli znajdującej się na str. 444.

płaszczyzny poślizgu

W wyniku iloczynu płaszczyzny symetrii oraz translacji powstaje płaszczyzna poślizgu. Działanie jej po-

lega na tym, że po wykonaniu odbicia zwierciadlanego każdego fragmentu struktury w płaszczyźnie symetrii, fragment ten zostaje przesunięty w kierunku równoległym do tej płaszczyzny o połowę odpowiedniego odcinka translacji. Płaszczyznę poślizgu oznacza się symbolem *a*, *b* lub *c*, w zależności od kierunku przesunięcia (wzdłuż osi krystalograficznej *X*, *Y* czy *Z*).

W strukturach kryształów występują również bardziej skomplikowane płaszczyzny poślizgu, działające w taki sposób, że odbiciu zwierciadlanemu towarzyszą dwa przesunięcia o $^1/_2$ lub $^1/_4$ odcinka translacji w kierunkach dwóch osi krystalograficznych. Płaszczyzna poślizgu typu *n* może mieć np. takie dwie składowe translacyjne: $^1/_2\tau_x$ i $^1/_2\tau_z$, a płaszczyzna typu *d* — $^1/_4\tau_x$ i $^1/_4\tau_z$ (τ_x i τ_z — odcinki translacji wzdłuż osi *X* i *Z*; zob. tab. Strukturalne elementy symetrii — płaszczyzny poślizgu i Przykłady działania niektórych osi śrubowych i płaszczyzn poślizgu).

Kombinacje elementów symetrii

Klasy i układy krystalograficzne

Kryształy, ich struktury czy sieci krystaliczne mogą mieć nie tylko pojedyncze elementy symetrii, jak np. tylko środek symetrii, tylko jedną *n*-krotną oś symetrii zwykłą, inwersyjną lub śrubową, tylko jedną płaszczyznę symetrii lub płaszczyznę poślizgu, ale także różne kombinacje (zespoły) elementów symetrii makroskopowych albo strukturalnych, lub i makroskopowych i strukturalnych. Działanie różnych kombinacji *n*-krotnych osi symetrii i płaszczyzn symetrii na ściany kryształów, dające wyobrażenie o kształtach kryształów przedstawiono na rys. 23.

klasy symetrii (grupy punktowe)

Różne możliwe w kryształach kombinacje makroskopowych elementów symetrii przecinających się w jednym punkcie tworzą klasy symetrii kryształów

Układy krystalograficzne i klasy symetrii

Układ krystalograficzny	Elementy symetrii charakteryzujące układ	Stałe sieciowe komórki elementarnej	Symbol klasy wg tablic międzynar.[a])	Symbol klasy wg Schoenfliesa[b])	Symbol klasy wg Kreutza[c])	Wszystkie elementy symetrii klasy	Klasa (nazwa)
Trójskośny	—	$a \neq b \neq c$, $\alpha \neq \beta \neq \gamma$	1	C_1	L^1	—	jednościanu
			1	C_1	C	C	dwuścianu
Jednoskośny	L^2 lub P	$a \neq b \neq c$, $\alpha = \gamma = 90°$, $\beta \neq 90°$	2	C_2	L^2_y	L^2	sfenoidu jednoskośnego
			m	C_s	P_y	P	daszka jednoskośnego
			2/*m*	C_{2h}	L^2_y, C	L^2PC	słupa jednoskośnego
Rombowy	$L^2 \perp L^2$ lub $L^2 \parallel P$	$a \neq b \neq c$, $\alpha = \beta = \gamma = 90°$	*mm*2	C_{2v}	L^2_z, P_x	L^22P	piramidy rombowej
			222	D_2	L^2_z, L^2_x	$3L^2$	czworościanu rombowego
			*mmm*2	D_{2h}	L^2_z, L^2_x, C	$3L^23PC$	podwójnej piramidy rombowej
Tetragonalny	L^4 lub A^4	$a = b \neq c$, $\alpha = \beta = \gamma = 90°$	$\bar{4}$	S_4	A^4_z	A^4	czworościanu tetragonalnego
			4	C_4	L^4_z	L^4	piramidy tetragonalnej
			4/*m*	C_{4h}	L^4_z, C	L^4PC	podwójnej piramidy tetragonalnej
			$\bar{4}2m$	D_{2d}	A^4_z, L^2_x	A^42L^22P	skalenoedru tetragonalnego
			4*mm*	C_{4v}	L^4_z, P_x	L^44P	piramidy dytetragonalnej
			422	D_4	L^4_z, L^2_x	L^44L^2	trapezoedru tetragonalnego
			4/*mmm*	D_{4h}	L^4_z, L^2_x, C	L^44L^25PC	podwójnej piramidy dytetragonalnej
Heksagonalny	L^6 lub L^3	$a = b \neq c$, $\alpha = \beta = 90°$, $\gamma = 120°$	3	C_3	L^3_z	L^3	piramidy trygonalnej
			$\bar{3}$	C_{3i}	L^3_z, C	L^3C	romboedru
			3*m*	C_{3v}	L^3_z, P_x	L^33P	piramidy dytrygonalnej
			32	D_3	L^3_z, L^2_x	L^33L^2	trapezoedru trygonalnego
			$\bar{3}m$	D_{3d}	L^3_z, L^2_x, C	L^33L^23PC	skalenoedru trygonalnego
			$\bar{6}$	C_{3h}	L^3_z, P_z	L^3P	podwójne piramidy trygonalnej
			$\bar{6}m2$	D_{3h}	L^3_z, L^2_x, P_z	L^33L^24P	podwójnej piramidy dytrygonalnej
			6	C_6	L^6_z	L^6	piramidy heksagonalnej
			6/*m*	C_{6h}	L^6_z, C	L^6PC	podwójnej piramidy heksagonalnej
			6*mm*	C_{6v}	L^6_z, P_x	L^66P	piramidy dyheksagonalnej
			622	D_6	L^6_z, L^2_x	L^66L^2	trapezoedru heksagonalnego
			6/*mmm*	D_{6h}	L^6_z, L^2_x, C	L^66L^27PC	podwójnej piramidy dyheksagonalnej
Regularny	$2L^3$	$a = b = c$, $\alpha = \beta = \gamma = 90°$	23	T	L^3_1, L^3_2	$4L^33L^2$	dwunastościanu tetraedryczno-pentagonalnego
			*m*3	T_h	L^3_1, L^3_2, C	$4L^33L^23PC$	dwunastościanu pentagonalnego podwójnego
			$\bar{4}3m$	T_d	A^4_z, A^4_x	$3A^44L^36P$	czworościanu poszóstnego
			432	O	L^4_z, L^4_x	$3L^44L^36L^2$	dwudziestoczterościanu pentagonalnego
			*m*3*m*	O_h	L^4_z, L^4_x, C	$3L^44L^36L^29PC$	czterdziestoośmiościanu

[a]) *International Tables for X-Ray Crystallography*, Birmingham 1952. [b]) A. M. Schoenflies — niemiecki matematyk (1853–1928). [c]) S. Kreutz — polski mineralog i krystalograf (1883–1941).

Uwaga: L^2_z oznacza dwukrotną oś symetrii równoległą do krystalograficznej osi *Z*; P_y oznacza płaszczyznę symetrii prostopadłą do krystalograficznej osi *Y*; 4/*m* oznacza płaszczyznę symetrii prostopadłą do czterokrotnej osi symetrii.

(klasy krystalograficzne lub grupy punktowe). Kombinacji takich istnieje tylko 32 i grupuje się je w 6 układów krystalograficznych. Kryształy dowolnego związku chemicznego można zawsze zaliczyć na podstawie ich symetrii makroskopowej do jednej z 32 klas symetrii. W tab. Układy krystalograficzne i klasy symetrii pokazano podział klas między układy krystalograficzne oraz podano elementy symetrii i stałe sieciowe charakteryzujące poszczególne układy. Nazwa klasy pochodzi od postaci prostej ogólnej występującej w danej klasie.

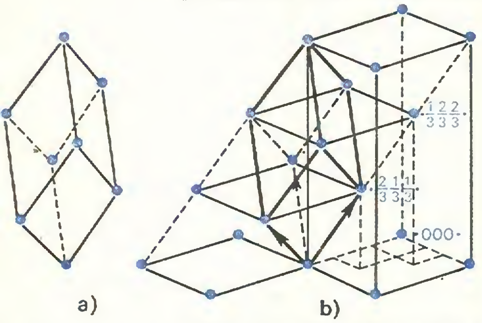
Przy ujęciu zagadnienia symetrii szerszym niż w krystalografii, tj. przy uwzględnieniu istnienia osi 5-, 7-, 8-, ...-krotnych, liczba klas symetrii staje się nieskończona. Łatwo jest wtedy utworzyć z klas symetrii szeregi, z których każdy kończy się klasą zawierającą oś symetrii o krotności nieskończonej (∞ lub L_∞), np.: 1, 2, 3, 4, 5, 6, ..., ∞; 2, 222, 32, 422, 52, 622, ..., ∞2; m, mm2, 3m, 4mm, 5m, 6mm, ..., ∞mm itp. Klasy symetrii (grupy punktowe), w których występują osie o krotności nieskończonej, nazywają się grupami granicznymi. Jak pokazał P. Curie, grup takich jest siedem; mają one symbole: ∞, ∞2, ∞/m, ∞mm, ∞/mm, $\infty\infty$, $\infty\infty m$. Grupy graniczne znajdują zastosowania przy opisywaniu fizycznych własności kryształów.

grupy graniczne

układy krystalograficzne

Układem krystalograficznym nazywa się zespół klas symetrii, których elementy symetrii powodują jednakowe ograniczenia stałych sieciowych (np. w układzie tetragonalnym obecność osi L^4 powoduje równość krawędzi a i b w komórce elementarnej oraz wymaga, by kąty α, β, γ były kątami prostymi). Wszystkie klasy danego zespołu można opisać przy użyciu takiego samego układu osi odniesienia scharakteryzowanego przez stałe sieciowe.

Jak pokazano w tabeli (Układy krystalograficzne i klasy symetrii) istnieje sześć układów krystalograficznych: trójskośny, jednoskośny, rombowy, tetragonalny, heksagonalny i regularny. Niekiedy (w fizyce ciała stałego i w mineralogii) heksagonalne klasy symetrii z trójkrotną osią symetrii wydziela się jako odrębny układ — trygonalny, w którym komórka elementarna ma kształt romboedru ($a = b = c$, $\alpha = \beta = \gamma \neq 90°$). Z punktu widzenia krystalografii strukturalnej nie jest to słuszne, gdyż sieć przestrzenna zarówno w obecności osi 6, jak i osi 3 jest taka sama, nie ma więc odrębnej sieci „trygonalnej", a poza tym prymitywną komórkę elementarną o kształcie rombo-



Rys. 25. Prymitywna komórka trygonalna (a) jako centrowana komórka heksagonalna typu R (c) i związek między tymi komórkami (b)

edru można zawsze przedstawić (rys. 25) jako podwójnie centrowaną komórkę heksagonalną typu R. Tak więc w krystalografii strukturalnej rozróżnia się 6 układów krystalograficznych, łącząc co najwyżej klasy z osią 3 w tzw. podukład trygonalny.

Zewnętrzne postacie kryształów

Jak już niejednokrotnie wspomniano, kryształy podczas swego wzrostu dążą do otaczania się płaskimi ścianami. W rezultacie powstają zwykle wielościanowe postacie kryształów, nieraz bardzo oryginalne, bardzo skomplikowane, a najczęściej bardzo ładne. Ściany na krysztale nie tworzą się w sposób dowolny, są one ściśle związane z sieciową budową wewnętrzną kryształów i ich symetrią.

postacie krystalograficzne

Zespoły ścian na krysztale nazywa się postaciami krystalograficznymi, wśród których rozróżnia się postacie proste i postacie złożone.

Postać prostą stanowi zespół ścian symetrycznie równoznacznych, tzn. związanych ze sobą elementami symetrii. Taki zespół ścian otrzymuje się w rezultacie poddania jednej ściany działaniu elementów symetrii danej klasy symetrii. Postaci prostych jest 47 (rys. 26). Postać prostą oznacza się symbolem, którym są wskaźniki jednej ze ścian tej postaci ujęte w nawias klamrowy: $\{hkl\}$; np. sześcian oznacza się symbolem {100}, a ośmiościan: {111}.

postacie proste ogólne i szczególne

Postacią prostą ogólną nazywa się postać, której ściany nachylone są skośnie względem elementów symetrii lub osi krystalograficznych. Ściany postaci ogólnej mają zawsze symbol typu (hkl), np. (321). Postacie proste szczególne mają ściany prostopadłe, równoległe, albo nachylone symetrycznie względem elementów symetrii lub osi krystalograficznych; symbolami takich ścian są np. (100), (110), (111), (221). W każdej klasie symetrii występuje jedna postać prosta ogólna $\{hkl\}$ i kilka postaci prostych szczególnych, np. {100}, {110}, $\{hhl\}$ itd. Ściany postaci prostych zamkniętych (np. czworościanu, sześcianu, podwójnych piramid) całkowicie ograniczają część przestrzeni, natomiast ściany postaci prostych otwartych (np. dwuścianu, słupów, piramid) nie zamykają przestrzeni.

postacie złożone

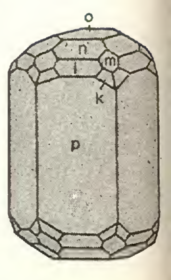
Kryształ, na którym występują ściany kilku postaci prostych, ma postać złożoną (rys. 27). Bardzo często spotykane są kryształy, których postać zewnętrzna jest kombinacją kilku czy kilkunastu nawet postaci prostych ogólnych i szczególnych.

W klasycznej nauce o postaciach prostych występują postacie o jednakowych nazwach i jednakowych kształtach (zarysach ścian), lecz różniące się symetrią. Wśród postaci prostych istnieje np. tylko jeden jednościan występujący zarówno na kryształach trójskośnych, jak i na kryształach jednoskośnych, rombowych, tetragonalnych czy heksagonalnych. A. W. Szubnikow zauważył, że chociaż w geometrii występuje tylko jeden sześcian, to w krystalografii jest ich pięć — o różnej symetrii, występujących w pięciu klasach symetrii układu regularnego. Podobnie i jednościan ma inną symetrię w układzie np. jednoskośnym, a inną w układzie rombowym, tetragonalnym czy heksagonalnym. Jeżeli dla poszczególnych postaci prostych, występujących w różnych klasach symetrii uwzględni się symetrię tych klas, to okazuje się, że postaci prostych jest nie 47, lecz 146, a nawet 193, gdy weźmie się jeszcze pod uwagę odmiany enancjomorficzne istniejące dla niektórych postaci (G. B. Bokij). Uwzględnienie przy charakteryzowaniu postaci prostych elementów symetrii grup przestrzennych doprowadziło do wyprowadzenia 1403 strukturalnych odmian postaci prostych (I. I. Szafranowskij). Dwie odmiany strukturalne jednej postaci prostej są różne wtedy, gdy ich ściany różnią się elementami symetrii swojej grupy przestrzennej. W ten sposób sześcian ma np. 36 odmian strukturalnych.

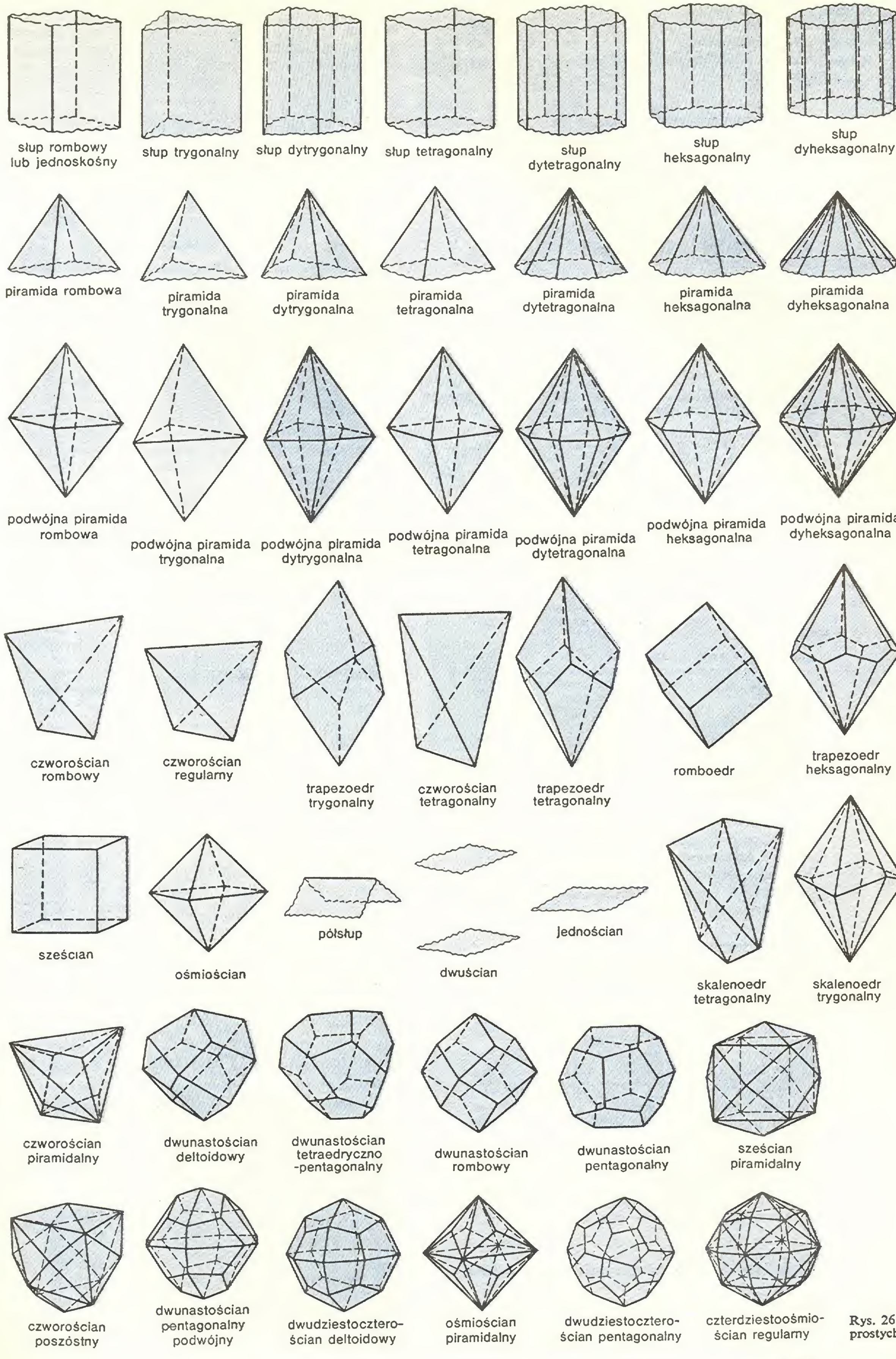
odmiany strukturalne

Pełniejszy opis kryształów rzeczywistych uwzględniający różne nieprawidłowości ich rozwoju wymagał rozszerzenia nauki o postaciach prostych w jeszcze inny sposób. Wprowadzono w tym celu pojęcie postaci prostej krawędziowej i postaci prostej wierzchołkowej. W tym ujęciu 47 krystalograficznych postaci prostych nazywa się postaciami ścianowymi.

Postać prostą wierzchołkową stanowi zespół naroży kryształu otrzymywanych jedno z drugiego za



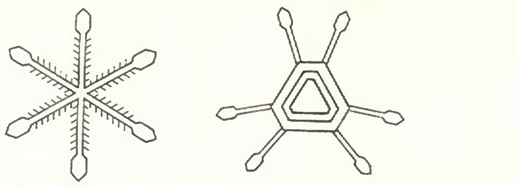
Rys. 27. Przykład kryształu o postaci złożonej: beryl — klasa symetrii $6/mmm$; postacie proste: k podwójna piramida dyheksagonalna, lmn podwójne piramidy heksagonalne, o dwuścian, p słup heksagonalny



Rys. 26. 47 postaci prostych

postacie proste wierzchołkowe

pomocą elementów symetrii. Postacie wierzchołkowe mogą być płaskie i przestrzenne. Postaci wierzchołkowych płaskich jest 10, a przestrzennych — 47. Przykładem płaskich postaci wierzchołkowych mogą być zgrubienia na końcach śnieżynek (rys. 28).



Rys. 28. Śnieżynki — przykłady postaci wierzchołkowych

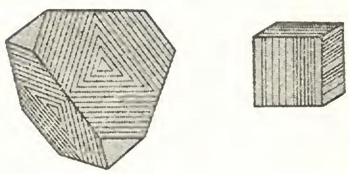
postacie proste krawędziowe

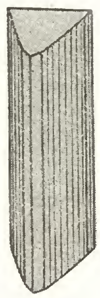
Postacią prostą krawędziową nazywa się zespół jednakowych krawędzi kryształu wyprowadzonych z jednej krawędzi za pomocą elementów symetrii (W. I. Michiejew, I. I. Szafranowskij). Postaci krawędziowych płaskich jest 27, a przestrzennych 303. Przykładem postaci krawędziowych są śnieżynki, kryształy cerusytu oraz wszystkie możliwe złożone zbruździenia ścian kryształów (rys. 29, 30).



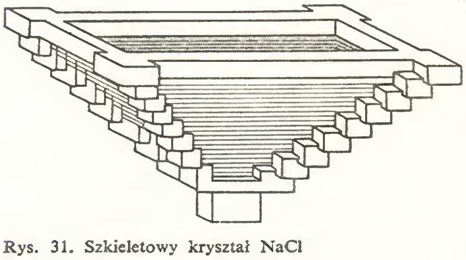
Rys. 29. Śnieżynki — przykłady postaci krawędziowych

Rys. 30. Zbruździenia ścian kryształów — przykłady postaci krawędziowych, a) sfaleryt, b) piryt, c) arsenopiryt





Między postaciami wierzchołkowymi, krawędziowymi i ścianowymi istnieją wzajemne powiązania. Widoczne jest to np. na szkieletowym krysztale soli kuchennej o kształcie piramidalnego lejka (tego rodzaju kryształy pływają na powierzchniach słonych



Rys. 31. Szkieletowy kryształ NaCl

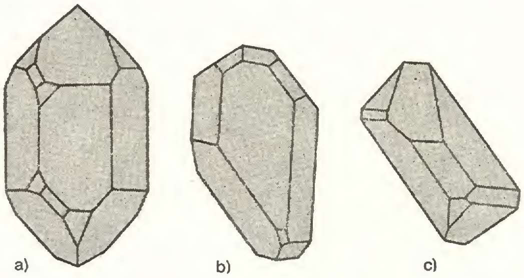
jezior). Maleńkie wierzchołkowe sześciany łączą się ze sobą podczas wzrostu kryształu i tworzą krawędziowe słupki, które nakładając się na siebie przechodzą w ścianki wklęsłego lejka — piramidki (rys. 31).

Kryształy idealne i kryształy rzeczywiste

kryształ idealny

Na krysztale idealnym wszystkie ściany należące do jednej postaci prostej mają ten sam kształt i tę samą wielkość oraz są jednakowo oddalone od środka kryształu. Sieć przestrzenna idealnego kryształu jest tworem jednorodnym, nie wykazującym zakłóceń w żadnym kierunku. Zewnętrzna i wewnętrzna budowa kryształów, zarówno naturalnych jak i otrzymywanych sztucznie najczęściej wykazuje wiele niedoskonałości i odbiega znacznie od idealnych modeli przedstawianych przez teorie symetrii i sieciowej budowy kryształów. Takie niedoskonałe kryształy nazywa się kryształami rzeczywistymi. O tym, jak może wyglądać zewnętrzna postać rzeczywistych kryształów kwarcu α, w porównaniu z kryształem idealnym, daje pewne wyobrażenie rys. 32.

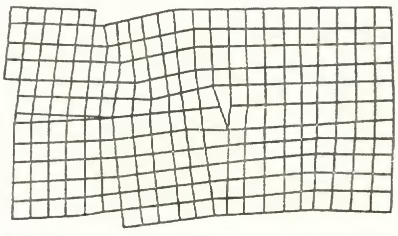
kryształ rzeczywisty



Rys. 32. Kwarc α: a) kryształ idealny, b) i c) kryształy rzeczywiste

Stwierdzono, że kryształy rzeczywiste zbudowane są najczęściej z bloków wielkości rzędu 10^{-5} cm (lub mniej) o prawidłowej budowie sieciowej, jednak nieznacznie skręconych względem siebie (rys. 33). Kryształ mający tego rodzaju budowę wewnętrzną nazywa się kryształem mozaikowym.

kryształ mozaikowy

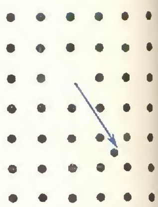


Rys. 33. Schemat mozaikowej budowy kryształu

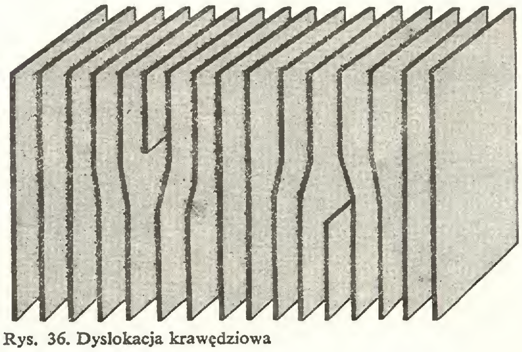
defekty punktowe

W budowie sieci krystalicznej występują lokalne zakłócenia prawidłowości jej budowy, zwane defektami punktowymi. Zasadniczymi typami tych defektów są defekty Frenkla, polegające na przesunięciu atomów (jonów) z węzłów sieci do przestrzeni międzywęzłowych (rys. 34), oraz defekty Schottky'ego (wakansje), polegające na nieobsadzeniu przez atomy (jony) części węzłów sieci (rys. 35).

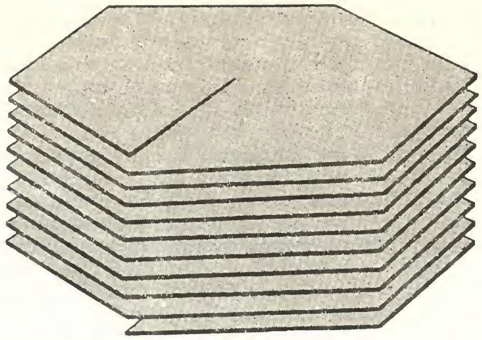
Innym rodzajem niedokładności w budowie sieci przestrzennej kryształów są dyslokacje, będące defektami liniowymi. Dyslokacje polegają na przesunięciu części płaszczyzn sieciowych w stosunku do innych, przy zachowaniu spójności sieci. Istnieją dyslokacje krawędziowe, śrubowe i mieszane. Jeżeli jedna z pła-



Rys. 34. Defekty Frenkla



Rys. 36. Dyslokacja krawędziowa



Rys. 37. Dyslokacja śrubowa

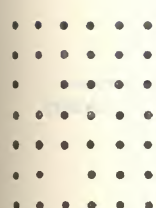
dyslokacje

szczyzn sieciowych urywa się we wnętrzu sieci przestrzennej, to wewnętrzny brzeg tej płaszczyzny tworzy dyslokację krawędziową (rys. 36). Dyslokacja śrubowa istnieje w krysztale wówczas, gdy płaszczyzny sieciowe jedynie w przybliżeniu są do siebie równoległe i połączone są ze sobą w taki sposób, że sieć przestrzenna kryształu stanowi jakby jedną płaszczyznę sieciową, tworzącą powierzchnię śrubową (rys. 37). W dyslokacjach mieszanych można rozróżnić dyslokacje składowe: krawędziową i śrubową.

Wiele właściwości fizycznych ciał krystalicznych zależy nie tylko od ich składu chemicznego i ich struktury, ale także od istniejących w nich rozmaitych defektów. Defekty mają wpływ m.in. na własności optyczne, półprzewodnikowe, magnetyczne, elektryczne i cieplne ciał krystalicznych oraz na ich wytrzymałość mechaniczną, a także na wzrost kryształów.

Grupy przestrzenne

Różne możliwe w strukturach ciał krystalicznych kombinacje makroskopowych i strukturalnych elementów symetrii tworzą grupy przestrzenne. Tak jak grupy punktowe (klasy symetrii) charakteryzują symetrię zewnętrznych postaci kryształów, tak grupy przestrzenne charakteryzują symetrię struktur kryształów. Istnieje 230 grup przestrzennych. Każdy układ krystalograficzny obejmuje pewną liczbę grup przestrzennych, podzielonych z kolei między poszczególne klasy symetrii.



Rys. 35. Defekty Schottky'ego

Grupy przestrzenne opisuje się zwykle za pomocą tzw. symboli międzynarodowych, składających się z dużej litery oznaczającej typ sieci Bravais'go, liczb oznaczających osie symetrii zwykłe, inwersyjne lub śrubowe i małych liter oznaczających płaszczyzny symetrii lub płaszczyzny poślizgu. Międzynarodowy symbol grupy przestrzennej jest tak skonstruowany, że na jego podstawie można wyznaczyć wszystkie elementy symetrii w danej grupie przestrzennej i określić ich rozmieszczenie w komórce elementarnej.

symbole grup przestrzennych

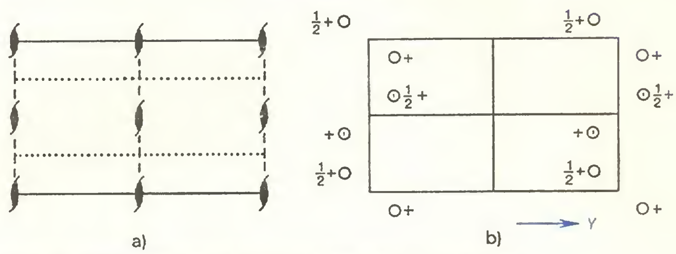
Każdej grupie przestrzennej odpowiada tylko jedna klasa krystalograficzna, którą można łatwo wyznaczyć dla danej grupy przestrzennej, zamieniając w tej ostatniej wszystkie osie śrubowe i płaszczyzny poślizgu na zwykłe osie i płaszczyzny oraz przesuwając równolegle wszystkie elementy symetrii tak, by przecinały się w jednym punkcie. W podobny sposób (zamieniając strukturalne elementy symetrii na makroskopowe i odrzucając symbol sieci Bravais'go) z symbolu grupy przestrzennej można uzyskać symbol klasy krystalograficznej, do której dana grupa przestrzenna należy, np.: $Pca2_1 \rightarrow mm2$ (to oznacza, że grupa $Pca2_1$ należy do klasy *mm*2 w układzie rombowym).

Każda grupa przestrzenna ma właściwe sobie tzw. zespoły pozycji symetrycznie równoznacznych, czyli przestrzenne układy punktów związanych ze sobą elementami symetrii. Każdy taki zespół punktów powstaje z jednego punktu wyjściowego (a więc mamy tu „rozmnożenie" punktu) przez poddanie go działaniu wszystkich elementów symetrii danej grupy przestrzennej. Liczba punktów znajdujących się w jednej komórce elementarnej, czyli tzw. liczebność lub krotność pozycji, może być różna w różnych zespołach pozycji danej grupy przestrzennej i jest zależna od położenia punktu wyjściowego względem elementów symetrii. W każdej grupie przestrzennej istnieje jeden zespół pozycji symetrycznie równoznacznych w położeniu ogólnym i kilka zespołów pozycji w położeniach szczególnych. Zwykle w kryształach tylko niektóre zespoły pozycji danej grupy przestrzennej są obsadzane przez atomy (jony, cząsteczki).

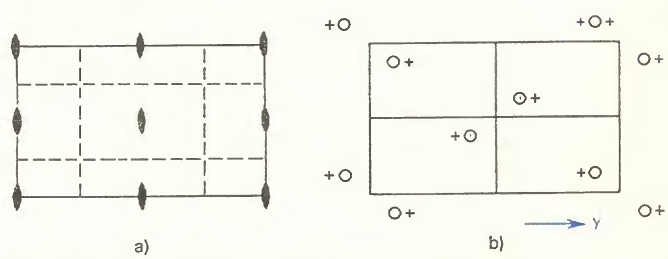
liczebność (krotność) pozycji

Symetrię grup przestrzennych i zespoły pozycji symetrycznie równoznacznych przedstawia się graficznie w rzucie na jedną ze ścian komórki elementarnej, zwykle na płaszczyznę *XY*. Takie rzuty dla trzech grup przestrzennych należących do układu rombowego pokazano na rys. 38, 39, 40. Na rysunkach tych wyraźnie widać różnice, w przestrzennych ugrupowaniach punktów i w liczebności pozycji ogólnej, zależne od zmiany elementów symetrii i typu sieci Bravais'go w obrębie jednej klasy symetrii.

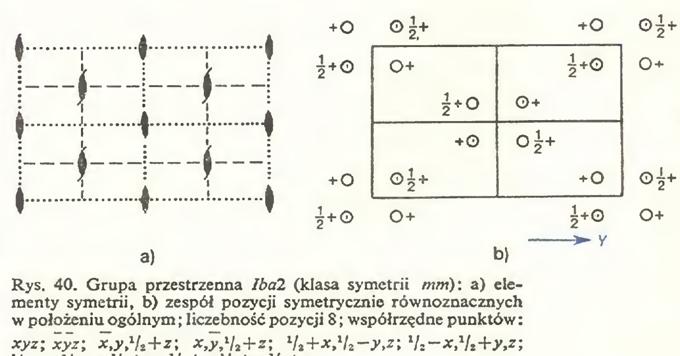
Badanie struktury każdego ciała krystalicznego rozpoczyna się od wyznaczenia: jego stałych siecio-



Rys. 38. Grupa przestrzenna $Pca2_1$ (klasa symetrii *mm*): a) Rozmieszczenie elementów symetrii w komórce elementarnej; wszystkie elementy symetrii są prostopadłe do płaszczyzny rysunku. b) Zespół pozycji symetrycznie równoznacznych w położeniu ogólnym; znak + oznacza, że punkt znajduje się nad płaszczyzną rysunku w odległości *z* od niej; przecinek wpisany w kółko oznacza, że dany punkt powstał w wyniku odbicia w płaszczyźnie symetrii (zwykłej lub poślizgu); $^1/_2+$ przy punkcie znaczy, że punkt znajduje się nad płaszczyzną rysunku na wysokości $z+{}^1/_2$ odcinka translacji prostopadłego do płaszczyzny rysunku, liczebność pozycji ogólnej 4; współrzędne punktów: xyz; $\bar{x},\bar{y},{}^1/_2+z$; ${}^1/_2-x, y, {}^1/_2+z$; ${}^1/_2+x,\bar{y},z$



Rys. 39. Grupa przestrzenna *Pba*2 (klasa symetrii *mm*): a) elementy symetrii, b) zespół pozycji symetrycznie równoznacznych w położeniu ogólnym; liczebność pozycji 4; współrzędne punktów: xyz, $\bar{x}\bar{y}z$; ${}^1/_2-x, {}^1/_2+y,z$; ${}^1/_2+x, {}^1/_2-y,z$



Rys. 40. Grupa przestrzenna *Iba*2 (klasa symetrii *mm*): a) elementy symetrii, b) zespół pozycji symetrycznie równoznacznych w położeniu ogólnym; liczebność pozycji 8; współrzędne punktów: xyz; $\bar{x}\bar{y}z$; $\bar{x},y,{}^1/_2+z$; $x,\bar{y},{}^1/_2+z$; ${}^1/_2+x,{}^1/_2-y,z$; ${}^1/_2-x,{}^1/_2+y,z$; ${}^1/_2-x,{}^1/_2-y,{}^1/_2+z$; ${}^1/_2+x,{}^1/_2+y,{}^1/_2+z$

wych, liczby atomów (jonów, cząsteczek) znajdujących się w komórce elementarnej oraz grupy przestrzennej. Bez znajomości grupy przestrzennej jest wręcz niemożliwe dalsze badanie struktury, tj. wyznaczenie pozycji atomów w komórce elementarnej. Z kolei bez znajomości współrzędnych $\cdot xyz \cdot$ atomów nie można określić długości wiązań, kątów między wiązaniami, wzajemnej konfiguracji atomów a w rezultacie i kształtu czy to np. kompleksowego jonu czy też całej cząsteczki.

znaczenie grup przestrzennych

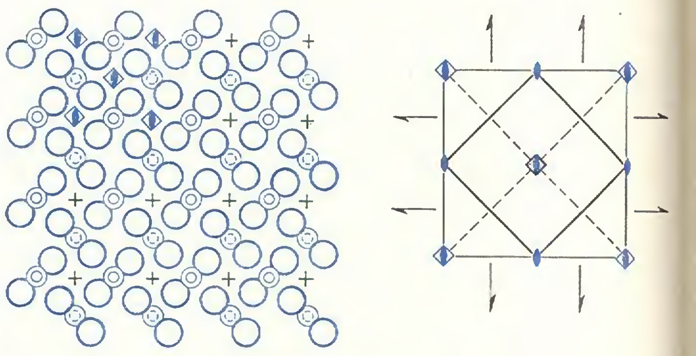
Dlaczego tak ważne jest wyznaczenie grupy przestrzennej? Otóż wydawałoby się, że atomy czy zespoły atomów mogą w komórkach elementarnych ciał krystalicznych zajmować miejsca zupełnie dowolne, najwyżej określone położeniem atomu w cząsteczce. Tymczasem tak nie jest. W komórkach elementarnych atomy zajmują miejsca w określonych zespołach pozycji symetrycznie równoznacznych i to najczęściej tak, aby liczba atomów jednego rodzaju była zgodna z liczebnością pozycji. Tak więc, znajomość grupy przestrzennej ciała krystalicznego zmniejsza nieograniczoną liczbę dowolnych położeń w komórce elementarnej dla każdego z atomów, do ściśle określonych zespołów pozycji. Przy badaniu struktury kryształu zadanie sprowadza się więc do określenia, który zespół pozycji zajmowany jest przez dany rodzaj atomów. W obrębie danego zespołu wystarczy już tylko wyznaczyć parametry x,y,z dla jednego (!) atomu. Współrzędne pozostałych atomów zajmujących tę pozycję dane są wówczas niejako automatycznie. Warto tu wspomnieć, że liczebności pozycji są nieraz bardzo wysokie, np. 16, 24, 32, a w niektórych grupach przestrzennych układu regularnego — 48, 96 a nawet 192(!). Należy zauważyć, że mogą istnieć i takie kryształy, w których kilka różnych rodzajów atomów może zajmować kilka pozycji ogólnych o różnych parametrach x,y,z, a są i takie kryształy, w których atomy zajmują wyłącznie pozycje szczególne (np. kryształy Au, NaCl). Struktury różnych związków chemicznych należące do jednej grupy przestrzennej mogą więc różnić się między sobą rodzajami i liczbą zajętych przez atomy zespołów pozycji symetrycznie równoznacznych, a także wartościami liczbowymi parametrów x,y,z punktu przyjętego za wyjściowy w danym zespole pozycji.

Zobaczymy teraz, jak realizowane są w konkretnych przypadkach podane wyżej definicje i zależności. Rozpatrzmy w tym celu strukturę związku chemicznego o wzorze $CO(NH_2)_2$, nazywanego mocznikiem lub karbamidem.

struktura mocznika

Mocznik krystalizuje w układzie tetragonalnym, w klasie skalenoedru tetragonalnego $\bar{4}2m$, w grupie przestrzennej $P\bar{4}2_1m$. Kryształy ograniczone są zwykle ścianami czworościanu tetragonalnego {111} i słupa tetragonalnego {110}. W komórce elementarnej o krawędziach a = 5,661 Å, c = 4,712 Å znajdują się dwie cząsteczki $CO(NH_2)_2$. Cząsteczka $CO(NH_2)_2$ jest płaska, tj. wszystkie jej atomy leżą w jednej płaszczyźnie. Symetria własna cząsteczki jest mm. W komórce elementarnej należy w sumie rozmieścić: dwa atomy tlenu, dwa atomy węgla, cztery atomy azotu i osiem atomów wodoru. Stwierdzono, że atomy tlenu zajmują pozycję szczególną o liczebności 2 o współrzędnych $0\frac{1}{2}z$ i $\frac{1}{2}0\bar{z}$ ($z_O = 0{,}5998$); atomy węgla zajmują tę samą pozycję, z tym tylko, że inna jest wielkość parametru z ($z_C = 0{,}3308$); atomy azotu zajmują pozycję szczególną o liczebności 4, o współrzędnych $x,\frac{1}{2}-x,z;\ \bar{x},\frac{1}{2}+x,z;\ \frac{1}{2}-x,\bar{x},\bar{z};\ \frac{1}{2}+x,x,\bar{z}$ ($x_N = 0{,}1419$, $z_N = 0{,}1857$). Atomy wodoru zajmują dwa te same zespoły pozycji co i atomy azotu, oczywiście przy innych parametrach x i z ($x_{H1} = 0{,}2390$, $z_{H1} = 0{,}2770$, $x_{H2} = 0{,}1240$, $z_{H2} = 0{,}0460$). Jak widać, w krysztale mocznika pozycja ogólna grupy przestrzennej wcale nie jest zajęta przez atomy.

Strukturę mocznika badano metodami rentgenostrukturalnymi, elektronograficznymi i neutronograficznymi. Dla znajdujących się w komórce elementarnej 16 atomów należało wyznaczyć 48 parametrów x,y,z. Jednakże ze względu na symetrię grupy przestrzennej, wystarczy w tej strukturze znaleźć położenia tylko jednego atomu tlenu, jednego atomu węgla, jednego atomu azotu i dwóch atomów wodoru (czyli położenia tylko 5 atomów (!); położenia pozostałych atomów wynikają automatycznie z zespołów pozycji symetrycznie równoznacznych), a więc w sumie wyznaczenie 15 parametrów x,y,z. Jak się okazało, ze względu na to, że atomy zajmują pozycje szczególne wystarczające było określenie tylko 8 (!) parametrów (z_O, z_C, x_N, z_N, x_{H1}, z_{H1}, x_{H2}, z_{H2}). Tak więc znajomość grupy przestrzennej zredukowała liczbę koniecznych do wyznaczenia parametrów z 48 do 8. Położenie



Rys. 41. Struktura mocznika $CO(NH_2)_2$: a) Rzut struktury na płaszczyznę XY; zaznaczono położenie osi czterokrotnych inwersyjnych w jednej komórce elementarnej. b) Elementy symetrii grupy przestrzennej $P\bar{4}2_1m$

cząsteczek mocznika w komórce elementarnej oraz usytuowanie ich względem elementów symetrii grupy przestrzennej $P\bar{4}2_1m$ pokazane jest na rys. 41. Na podstawie współrzędnych xyz atomów można łatwo obliczyć długości wiązań między atomami, kąty między wiązaniami oraz określić kształt cząsteczki (rys. 22). Cząsteczki połączone są między sobą wiązaniami wodorowymi: każdy atom tlenu tworzy cztery wiązania wodorowe z atomami azotu należącymi do trzech różnych cząsteczek (rys. 2).

Jak wynika z opisu struktury mocznika, w komórce elementarnej pewna liczba atomów (5 — w opisywanym przypadku) tworzy swego rodzaju zespół powtarzany elementami symetrii grupy przestrzennej w całej objętości komórki elementarnej (no i w całym krysztale). Taka grupa atomów nazywa się motywem struktury, a część komórki elementarnej, w której znajduje się motyw, jest asymetryczną częścią komórki. Znając grupę przestrzenną danego ciała krystalicznego wystarczy wyznaczyć współrzędne atomów tylko w motywie struktury, a następnie można powtórzyć ten motyw w całym krysztale.

motyw struktury

Podobne zależności występują także w strukturach innych pierwiastków czy związków chemicznych.

Problemy krystalografii współczesnej

Przedstawiona teoria symetrii jest już dziś teorią klasyczną, dotyczącą kryształów idealnych. Niewzruszoność praw klasycznej krystalografii zostaje poważnie zachwiana, gdy budowa kryształów odbiega od idealnej symetrii sieciowej. Przyjmuje się, że w kryształach idealnych każdy tego samego rodzaju atom, pełniący w strukturze tę samą funkcję, znajduje się w jednakowych warunkach geometrycznych i fizycznych. W rzeczywistości takie kryształy w przyrodzie nie występują i bardzo trudne jest ich otrzymanie sztuczne. Idealny kryształ musiałby być w zasadzie kryształem nieskończonym, stwierdzono bowiem, że

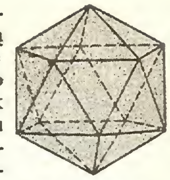
na powierzchni kryształu atomy układają się inaczej niż w jego wnętrzu.

struktury biologiczne

W latach trzydziestych rentgenostrukturalne metody badania kryształów zaczęto stosować do badania struktur biologicznych, m.in. kryształów białek. I tu zaczęły się dziać dziwne rzeczy, gdyż wkrótce okazało się, że metody klasycznej krystalografii są niewystarczające dla interpretacji wyników tych badań. Ogólnie przyjmuje się, że kryształami są struktury uporządkowane w trzech kierunkach przestrzeni. Możliwe są jednak i struktury uporządkowane tylko w jednym lub dwu wymiarach. Struktury takie nie podlegają ograniczeniom klasycznej krystalografii. Mogą w nich występować np. 5- i 7-krotne osie symetrii, a jednostkami strukturalnymi mogą być prócz atomów, jonów czy cząsteczek np. wirusy (il. 64, tabl. 17). Wiele struktur biologicznych tworzy kryształy dwuwymiarowe. Powierzchniami dwuwymiarowych kryształów są otwarte cylindry i zamknięte kule, na których rozmieszczane są równoważne cząstki. Doświadczalnie, metodą dyfrakcji promieni rentgenowskich stwierdzono istnienie wirusów mających postać kul, cylindrów oraz wielościanów.

W wyniku badań rentgenostrukturalnych okazało się, iż symetria helisków α w białkach fibrylarnych nie mieści się w klasycznej teorii symetrii kryształów, gdyż np. w polipeptydzie na jeden zwój heliksu nie przypada całkowita liczba grup aminokwasowych (jest to sprzeczne z własnościami klasycznej osi śrubowej). Bywa np., że 18 reszt aminokwasowych mieści się w 5 zwojach heliksu, co w sumie daje osiemnastokrotną oś śrubową o składowej translacyjnej $^5/_{18}$. Symbol takiej osi: 18_5.

Wirusy globularne krystalizują na tyle dobrze, że można otrzymać obrazy dyfrakcyjne ich monokryształów. Kryształy wirusa zarazy pomidorowej mają komórkę elementarną w kształcie sześcianu ($a = 386$ Å), lecz symetria tej komórki nie należy do układu regularnego. Dla poszczególnych cząsteczek wirusa charakterystyczna jest symetria ikosaedru i w związku z tym nie mają one symetrii krystalograficznej. Ikosaedr (rys. 42) jest wielościanem zbudowanym z 20 ścian będących jednakowymi równobocznymi trójkątami. W ikosaedrze istnieje 6 pięciokrotnych osi symetrii, 10 trójkrotnych i 15 dwukrotnych osi symetrii. Taka symetria charakterystyczna jest dla wszystkich „sferycznych" wirusów roślinnych i niektórych wirusów ssaków.



Rys. 42. Ikosaedr

Z tych przykładów widać, że zaistniała potrzeba stworzenia nowej, uogólnionej, rozszerzonej krystalografii, w której proste i płaszczyzny sieciowe zostałyby zastąpione dowolnymi krzywymi i powierzchniami pełniącymi ich rolę w strukturach. Klasyczna krystalografia, z jej prostymi i płaszczyznami sieciowymi byłaby szczególnym przypadkiem tej nowej, uogólnionej krystalografii (→ Współczesne teorie symetrii w krystalografii).

I. D. Bernal, S. Ch. Karlile *Kristallografija*, 13, 927 (1968); G. B. Bokij, *Kristallochimija*, Moskwa 1971; M. J. Buerger *Elementary Crystallography*, New York 1956; C. W. Bunn *Chemical Crystallography*, Oxford 1961; G. Burns, A. M. Glazer *Space Groups for Solid State Scientists*, New York, London, 1978; J. Chojnacki *Elementy krystalografii chemicznej i fizycznej*, Warszawa 1973; A. I. Kitajgorodskij *Organiczeskaja kristallochimija*, Moskwa 1955; I. Kostow *Kristallografija*, Moskwa 1965; S. C. Nyburg *X-Ray Analysis of Organic Structures*, New York; London, 1941; I. I. Szafranowskij *Kristallografija*, 2, 326 (1957), I. I. Szafranowskij *Lekcii po kristallomorfologii*, Moskwa 1968; B. K. Wajnsztejn *Sowriemiennaja kristallografija*, Tom I, Moskwa 1979.

Otrzymywanie monokryształów

Zdzisław Sołtys

Duże, przejrzyste pojedyncze kryształy zawsze wzbudzały podziw ludzi symetrią swojego kształtu, zdolnością załamywania światła i barwami, a jeśli odznaczały się przy tym twardością, nazywano je kamieniami szlachetnymi. Niektóre kamienie szlachetne, pojedyncze kryształy minerałów, szczególnie po odpowiedniej obróbce szlifierskiej, oprawione w złoto lub platynę używane były i są jako cenne ozdoby. Najpiękniejsze i najdroższe okazy kryształów od wielu wieków przechowywane są w skarbcach władców, w prywatnych kolekcjach miłośników minerałów lub w muzeach.

Rozwój nauk przyrodniczych, głównie fizyki ciała stałego, stworzył zapotrzebowanie na duże pojedyncze kryształy wielu substancji nieorganicznych i organicznych. W krótkim czasie okazało się, że naturalne źródła są niewystarczające lub w ogóle nie istnieją. Stan taki spowodował rozwój laboratoryjnych i przemysłowych metod wytwarzania dużych, pojedynczych kryształów zwanych monokryształami. Właściwy rozwój metod wytwarzania monokryształów rozpoczął się od czasu wynalezienia tranzystora przed 30 laty. Zastosowanie monokryształów jako materiałów konstrukcyjnych głównie w elektronice, a także w innych gałęziach techniki, stało się początkiem tzw. rewolucji materiałowej (tabela).

mono-kryształy

Obecnie coraz mniej stosuje się materiałów w postaci naturalnej lub drobnokrystalicznej zastępując je monokryształami. Materiały monokrystaliczne odznaczające się doskonałą jednorodnością fizyczną i chemiczną ujawniły nowe właściwości, nie wykazywane przez te same materiały w postaci naturalnej — zanieczyszczonej lub drobnokrystalicznej. (Szczególnie interesujące mogą okazać się właściwości polimerów organicznych w postaci monokrystalicznej). Od monokryształów nie wymaga się zewnętrznego podobień-

zastosowanie monokryształów

Zastosowanie niektórych materiałów monokrystalicznych

Zastosowanie	Monokryształy*
Przyrządy półprzewodnikowe: diody tranzystory, układy scalone	Si, Ge, GaAs
Lasery półprzewodnikowe	rubin Al_2O_3:Cr, $CaWO_4$, CaF_2, GaAs
Generatory ze stabilizacją częstotliwości	kwarc α-SiO_2
Modulatory i rezonatory laserowe wzmacniacze parametryczne	KH_2PO_4, $LiNbO_3$, $LiTaO_3$, (Ba, Na)$(NbO_3)_3$
Przetworniki elektromechaniczne (piezoelektryczne)	kwarc α-SiO_2, $BaTiO_3$, GaAs, sól Seignette'a $NaK(C_4H_4O_6)\cdot 4H_2O$
Przyrządy optyczne: soczewki, pryzmaty, polaryzatory	kalcyt $CaCO_3$, CaF_2, SiO_2, $NaNO_3$
Optyka w podczerwieni: soczewki, pryzmaty	Ge, Si, LiF, KBr
Przyrządy luminescencyjne	GaP, Ga(As, P), GaAs
Detektory promieniowania	antracen $C_{14}H_{10}$, KCl, Si, GaAs, NaJ:Tl, Ge:Li, $NH_4H_4PO_4$
Wzmacniacze ultradźwiękowe	CdS
Wzmacniacze mikrofalowe	rubin Al_2O_3:Cr
Podzespoły mikrofalowe magnet.	granaty, np. $Y_3Fe_5O_{12}$ (YIG)
Elementy pamięci maszyn cyfrowych	granat $Eu_{1,9}Gd_{1,1}Al_{0,5}Fe_{4,5}O_{12}$ na podłożu granatu $Gd_3Al_5O_{12}$
Monochromatory, filtry strumienia neutronów	Cu, Zn, stop 0,92 Co + + 0,08Fe
Materiały ścierne i elementy tnące	korund α-Al_2O_3, szafir Al_2O_3:Fe, Ti, diament C, węglik krzemu SiC
Łożyska w przyrządach pomiarowych i mechanizmach precyzyjnych	szafir, korund, rubiny
Klejnoty i ozdoby	rubiny, szafiry

*Zapis Al_2O_3:Cr oznacza, że chrom jest niewielką domieszką w krysztale, celowo wprowadzonym zanieczyszczeniem. Zapis typu Ga(As, P) oznacza, że pewna część atomów arsenu jest zastąpiona przez atomy fosforu w krysztale arsenku galu GaAs.

stwa (choć nie zawsze) do naturalnych pojedynczych kryształów, ale za to wysokiej czystości chemicznej (wymagana dopuszczalna zawartość zanieczyszczeń czasem jest znacznie poniżej 10^{-8} %) oraz dużej doskonałości krystalograficznej (jednolita sieć przestrzenna bez dyslokacji i innych defektów w rozmieszczeniu atomów w krysztale, → Budowa kryształów), braku naprężeń mechanicznych itp. Uzyskiwane monokryształy powinny zawierać celowo wprowadzone w żądanej ilości domieszki różnych atomów rozmieszczonych w przestrzeni monokryształu równomiernie lub w inny, z góry określony sposób. Wszystkie te wymagania spełnia współczesna technologia wytwarzania monokryształów rozmaitych substancji, mimo że procesy monokrystalizacji (wzrostu monokryształów) nie są jeszcze teoretycznie dobrze wyjaśnione. Umiejętność „hodowli" monokryształów jest jeszcze do dziś bardziej sztuką niż nauką.

Monokryształy tworzą się w procesie powstawania fazy stałej (krystalicznej) z fazy gazowej albo z fazy ciekłej lub w procesie przemiany jednej fazy stałej w inną. Powstanie fazy stałej w środowisku gazowym lub ciekłym polega na wytworzeniu trwałego zarodka fazy stałej i utrzymywaniu w tym ośrodku warunków, w których zarodek ten może rosnąć. Warunkiem koniecznym do wytworzenia się zarodków fazy stałej jest powstanie przesycenia w ośrodku gazowym lub w roztworze, a przechłodzenia — w ośrodku ciekłym. Miarą przesycenia jest różnica ciśnienia (stężenia) ponad ciśnienie (stężenie) równowagowe w danej temperaturze. Miarą przechłodzenia jest różnica temperatury układu poniżej temperatury topnienia fazy stałej. Im większe przesycenie lub przechłodzenie, tym łatwiej i tym więcej powstaje zarodków drobnych. Dla uzyskania kilku trwałych i względnie dużych zarodków i ich dalszego wzrostu konieczne jest małe przesycenie (przechłodzenie).

powstawanie zarodków

W hodowli monokryształów bardzo często, a w niektórych wypadkach wyłącznie, wykorzystuje się nie przypadkowe powstawanie zarodków fazy stałej wskutek przechłodzenia czy przesycenia, lecz przeprowadza się proces polegający na powiększeniu objętości niewielkiego monokryształu, zwanego zarodzią, umieszczonego w ośrodku dostarczającym materiału do jego wzrostu. Zarodzie otrzymuje się z przypadkowo powstałych w przesyconym lub przechłodzonym roztworze zarodków, jednakże tak kieruje się procesem, aby powstało ich niewiele. Często duże zarodzie kształtuje się przez rozcinanie lub szlifowanie w postać pręcików o wybranej orientacji krystalograficznej powierzchni czołowej zarodzi. Wtedy dalszy wzrost zarodzi monokryształu może odbywać się tylko zgodnie z tym wybranym kierunkiem.

zarodź kryształu

Wzrost monokryształów z roztworów ciekłych

Proces wzrostu monokryształów z roztworów wodnych

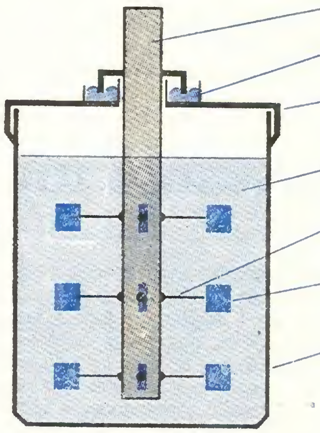
Najstarszą metodą otrzymywania substancji w postaci krystalicznej jest hodowla kryształów z roztworów wodnych. Tak otrzymuje się sól kuchenną z solanek, nawozy sztuczne itp. Tak też wytwarza się cukier, pierwszą drobnokrystaliczną substancję otrzymywaną przemysłowo na wielką skalę, a nie występującą w przyrodzie w postaci krystalicznej. Zwykłe przemysłowe procesy krystalizacji mają na celu wydzielenie z roztworu w maksymalnej ilości jednorodnego produktu w postaci drobnych kryształów o prawie jednakowych rozmiarach, zaś doskonałość otrzymanych kryształów jest bez znaczenia.

W procesie wytwarzania dużych kryształów o dużej doskonałości krystalograficznej z roztworów ciekłych zasadniczo nie wykorzystuje się spontanicznego powstawania zarodków w roztworze przesyconym i ich dalszego wzrostu, lecz kontroluje się wzrost kryształu na zarodzi zawieszonej w roztworze o stałym przesyceniu (stałe przesycanie roztworu uzyskuje się przez bardzo powolne w miarę wzrostu kryształu ochładzanie roztworu według odpowiedniego programu). Szybkość wzrostu poszczególnych płaskich ścian kryształu zawieszonego w roztworze zależy od wielu czynników, m.in. od stężenia, temperatury, lepkości roztworu, dyfuzji substancji rozpuszczonej do rosnącej powierzchni kryształu, zawartości zanieczyszczeń oraz celowo dodanych domieszek. Od parametrów roztworu zależą także kształt otrzymywanych kryształów i ich doskonałość. Na przykład podczas krystalizacji chlorku sodowego NaCl z czystego roztworu powstają małe nieprzejrzyste kryształy, ale dodanie 0,1% jonów Pb^{2+} do zakwaszonego roztworu NaCl powoduje, że w temperaturze 75°C otrzymuje się spore (6 × 6 × 6 mm) przejrzyste monokryształy NaCl o dużej doskonałości.

hodowla soli kuchennej

Najprostszy sposób otrzymywania monokryształów polega na zawieszeniu zarodzi na nici (druciku) w roztworze nasyconym w temperaturze wyższej od temperatury otoczenia; zarodź ta podczas naturalnego swobodnego stygnięcia roztworu powiększa swoje rozmiary. Doskonalsze kryształy uzyskuje się na zarodziach ruchomych. W krystalizatorze Holdena (rys. 1) zarodzie mocowane są na końcach prętów wolframowych zaciśniętych w ramionach obracającego się bardzo powoli (i ze zmianą kierunku obrotu) mieszadła zanurzonego w naczyniu z roztworem nasyconym. Delikatne mieszanie powoduje wyrównywanie się stężenia roztworu ochładzanego w sposób ciągły o 0,1–1,5°C na dobę. W dużych przemysłowych krystalizatorach typu Holdena otrzymuje się np. monokryształy fosforanu dwuwodoroamonowego $NH_4H_2PO_4$ o masie ok. 20 kg; czas wzrostu takiego monokryształu wynosi 4 miesiące.

metoda Holdena



Rys. 1. Schemat krystalizatora Holdena

Ogólna masa monokryształów związków nieorganicznych i organicznych hodowanych na świecie z roztworów wodnych w zwykłych temperaturach przekracza kilkaset ton rocznie.

Proces wzrostu monokryształów z topnika

W podwyższonych temperaturach stopione sole, tlenki, wodorotlenki lub metale stają się rozpuszczalnikami (topnikami) innych substancji. Mieszaninę substancji, która ma być otrzymana w postaci monokryształu, i topnika, umieszcza się w tyglu odpornym na działanie stopu i ogrzewa się ją do stopienia. Skład mieszaniny dobiera się tak, aby uzyskać w topniku roztwór nasycony substancji krystalizującej. Początkowo tygiel ochładza się powoli z szybkością

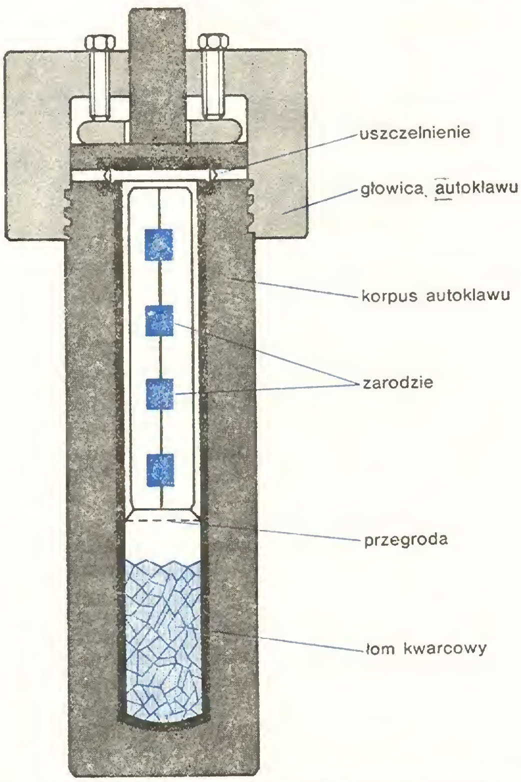
kilku, kilkunastu stopni na godzinę. W pewnym momencie powstaje przesycenie i rozpoczyna się tworzenie zarodków a następnie dalszy ich wzrost, ewentualnie wzrost umieszczonej w tyglu zarodzi. Przez cały czas krystalizacji, zachodzącej zwykle w temperaturze 800–900°C, temperaturę tygla obniża się bardzo powoli, po czym ochładza się go szybciej. Chłodny topnik i kryształy wymywa się z tygla gorącą wodą, roztworami kwasów lub zasad w zależności od rodzaju kryształów i rodzaju topnika, a następnie wytworzone kryształy oddziela się od siebie. Monokryształy otrzymane tą metodą mogą zawierać pewną ilość zatrzymanego (okludowanego) topnika.

hodowla tytanianu baru

Z roztworów w topniku otrzymuje się np. monokryształy tytanianu baru $BaTiO_3$ ze stopionego fluorku potasowego KF (zakres temperatury krystalizacji 1200–850°C, szybkość studzenia 1–5°C na godzinę); monokryształy fosforku galu GaP z roztworu w galu (zakres temperatury krystalizacji 1200–900°C, szybkość studzenia 1°C na godzinę).

Krystalizacja hydrotermalna

Dość dawno zauważono że kwarc, znany jako substancja nierozpuszczalna w wodzie w temperaturach zwykłych, wykazuje pewną rozpuszczalność w temperaturach w pobliżu lub powyżej jej temperatury krytycznej (374,2°C), szczególnie jeśli do rozpuszczania



Rys. 2. Autoklaw do monokrystalizacji kwarcu metodą hydrotermalną

użyje się słabo zasadowego roztworu. Proces nosi nazwę krystalizacji hydrotermalnej. Przebiega on w grubościennych stalowych autoklawach, pod ciśnieniem 145 MPa (rys. 2). W autoklawie umieszcza się na dnie złom kwarcowy o odpowiednim uziarnieniu.

hodowla kwarcu α

Następnie autoklaw napełnia się roztworem wodorotlenku sodowego NaOH i ogrzewa w taki sposób, aby temperatura dołu autoklawu wynosiła ok. 400°C, a temperatura obszaru krystalizacji wynosiła 360°C. Monokryształy kwarcu rosną w tych warunkach ok. 6 mm dziennie w kierunku [0001].

Monokryształy kwarcu wyhodowane metodą hydrotermalną w dużym autoklawie w Western Electric Company pokazane na il. 3 (tabl. 1) mają masę ok. 800 g każdy. Metodą hydrotermalną otrzymuje się i inne monokryształy, np. korundu α-Al_2O_3 z wodnego 1 molowego roztworu węglanu sodowego Na_2CO_3 (temperatura krystalizacji 405°C, temperatura rozpuszczania 435°C), a kryształy magnetytu Fe_3O_4 z 0,5 molowego roztworu wodnego chlorku amonowego NH_4Cl (odpowiednie temperatury 530°C i 515°C).

hodowla korundu i magnetytu

Wzrost monokryształów podczas krzepnięcia substancji stopionej

Proces krzepnięcia stopionej substancji czyli proces krystalizacji przebiega pod warunkiem odprowadzania ciepła z fazy ciekłej i rozpoczyna się od powstania unoszących się w jej objętości trwałych zarodków fazy stałej (kryształów), jeżeli występuje, przynajmniej lokalnie, temperatura nieco niższa od temperatury topnienia, tzn. jeżeli występuje lokalnie przechłodzona faza ciekła. Dalszy wzrost kryształów na tych zarodkach ewentualnie na zarodziach umieszczonych w stopionej substancji następuje w wyniku umiejscawiania się atomów (jonów, cząsteczek) fazy ciekłej w węzłach sieci kryształu na powierzchni zarodków trwałych lub zarodzi. Proces taki może przebiegać w sposób naturalny lub może być sterowany. Wszystkie procesy otrzymywania monokryształów na skalę techniczną są procesami sterowanymi. Stosowana metoda wzrostu monokryształu w procesie przemiany fazy ciekłej w stałą zależy od właściwości fizycznych substancji krystalizowanej.

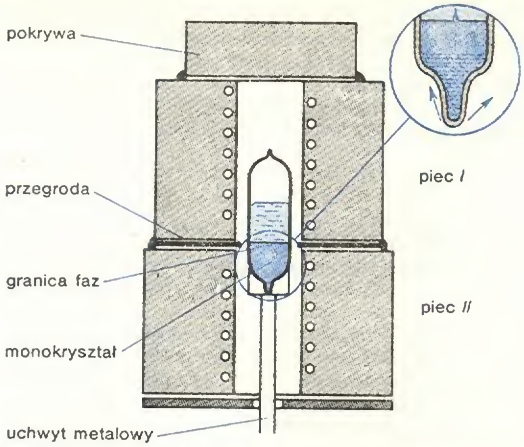
postać otrzymywanych monokryształów

Monokryształy otrzymywane podczas krzepnięcia substancji roztopionej mają zwykle kształt prętów lub wydłużonych bloków. W temperaturze krzepnięcia substancja stopiona ma dużą lepkość, tak że nie mogą wytworzyć się płaskie powierzchnie ścian charakterystyczne dla substancji krystalicznych, uzyskiwanych podczas krystalizacji z roztworów lub powstające podczas kondensacji fazy gazowej. Otrzymywane pręty monokrystaliczne i bloki są właściwie krystalitami, jednakże powszechnie nazywa się je monokryształami. Niemniej cały taki krystalit — monokryształ ma jednolitą sieć przestrzenną, wykazującą tę samą symetrię i stałe sieciowe, jaką wykazuje kryształ z wykształconymi ścianami. Średnice monokryształów wytwarzanych obecnie wynoszą od kilkunastu a nawet kilkudziesięciu milimetrów, a długości od kilku do kilkudziesięciu centymetrów. Masa pojedynczych bloków monokryształów, np. krzemu, dochodzi do ok. 20 kg (cena takiego monokryształu wynosi nawet ponad 20 tys. dolarów za 1 kg). Wytwarza się obecnie na większą skalę monokryształy ponad 50 pierwiastków, bardzo wielu związków chemicznych, a nawet monokryształy niektórych stopów. W większości metod otrzymywania monokryształów podczas krzepnięcia stopionej substancji wykorzystuje się gotowe zarodzie.

Monokrystalizacja metodą Bridgmana–Stockbargera

Najpowszechniej stosowaną metodą wytwarzania zarodków kryształu we wstępnej fazie procesu jest metoda Bridgmana–Stockbargera. Sterowanie wzrostem zarodków powstających w obszarze gradientu temperatury w ampule-tyglu polega na odpowiednim ukształtowaniu dna tygla. Tylko zarodek ułożony równolegle do osi pionowej może rosnąć i wypełnić całą ampułę. Wszystkie inne zarodki kończą swój wzrost na ściance wydłużonej części dna ampuły-tygla. Po zakończeniu tej fazy tygiel opuszcza się mechanicznie w dół pieca — w obszar o niższej temperaturze —

z prędkością kilku milimetrów na godzinę i uzyskuje się ostatecznie monokryształ wypełniający ampułę-tygiel (rys. 3). Ampułę (łódkę) można również umieścić



Rys. 3. Schemat urządzenia do otrzymywania monokryształów metodą Bridgmana–Stockbargera

w piecu poziomo. Substancje łatwo parujące lub rozkładające się w temperaturze niższej niż temperatura topnienia umieszcza się w szczelnie zespawanych ampułach, zwykle kwarcowych, ustawionych poziomo.

zastosowanie metody Bridgmana–Stockbargera

Metodę Bridgmana–Stockbargera stosuje się przy wytwarzaniu monokryształów związków półprzewodnikowych łatwo rozkładających się poniżej temperatury topnienia, większości metali oraz niektórych związków nieorganicznych i organicznych (tabela).

Przykłady monokrystalizacji metodą Bridgmana–Stockbargera

Substancja	Temperatura topnienia °C	Gradient temperatury °C/mm	Szybkość wzrostu (opuszczania tygla) mm/h	Materiał tygla (ampuły)	Atmosfera
Argon Ar	−189,4	0,5	60	szkło	argon
Miedź Cu	1083	12	16	grafit	azot
Fluorek wapnia CaF_2	1392	10	1	grafit	próżnia
Arsen As	814 (3,6 MPa)	8 (poziomy)	ok.10	kwarc	pary arsenu (6 MPa)
Arsenek galu GaAs	1238 (rozkład)	30 (poziomy)	20	kwarc	pary arsenu (0,09 MPa)

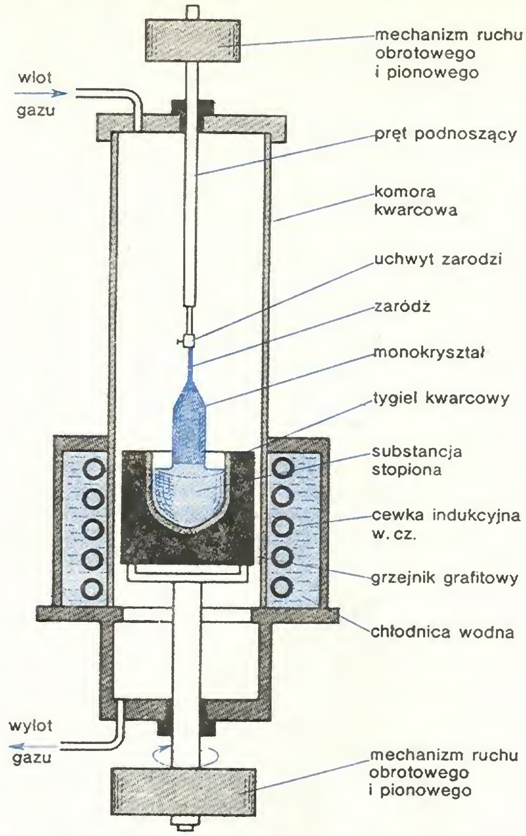
Monokrystalizacja metodą Czochralskiego

W metodzie Czochralskiego również wykorzystuje się wzrost monokryształu na gotowej zarodzi umocowanej w uchwycie i stykającej się z powierzchnią substancji stopionej. W miarę odprowadzania ciepła przez zaródź i przez uchwyt na granicy faz ciecz–zaródź narasta monokryształ powoli unoszony uchwytem zarodzi. Podczas obserwacji tego procesu odnosi się wrażenie, że monokryształ jest wyciągany z fazy stopionej znajdującej się w tyglu; stąd druga nazwa metody — metoda wyciągania monokryształu (rys. 4).

metoda wyciągania

Jeśli szybkość krystalizacji jest większa niż szybkość wyciągania, średnica powstającego kryształu jest coraz większa i odwrotnie. Podobny efekt uzyskuje się w wyniku obniżenia temperatury przy granicy faz przez obniżenie temperatury substancji stopionej znajdującej się w tyglu. Natomiast gdy szybkości wyciągania i krystalizacji są sobie równe, a ponadto uchwyt z monokryształem wprawiony jest w ruch obrotowy, to uzyskujemy monokryształ o stałej średnicy.

Istnieją liczne odmiany metody Czochralskiego dostosowane do właściwości substancji (tabela). Proces wyciągania kryształów prowadzi się w atmosferze



Rys. 4. Schemat urządzenia do wyciągania monokryształów metodą Czochralskiego

gazu ochronnego (argon, azot, wodór, hel), w próżni lub w układzie otwartym na powietrzu. Monokryształy substancji rozkładających się w pobliżu temperatury topnienia otrzymuje się w komorze pod odpowiednim ciśnieniem par składników, mniejszym lub większym od ciśnienia atmosferycznego. Tygle do topienia ogrzewane są grzejnikami oporowymi. Dość

Przykłady monokrystalizacji metodą Czochralskiego

Substancja	Temperatura topnienia °C	Prędkość wyciągania cm/h	Atmosfera, warunki procesu	Rozmiary monokryształu		
				średnica mm	długość mm	masa kg
Bizmut Bi	271	3	próżnia	75	180	
German Ge	938	0,6–36	próżnia, tygiel grafitowy	35	300	
Krzem Si	1410	3–6	próżnia lub argon	100	1000	20
Arsenek galu GaAs	1238 (rozkład)	1,5–3,5	ciśnienie As 0,09 MPa lub powierzchnia pokryta ochronnie stopionym boraksem i ciśn. 0,09 MPa azotu	35	200	
Wolframian wapnia CaWO	1530	0,5–2	powietrze, grzanie prądami w.cz., tygiel rodowy			

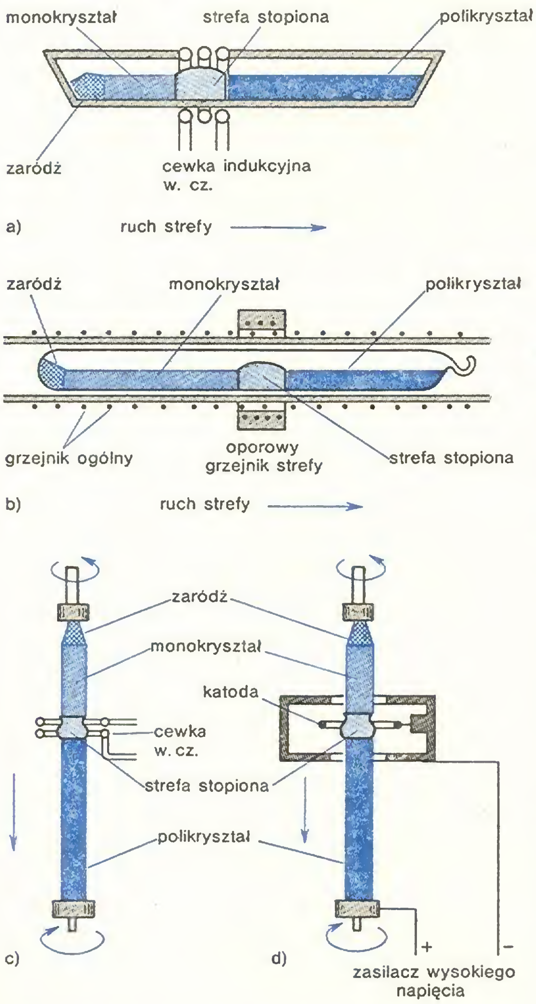
wygodne jest stosowanie ogrzewania, nawet do wysokich temperatur, za pomocą indukcyjnych prądów wielkiej częstotliwości (w.cz.). W wypadku gdy topimy metale lub inne substancje przewodzące, one same są odbiornikami energii wytwarzanej przez pole wielkiej częstotliwości i mogą zostać ogrzane aż do stopienia w tyglu z materiału nieprzewodzącego prądu elektrycznego. Jeśli są to substancje nieprzewodzące, tygiel kwarcowy umieszcza się w dodatkowym tyglu grafitowym.

zastosowanie metody Czochralskiego

Monokryształy otrzymane metodą Czochralskiego odznaczają się bardzo wysoką doskonałością krystalograficzną; niektóre są bezdyslokacyjne, nie obserwuje się także występowania w nich naprężeń mechanicznych. Ponieważ podczas procesu krzepnięcia granica między fazą ciekłą i stałą jest płaska, monokryształy otrzymane tą metodą są szczególnie przydatne do produkcji cienkich płytek, które tnie się z dużego monokryształu prostopadle do osi wzrostu. Płytki te nie wykazują niejednorodności własności fizycznych wzdłuż promienia.

Topienie strefowe

Monokryształy pierwiastków i niektórych stopów o wysokich temperaturach topnienia (ponad 1400°C) można otrzymać w procesie topienia strefowego. Na jednym końcu łódki (tygla) umieszcza się zaródź,



Rys. 5. Sposoby wytwarzania strefy stopionej w procesie monokrystalizacji a) w otwartej łódce — grzanie prądami w.cz.; b) w zatopionej ampule — grzejnik oporowy strefy lub prądy w.cz.; c) sposób beztyglowy — grzanie prądami w.cz.; d) sposób beztyglowy — bombardowanie strefy elektronami w próżni

a w jej pozostałej części materiał polikrystaliczny. Między zarodzią i polikryształem wytwarza się łączącą je strefę stopioną i przesuwa się tę strefę z prędkością krystalizacji w kierunku od zarodzi (rys. 5a i b). W ten sposób uzyskuje się monokryształ o przekroju poprzecznym nadanym przez kształt łódki.

metoda pływającej strefy

Najbardziej interesująca jest beztyglowa odmiana tej metody, polegająca na tym, że w czasie procesu monokrystalizacji pręt polikrystaliczny i zaródź znajdują się w pozycji pionowej. Stopioną strefę między dwoma odcinkami „pręta" utrzymują siły napięcia powierzchniowego i jej grubość maksymalna jest ograniczona. W metalach i substancjach przewodzących prąd elektryczny stopioną strefę uzyskuje się najczęściej za pomocą cewki indukcyjnej wielkiej częstości w atmosferze gazu ochronnego lub przez bombardowanie wąskiej strefy na obwodzie pręta odpowiednio ukształtowaną wiązką elektronów w próżni (rys. 5c i d). Otrzymywane tą metodą pręty monokrystaliczne mają przekrój kołowy, podobnie jak pręty otrzymywane metodą Czochralskiego. W zależności od wielkości napięcia powierzchniowego w temperaturze topnienia na powierzchni zewnętrznej pręta mogą się zaznaczyć słabiej lub mocniej ślady symetrii krystalograficznej w stosunku do kierunku wybranego dla zarodzi. Na przykład dla kierunku [111] wybranego dla zarodzi germanu przekrój pręta mono-

Przykłady monokrystalizacji metodą pływającej strefy

Substancja	Temperatura topnienia, °C	Sposób wytworzenia strefy	Prędkość przesuwania strefy, cm/h	Rozmiary strefy	
				średnica mm	wysokość mm
Krzem Si	1410	cewką w.cz.	3	25	17
Nikiel Ni	1450	bombardowanie elektronami	18	6	3
Tytan Ti	1725	cewką w.cz. w atmosferze argonu	6	10	10
Wofram W	3370	bombardowanie elektronami	18	3–6	3–6

krystalicznego germanu wykazuje wyraźniej symetrię osi trójkrotnej [111] niż przekrój pręta krzemu (il. 99, tabl. 25). Metoda topienia strefowego beztyglowego nosi nazwę metody pływającej strefy (tabela).

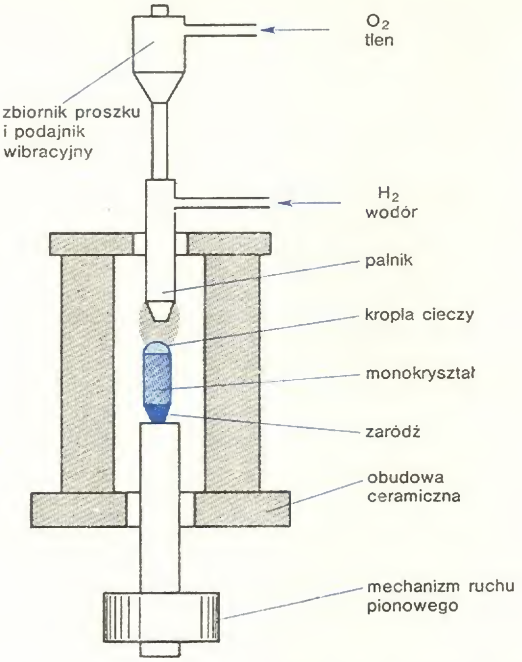
Monokrystalizacja metodą Verneuila

Modyfikacją metody monokrystalizacji na zarodzi w temperaturze krzepnięcia substancji jest metoda Verneuila. Jej oryginalność polega na dodawaniu materiału w postaci topiącego się proszku do kropli cieczy znajdującej się na wierzchołku wytwarzanego monokryształu (rys. 6). W pierwszym rozwiązaniu technicznym Verneuila (1902) służącym do otrzymywania monokryształów korundu (α-Al_2O_3), szafiru (Al_2O_3: Ti i Fe) i rubinu (Al_2O_3: Cr) posługiwano się palnikiem tlenowo-wodorowym do wytworzenia kropli i do topienia drobnokrystalicznego tlenku glinu (α-Al_2O_3) z ewentualnymi domieszkami lub mieszaniny tlenku magnezu MgO i tlenku glinu do wytworzenia monokrystalicznego spinelu $Mg(AlO_2)_2$. Monokryształy szafiru, korundu czy rubinu otrzymywane tą metodą były z reguły mało doskonałe i pękały po ostudzeniu wskutek naprężeń wywołanych występowaniem dużych gradientów temperatury podczas wzrostu kryształu będącego słabym przewodnikiem ciepła. Sztuczne rubiny stosowano początkowo głównie w jubilerstwie, jako łożyska (kamienie) w zegarkach lub jako pryzmaty do czułych wag.

korund, szafiry i rubiny

Dopiero rozwój fizyki i elektroniki ciała stałego po II wojnie światowej wpłynął na udoskonalenie metody

Verneuila i otrzymywane obecnie tą metodą monokryształy odznaczają się lepszymi właściwościami niż kryształy naturalne. Zastosowano precyzyjne urzą-



Rys. 6. Schemat urządzenia do otrzymywania monokryształów metodą Verneuila

dzenia do sterowania procesem wzrostu i usuwania naprężeń w kryształach oraz wprowadzono kilka nowych układów ogrzewania (palnik plazmowy, ogrzewanie prądami indukcyjnymi wielkiej częstotliwości, ogrzewanie zogniskowaną energią promieniowania łuku elektrycznego). Umożliwiło to otrzymywanie monokryształów nie tylko korundu, szafiru czy rubinu, ale też i in. trudno topliwych tlenków jak rutylu TiO_2 (temperatura topnienia $T_t = 1830°C$), tlenku cyrkonu ZrO_2 ($T_t = 2700°C$), tlenku magnezu MgO ($T_t = 2640°C$), tlenku wapnia CaO ($T_t = 2570°C$), tlenku niklu NiO ($T_t = 2090°C$) oraz ferrytów ($T_t = 1200$–$1800°C$).

zastosowanie metody Verneuila

Metodą Verneuila uzyskano też monokryształy węglików, azotków, borków, krzemków, berylków metali, substancji bardzo łatwo utleniających się i bardzo trudno topliwych ($T_t = 2000$–$4000°C$). Metodą Verneuila wytwarza się obecnie monokryształy ok. 100 substancji.

Wzrost monokryształów przez przemianę w fazie stałej

Otrzymywanie dużych monokryształów przez przemianę w fazie stałej stosuje się wtedy, gdy zależy nam na uzyskaniu monokryształów o określonym kształcie, np. w postaci cienkich taśm, drutów o małej średnicy, albo gdy ma być zachowana jednorodność składu stopu lub rozmieszczenie atomów domieszki. Jeśli substancja podlega przemianie fazowej w fazie stałej między temperaturą topnienia i temperaturą pokojową, metoda ta jest w zasadzie jedyną metodą otrzymywania monokryształu o określonej strukturze, szczególnie, gdy substancja ma wysoką temperaturę topnienia. Proces monokrystalizacji przez przemianę w fazie stałej jest możliwy dzięki naturalnej tendencji każdego układu do uzyskania minimum energii (monokryształ ma mniejszą energię niż polikryształ).

Wzrost monokryształów substancji niemetalicznych przez przemianę w fazie stałej jest procesem bardzo złożonym i zależy od wielu czynników. Przeważnie stosuje się spiekanie proszków sprasowanych w kształtki. Spiekanie odbywa się w temperaturze niewiele niższej od temperatury topnienia substancji. Uzyskuje się zwykle kilka (kilkanaście) kilkumilimetrowych ziaren monokrystalicznych w objętości próbki.

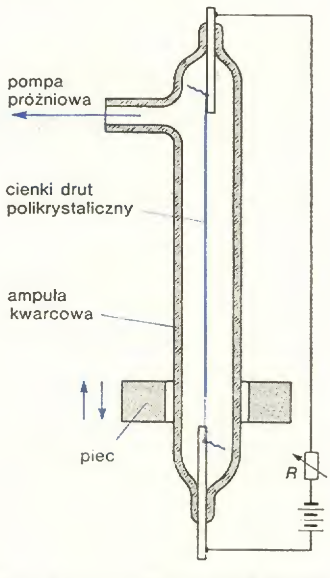
spiekanie proszków

Monokrystalizacja przez wyżarzanie

W wypadku metali poddaje się je najpierw odkształceniu prawie do granicy plastyczności przez walcowanie, zgniatanie, rozciąganie, kucie, przeciąganie itp. Nagromadzona w ten sposób w materiale energia wystarcza do przemiany polikrystalicznego metalu lub stopu w monokryształ. W zwykłych temperaturach przemiana taka jest bardzo powolna, podgrzewa się więc układ do odpowiedniej temperatury dla przyspieszenia przemiany i wywołania wzrostu jednego z ziaren polikryształu w duże ziarno monokrystaliczne (lub kilka dużych ziaren) wypełniające całą objętość użytej próbki. Małe ziarna zostają „pożarte" przez sąsiadujące z nimi ziarno rosnące.

Proces monokrystalizacji przez wyżarzenie odkształceń plastycznych bardzo często można przeprowadzić w aparaturze takiej samej jak przy metodzie Bridgmana–Stockbargera, z tym że stosujemy odwrotny gradient temperatury w stosunku do ruchu próbki w piecu i oczywiście nie używamy tygla, gdyż przemiana zachodzi w fazie stałej. Aby przyspieszyć pierwszą fazę procesu (wytwarzanie zarodków), cienki koniec próbki odkształcamy dodatkowo przez zgniatanie bezpośrednio przed wyżarzaniem. Druty metaliczne Mo, W, Ta, Nb, Fe o średnicach nawet do kilkunastu mm przeciąga się przez piec i dodatkowo podgrzewa się je prądem elektrycznym w celu wytworzenia odpowiedniego gradientu temperatury — me-

metoda Andrade'a



Rys. 7. Schemat urządzenia do otrzymywania cienkich drutów monokrystalicznych przez przemianę w fazie stałej (metodą Andrade'a)

toda Andrade'a (rys. 7). Wymagane odkształcenia powstają podczas przeciągania drutu w przeciągarkach; stosuje się też dodatkowe naciąganie podczas wyżarzania.

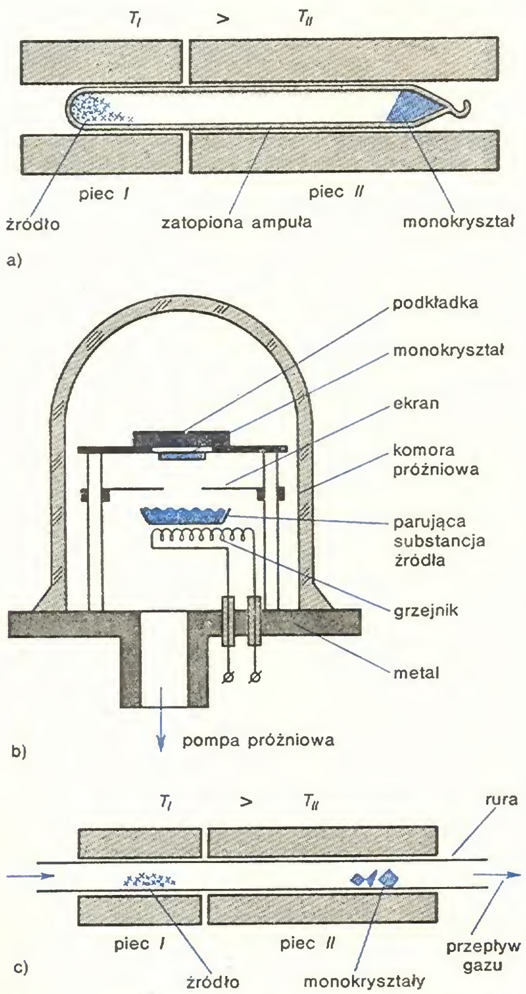
Głównymi wadami metody monokrystalizacji przez przemianę w fazie stałej są trudności z zarodkowaniem i wytworzeniem tylko jednego dobrze wykształconego monokryształu o żądanej orientacji.

Wzrost monokryształów z fazy gazowej

W procesie hodowania monokryształów z fazy gazowej wykorzystuje się zjawisko sublimacji i kondensacji lub pewien typ reakcji chemicznych. Proces może być prowadzony w zamkniętym naczyniu lub w układzie otwartym z przepływem gazu nośnego w zależności od właściwości krystalizowanej substancji.

układ zamknięty

Drobnokrystaliczną substancję, tzw. źródło, z którego mają powstać duże monokryształy, umieszcza się w ampule, zwykle kwarcowej, wytwarza się w niej próżnię lub napełnia się ją gazem obojętnym do odpowiedniego ciśnienia i zatapia (rys. 8a). Ampułę



Rys. 8. Schematy urządzeń do otrzymywania monokryształów z fazy gazowej a) w układzie zamkniętym w zatopionej ampule; b) w układzie zamkniętym w komorze z próżnią dynamiczną; c) w układzie otwartym, w przepływie gazu nośnego

umieszcza się w piecu mającym dwie strefy temperatury (wyższej i niższej) dostosowane do właściwości substancji krystalizowanej. W strefie źródła następuje sublimacja (lub parowanie) substancji, a w strefie monokrystalizacji — kondensacja jej pary. Pomiędzy strefą źródła i strefą monokryształu powstaje różnica ciśnień, która utrzymuje stałe przesycenie. Krystalizacja zachodzi na zarodku lub zarodkach utworzonych w najwęższym miejscu ampuły albo na zarodzi umieszczonej tam przed zatopieniem ampuły. Jeśli ogólne ciśnienie gazu (prężność pary) w ampule w temperaturze procesu jest niewielkie, o szybkości procesu decyduje proces dyfuzji cząsteczek pary (gazu) wewnątrz rury; przy większych ciśnieniach lub w razie napełnienia ampuły gazem obojętnym odgrywają rolę procesy konwekcji. Odmianą układu zamkniętego jest układ z próżnią dynamiczną (rys. 8b) stosowany do uzyskiwania monokrystalicznych warstw metali i półprzewodników na płytkach podłożowych umieszczonych nad silnie ogrzanym źródłem parującym w próżni zwykle rzędu 10^{-4} Pa i lepszej (tabela).

układ z próżnią dynamiczną

Przykłady monokrystalizacji z fazy gazowej

Substancja	Temperatura topnienia, °C	Układ	Temperatura, °C		Gaz nośny lub atmosfera
			sublimacji	kondensacji	
Cynk Zn	419	zamknięty (ampuła) lub otwarty	375–475	350–380	argon, hel
Glin Al	660	zamknięty (komora próżniowa)	1000	600	próżnia
Węglik krzemu SiC	sublimacja ok. 2500	zamknięty (komora reakcyjna)	> 2500	< 2500	gaz obojętny

układ otwarty

Mechanizmy procesu w obu układach zamkniętych i w układzie otwartym (rys. 8c) są analogiczne, jednakże przepływający gaz nośny przenosi znacznie większe ilości pary parującego źródła i szybkość wzrostu w układzie otwartym jest znacznie większa, niż w zamkniętej ampule czy w próżni. Otrzymanie dużych pojedynczych kryształów w metodzie otwartej jest jednak trudne. Znaczne stężenia substancji ułatwiają powstawanie wielu nowych zarodków podczas trwania procesu, a utrudniają wzrost monokryształów na już istniejących zarodkach.

Wykorzystanie reakcji chemicznych

Substancją służącą do budowy monokryształu z fazy gazowej mogą być też produkty reakcji chemicznej. Jedną z metod otrzymywania monokrystalicznego korundu α-Al_2O_3 jest wykorzystanie nieodwracalnej reakcji par chlorku glinu $AlCl_3$ z wodorem i dwutlenkiem węgla. Procesy tego typu prowadzi się zwykle w układzie otwartym.

metoda transportu chemicznego

Inną metodą wykorzystującą reakcje chemiczne do przenoszenia substancji źródła w obszar monokryształu jest sposób zwany metodą transportu chemicznego. Polega ona na dodaniu do układu niewielkiej ilości substancji, zwanej transporterem, która reagując z substancją źródła wytwarza wyłącznie produkty gazowe. W innym miejscu układu, w innej temperaturze, produkty te reagują ze sobą ponownie i wytwarzają w reakcji odwracalnej trwałą fazę sta-

Przykłady monokrystalizacji metodą transportu chemicznego w układzie otwartym i zamkniętym

Substancja	Temperatura topnienia, °C	Równanie reakcji transportu chemicznego	Układ	Temperatura procesu, °C
Tlenek glinu Al_2O_3	2050	$2AlCl_{3(g)} + 3H_{2(g)} + 3CO_{2(g)} \rightarrow Al_2O_{3(s)} + 3CO_{(g)} + 3HCl_{(g)}$	otwarty	1700–1800
Żelazo Fe	1535	$Fe_{(s)} + 2HCl_{(g)} \rightleftarrows FeCl_{2(g)} + H_{2(g)}$	zamknięty	źródła 1000, monokryształu 730
Arsenek galu GaAs	1238 (rozkład)	$2GaAs_{(s)} + 3I_{2(g)} \rightleftarrows 2GaI_{3(g)} + As_{2(g)}$	zamknięty	źródła 1100, monokryształu 900
Krzem Si	1410	$SiCl_{4(g)} + 2H_{2(g)} \rightarrow Si_{(s)} + 4HCl_{(g)}$	otwarty	1100–1250

łą — monokryształ substancji źródła. Uwalnia się przy tym substancja–transporter przenoszona przez konwekcję lub dyfuzję do obszaru źródła — jeśli proces odbywa się w układzie zamkniętym — lub unoszona przez gaz nośny poza układ otwarty. Metody transportu chemicznego znalazły zastosowanie do otrzymywania monokryształów pierwiastków i związków o niskich prężnościach par lub do związków bardzo łatwo rozkładających się po ogrzaniu, których utrzymanie w układzie zamkniętym jednoskładnikowym (bez reakcji chemicznej) jest niemożliwe lub zbyt powolne (tabela).

Reakcje transportu chemicznego, odwracalne i nieodwracalne, znalazły zastosowanie zarówno do wytwarzania dużych monokryształów, jak i tzw. warstw epitaksjalnych. Są to cienkie, o grubości kilku do kilkudziesięciu mikrometrów, monokrystaliczne warstwy „narośnięte" na płaskim podłożu monokrystalicznym będącym zarodzią. Warstwy te powtarzając strukturę podłoża-zarodzi mogą mieć nieco zmieniany skład chemiczny przez wprowadzanie do rosnącej warstwy niewielkiej ilości domieszki. Na granicy podłoże–warstwa uzyskuje się w ten sposób zamierzony skok właściwości substancji monokryształu. Metoda epitaksjalnej monokrystalizacji krzemu stała się podstawą rozwoju technologii krzemowych układów scalonych stosowanych powszechnie w kalkulatorach, komputerach, coraz powszechniej w odbiornikach radiowych i telewizyjnych, a także w układach regulacji i sterowania automatycznego procesów przemysłowych (→ Mikroelektronika).

warstwy epitaksjalne

W. D. LAWSON i S. NIELSEN *Otrzymywanie monokryształów*, Warszawa 1962; *The Art and Science of Growing Crystals*, J. J. GILMAN (ed.), London 1963.

Dyslokacje w kryształach

Tadeusz Figielski

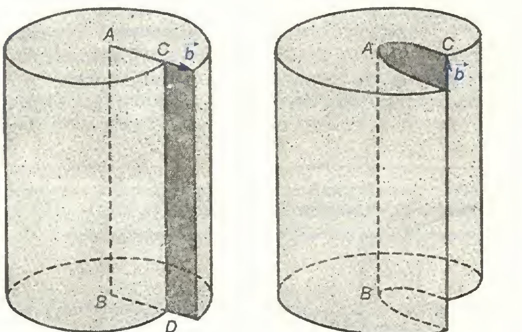
Dyslokacje są liniowymi defektami struktury krystalicznej. Wpływają one w zasadniczy sposób na własności mechaniczne ciał stałych, również ważną rolę odgrywają w materiałach półprzewodnikowych stosowanych w elektronice.

Zdumiewa nas fakt, że jeszcze niedawno — przed wysunięciem koncepcji dyslokacji, gdy już została rozszyfrowana kwantowa struktura atomu — fizycy nie rozumieli, dlaczego sztabka czystego metalu jest plastyczna zaraz po jej wytopieniu, a staje się twarda i wytrzymała po kuciu lub walcowaniu. Wyrażając się ściślej, nie rozumiano, dlaczego wytrzymałość kryształu na ścinanie jest o wiele rzędów niższa niż ta, jaką przewidywała najprostsza teoria plastyczności, w której zakładano, że całe płaszczyzny atomowe kryształu ślizgają się jedne po drugich. Rozwiązanie tej zagadki dali trzej badacze G. J. Taylor, E. Orowan i M. Polanyi, którzy niezależnie od siebie zaproponowali w 1934 r. koncepcję dyslokacji w krysztale.

wektor Burgersa

Rozważmy na początku pojęcie dyslokacji w ośrodku ciągłym. Wyobraźmy sobie doskonale sprężysty materiał w postaci walca, w którym wykonano nacięcie wzdłuż płaszczyzny *ABCD* aż do jego osi (rys. 1). Przemieśćmy następnie dwie przylegające części walca wzdłuż płaszczyzny przecięcia o wektor $\vec{b}$. Ten rodzaj zakłócenia, rozciągający się wzdłuż linii *AB* (osi), nazywa się dyslokacją, a wektor $\vec{b}$ — wektorem Burgersa dyslokacji. W sieci krystalicznej wektor Burgersa może przyjmować jedynie pewne wartości dyskretne; dyslokacja z najkrótszym możliwym wektorem Burgersa nazywa się dyslokacją jednostkową.

dyslokacja jednostkowa



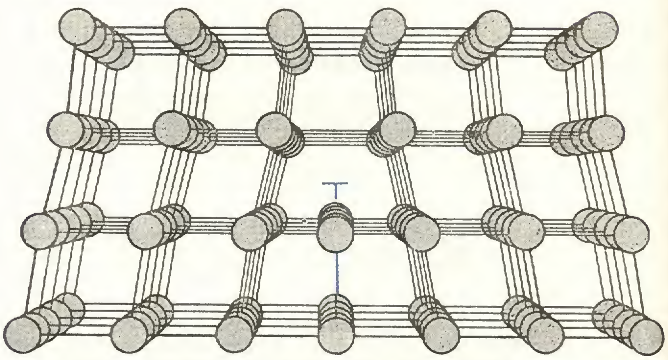
Rys. 1. Dyslokacja w ośrodku ciągłym: a) krawędziowa, b) śrubowa

Rozróżnia się dwa zasadnicze typy dyslokacji: dyslokację krawędziową (rys. 1a), w której wektor Burgersa jest prostopadły do linii (osi) dyslokacji oraz dyslokację śrubową (rys. 1b) — gdy wektor Burgersa jest równoległy do osi. Ogólnie dyslokacja może mieć zarówno składową krawędziową, jak i śrubową, czyli może być typu mieszanego. Przykładem takiej dyslokacji jest dyslokacja sześćdziesięciostopniowa występująca w kryształach o strukturze diamentu (rys. 12; 60° jest kątem między wektorem Burgersa i osią dyslokacji).

dyslokacja krawędziowa i śrubowa

Istnieje ogólne prawo ułatwiające rozpatrywanie geometrycznych zagadnień związanych z dyslokacjami: wektor Burgersa musi być zachowany wzdłuż całej linii dyslokacyjnej. Z prawa tego wynika, że dyslokacja nie urywa się wewnątrz kryształu; może ona kończyć się jedynie na jego powierzchniach lub tworzyć zamknięte pętle. (il. 80, tabl. 21).

Dyslokacja krawędziowa z jednostkowym wektorem Burgersa w najprostszej sieci krystalicznej jest równoważna umieszczeniu dodatkowej płaszczyzny atomowej między dwiema sąsiadującymi płaszczyznami krystalicznymi (rys. 2). Można zauważyć, że każdy atom na krawędzi tej płaszczyzny (będącej



Rys. 2. Dyslokacja krawędziowa w najprostszej strukturze krystalicznej

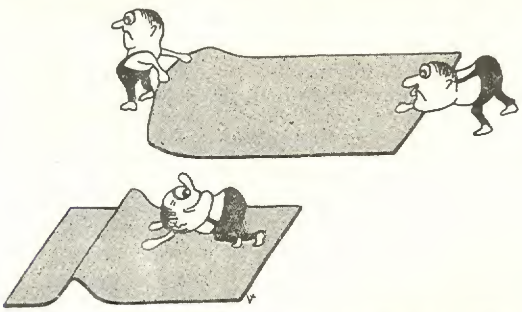
linią dyslokacji) ma mniej sąsiadów niż atom w sieci doskonałej. A zatem w kryształach o wiązaniu kowalencyjnym jedno wiązanie w każdym atomie krawędziowym jest wolne; jest ono nazywane wiązaniem wiszącym i odgrywa bardzo ważną rolę w elektronowych własnościach dyslokacji.

wiązanie wiszące

Rola pojęcia dyslokacji w teorii plastyczności jest oczywista. Zamiast jednoczesnego przemieszczania podczas deformacji kryształu całych płaszczyzn ato-

ruch poślizgowy dyslokacji

mowych, wystarczy przemieszczanie kolejnych rzędów atomów znajdujących się na linii dyslokacyjnej. Ten rodzaj ruchu jest dobrze znany z praktyki domowej np. ułatwia przesuwanie ciężkiego dywanu (rys. 3).

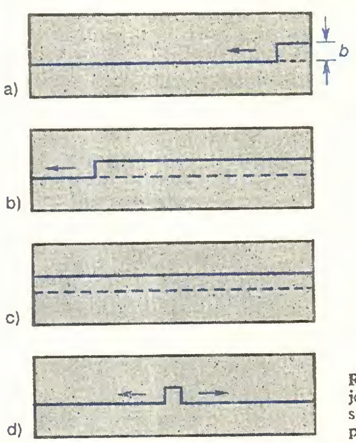


Rys. 3. Mechanizm dyslokacji może być użyteczny przy przesuwaniu ciężkiego dywanu

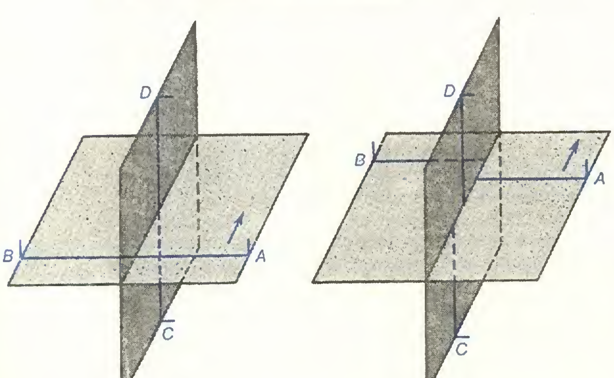
W krysztale taki proces może być opisany przez ruch dyslokacji; jest to ruch poślizgowy dyslokacji zachodzący w płaszczyźnie poślizgu. Uprzywilejowanymi płaszczyznami poślizgu są płaszczyzny krystalograficzne o najgęstszym upakowaniu atomów.

ruch przegięcia

W kryształach o wiązaniu kowalencyjnym naprężenie niezbędne do wywołania poślizgu całej linii dyslokacyjnej jest stosunkowo duże i wówczas bardziej uprzywilejowanym procesem może być ruch dyslokacji zachodzący stopniowo wzdłuż krótkiego jej odcinka. Taki proces jest możliwy, gdy dyslokacja zawiera przegięcie leżące w tej samej płaszczyźnie poślizgu i o długości równej zazwyczaj jednostkowemu wektorowi Burgersa (rys. 4). Przegięcie może ślizgać się wzdłuż dyslokacji, a jego przemieszczenie się z jednego końca dyslokacji do drugiego jest równoważne przesunięciu całej dyslokacji o jednostkowy wektor Burgersa. Podwójne przegięcia mogą być wzbudzane spontanicznie przez drgania cieplne sieci (rys. 4d), jak również mogą być wytwarzane w wyniku przecięcia się dwóch dyslokacji leżących w różnych płaszczyznach poślizgu (rys. 5).



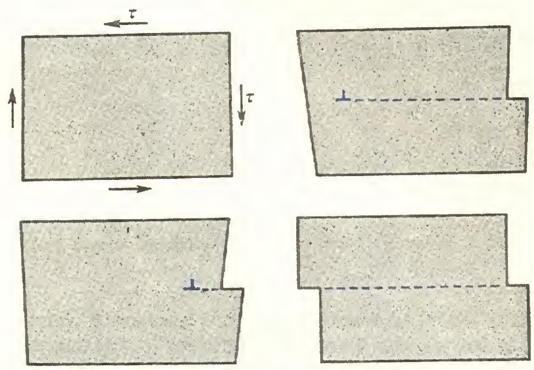
Rys. 4. Przegięcie i jego ruch wzdłuż dyslokacji (a, b, c), d) podwójne przegięcie



Rys. 5. Przecięcie się dwóch dyslokacji krawędziowych, prowadzące do wytworzenia przegięć

„kwant" deformacji plastycznej

Pod wpływem działania na kryształ stałego zewnętrznego naprężenia τ dyslokacja wykonuje ruch poślizgowy i w końcu opuszcza kryształ pozostawiając po sobie uskok — „kwant" plastycznej deformacji (rys. 6). Taki prosty model prowadzi oczywiście do wystąpienia pewnej górnej granicy stopnia deformacji plastycznej kryształu, która byłaby określona gęstością

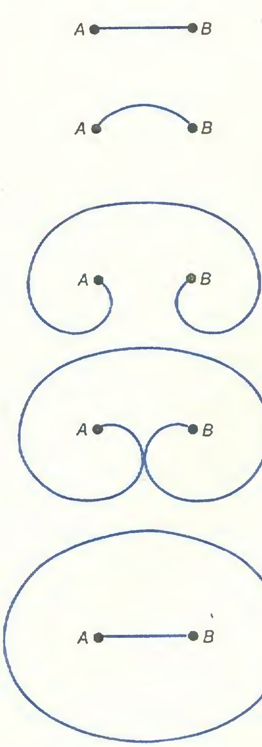


Rys. 6. Ruch poślizgowy dyslokacji jest równoważny deformacji plastycznej kryształu

dyslokacji początkowo istniejących w krysztale. Ponieważ doświadczalnie nie obserwuje się takiej granicy, należy wnioskować, że podczas plastycznego „płynięcia" kryształu działają pewne mechanizmy wytwarzania (generacji) lub powielania dyslokacji

rozmnażanie dyslokacji

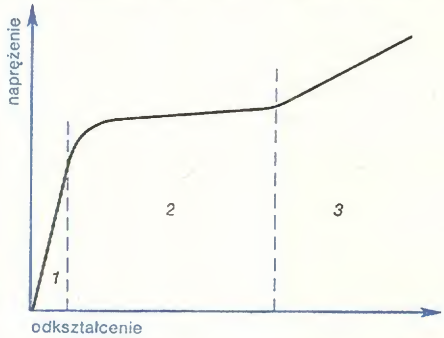
Jednym z bardziej znanych jest mechanizm generacji Franka-Reada, który zachodzi wówczas, gdy w płaszczyźnie poślizgu występują przeszkody w ruchu dyslokacji — tzw. punkty zaczepienia. Działanie tego typu mechanizmu przedstawia rys. 7.



Rys. 7. Mechanizm Franka-Reada generacji dyslokacji. Odcinek linii dyslokacyjnej, unieruchomiony między punktami *A* i *B* leżącymi w płaszczyźnie poślizgu, wygina się pod wpływem przyłożonego naprężenia, a następnie zamyka się wytwarzając pętle dyslokacyjne

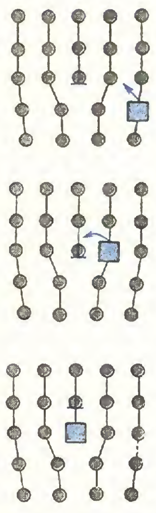
Podatność kryształu na działanie sił mechanicznych można zobrazować za pomocą wykresu, na którym na osi rzędnych odłożone jest działające naprężenie (np. siła rozciągająca na jednostkę przekroju poprzecznego kryształu), natomiast na osi odciętych — odkształcenie (np. względna zmiana długości kryształu). Przykładowa zależność tego rodzaju przedstawiona jest na rys. 8, gdzie można wyróżnić

wspinanie się dyslokacji



Rys. 8. Zależność naprężenia od odkształcenia monokryształu metalu

trzy charakterystyczne obszary. W obszarze *1* kryształ zachowuje się doskonale sprężyście. Odkształcenie jest tu proporcjonalne do naprężenia i jest całkowicie odwracalne. W obszarze *2* kryształ zaczyna plastycznie (nieodwracalnie) „płynąć". Małe przyrosty naprężenia wywołują duże zmiany długości kryształu. Jest to obszar, w którym zachodzi intensywny ruch poślizgowy i rozmnażanie dyslokacji. Przy dalszej deformacji osiąga się trzeci obszar, w którym następuje utwardzanie kryształu. Liczba dyslokacji leżących w różnych płaszczyznach poślizgu jest wówczas tak duża, że blokują one sobie wzajemnie możliwości ruchu, utrudniając deformację plastyczną.



Rys. 9. Wspinanie się dyslokacji wywołane dopływem luk

Ruch poślizgowy dyslokacji jest ruchem zachowawczym, co oznacza, że nie towarzyszy mu transport masy. W wyższych temperaturach może stać się istotny niezachowawczy rodzaj ruchu dyslokacji, który jest wywołany dopływem lub odpływem luk lub atomów międzywęzłowych; ruch taki nazywany jest wspinaniem się dyslokacji (rys. 9).

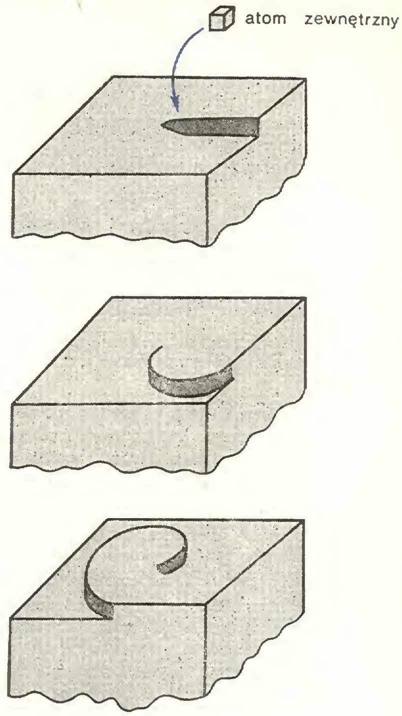
Przez długi czas dyslokacje były koncepcją czysto teoretyczną, zanim nauczono się je uwidaczniać w krysztale. Pierwszą i najprostszą metodą było użycie specjalnych środków trawiących, atakujących najbardziej obszary kryształu, w których dyslokacje wychodzą na powierzchnię. Ilustracja 79 (tabl. 21) przedstawia jedną z pierwszych identyfikacji jamek trawienia na germanie z pojedynczymi dyslokacjami. Dyslokacje są tu uszeregowane liniowo tworząc tzw. małokątową granicę ziaren (il. 82, tabl. 21). Obecnie możemy oglądać dyslokacje wewnątrz kryształu stosując topografię rentgenowską lub — bardziej szczegółowo — wykorzystując elektronową mikroskopię prześwietleniową (→ Mikroskopia elektronowa i il. 87, tabl. 22). Za pomocą tej metody można również obserwować dyslokacje podczas ich ruchu. Badania elektronomikroskopowe wykazują, że dyslokacje bywają często rozszczepione na dyslokacje częściowe z ułamkowym wektorem Burgersa. Płaski defekt rozciągający się między rozszczepioną parą nazywa się błędem ułożenia.

obserwacje dyslokacji

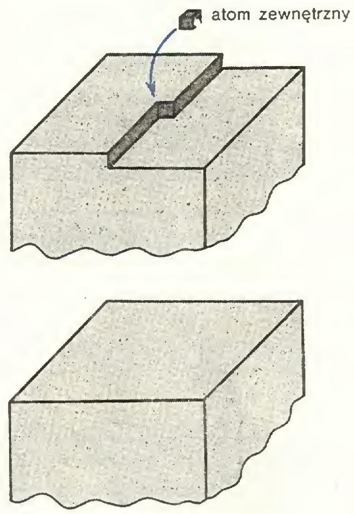
Wokół linii dyslokacyjnej istnieje złożone pole naprężeń mechanicznych. W pobliżu dyslokacji krawędziowej deformacja sieci ma charakter dylatacji bądź kompresji w zależności od położenia obszaru w stosunku do płaszczyzny poślizgu dyslokacji. To pole naprężeń jest przyczyną silnego oddziaływania dyslokacji z defektami punktowymi i atomami domieszkowymi w krysztale. W szczególności domieszki mogą działać jako punkty zaczepienia dyslokacji, hamując ich ruch poślizgowy.

wzrost kryształów

Dyslokacje odgrywają istotną rolę w mechanizmie wzrostu kryształów. Na rys. 10 pokazana jest dyslokacja śrubowa, której wyjście na powierzchnię stanowi niewysycający się zarodek szybkiego wzrostu kryształu. Jego działanie polega na tym, że zewnętrzne atomy łatwo wbudowują się w kryształ w miejscach, gdzie mogą się one wiązać z więcej niż jednym atomem. Gdy na powierzchni kryształu istnieje prosty stopień (rys. 11), mechanizm szybkiego wzrostu przestaje działać po skompletowaniu pełnej warstwy atomowej.



Rys. 10. Dyslokacja śrubowa i spiralny wzrost kryształu



Rys. 11. Schodek na krysztale nie podtrzymuje stałego szybkiego wzrostu kryształu

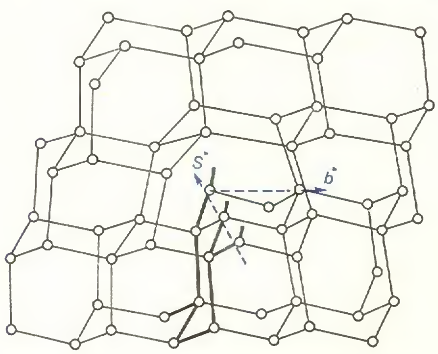
Energia związana z naprężeniami wokół dyslokacji jest tak duża, że dyslokacja, w przeciwieństwie do defektów punktowych (luk i atomów międzywęzłowych), nie może powstać samoistnie w krysztale w wyniku drgań cieplnych atomów; dyslokacja nie jest więc defektem termodynamicznie równowagowym. Wynika stąd, że dyslokacje mogą być w zasadzie całkowicie usunięte z kryształu. Obecnie tzw. kryształy bezdyslokacyjne półprzewodników germanu i krzemu produkowane są w skali przemysłowej.

kryształy bezdyslokacyjne

wpływ dyslokacji na domieszkowanie

W materiałach półprzewodnikowych dyslokacje wpływają w sposób istotny zarówno na procesy atomowe jak i elektronowe. Wpływ ten jest w zasadzie ujemny. Wzdłuż dyslokacji zachodzi wzmożona dyfuzja domieszek psująca ich rozkład przestrzenny w przyrządach półprzewodnikowych. Dyslokacje w sposób drastyczny obniżają czas życia mniejszościowych nośników prądu, działając jako efektywne centra rekombinacji (→ Fizyka przyrządów półprzewodnikowych). Wreszcie dyslokacje zapoczątkowują lawinowe przebicie w złączach *n-p*, obniżając tym samym wartość maksymalnego napięcia zaporowego, przy którym mogą pracować diody półprzewodnikowe. Jest zatem sprawą oczywistą, że technologowie czynią wielki wysiłek, ażeby uniknąć dyslokacji w materiałach i strukturach półprzewodnikowych.

Fizyka zjawisk elektronowych związanych z dyslokacjami w półprzewodnikach jest bardzo interesująca. Na krawędzi dodatkowej półpłaszczyzny atomowej, jaką stanowi dyslokacja, zlokalizowane są niewysycone, „wiszące" wiązania chemiczne. W wypadku półprzewodników atomowych, germanu i krzemu, są to ściśle ukierunkowane wiązania kowalencyjne (rys. 12). Sterczą one prostopadle do osi dyslokacji i są



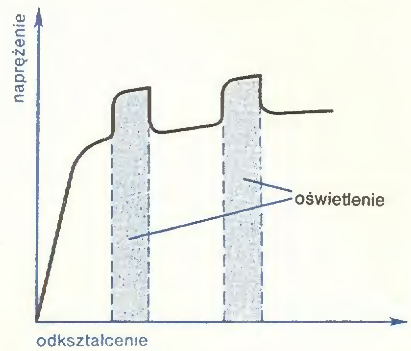
Rys. 12. Dyslokacja sześćdziesięciostopniowa w krysztale o strukturze diamentu

bariera potencjału wokół dyslokacji

odległe od siebie jedynie o kilka Å. Dodatkowy elektron może być wychwycony z pasma przewodnictwa półprzewodnika na wiszące wiązanie, lub też wiszący elektron może być przeniesiony do pasma. W pierwszym wypadku dyslokacja staje się naładowana ujemnie, w drugim — dodatnio. Proces ładowania dyslokacji zachodzi spontanicznie aż do momentu osiągnięcia równowagi termodynamicznej układu. Znak i wielkość ładunku na dyslokacji zależy od energii Fermiego w krysztale. Ten ładunek powoduje wytworzenie potencjału elektrostatycznego na dyslokacji. Różnica potencjałów powstająca między linią dyslokacyjną a otaczającym ją materiałem odgrywa zasadniczą rolę w procesach rekombinacji nośników prądu w półprzewodnikach. Od niej zależy dopływ nośników do dyslokacji będących bardzo efektywnymi centrami rekombinacji.

fotoplastyczność

Oświetlenie półprzewodnika światłem generującym dodatkowe nośniki prądu zmniejsza ładunek dyslokacji. Efekt ten jest prawdopodobnie przyczyną obserwowanego w związkach półprzewodnikowych zjawiska fotoplastyczności. Zjawisko to polega na



Rys. 13. Zjawisko fotoplastyczne w krysztale tellurku kadmu (CdTe). Naprężenie niezbędne do wywołania określonej deformacji plastycznej wzrasta przy oświetleniu kryształu

zmianie efektywnej ruchliwości dyslokacji pod wpływem oświetlenia, co powoduje zmianę twardości kryształu (rys. 13).

dyslokacje — centra paramagnetyczne

Wiszące elektrony wzdłuż dyslokacji krawędziowej mają niesparowane spiny i związany z nimi moment magnetyczny. Dyslokacje są zatem centrami paramagnetycznymi w krysztale. Moment magnetyczny pochodzący od dyslokacji został rzeczywiście wykryty doświadczalnie przy badaniu elektronowego rezonansu paramagnetycznego w krzemie.

J. Weertman, J. R. Weertman *Podstawy teorii dyslokacji*, Warszawa 1969.

Badanie struktury kryształów

Teoria wewnętrznej budowy kryształów, obejmująca teorię sieciową A. Bravais'go i teorię grup przestrzennych J. S. Fiodorowa, A. Schoenfliesa i W. Barlowa powstała w XIX w.

Słuszność tych teorii potwierdzono doświadczalnie dopiero w r. 1912, gdy przy próbach wyjaśnienia natury (fale czy cząstki?) promieniowania rentgenowskiego M. von Laue, W. Friedrich i P. Knipping wykazali, że promienie rentgenowskie ulegają dyfrakcji na sieciach przestrzennych kryształów. Rok 1912 stał się rokiem przełomowym dla rozwoju krystalografii, gdyż odkrycie Lauego, Friedricha i Knippinga dało początek nowym jej gałęziom, związanym z badaniami rodzajów sieci przestrzennych występujących w ciałach krystalicznych i badaniem rozmieszczeń atomów, jonów lub cząsteczek w komórkach elementarnych: krystalografii rentgenowskiej i analizie strukturalnej kryształów.

Dwanaście lat później, L. de Broglie sformułował hipotezę o falowym charakterze ruchu cząstek. Doświadczalnym potwierdzeniem tej hipotezy było pokazanie w 1927 r., przez C. I. Davissona i L. H. Germera, że elektrony ulegają dyfrakcji na sieciach przestrzennych ciał krystalicznych, podobnie jak promienie rentgenowskie. Pomyślnie zakończone doświadczenia nad dyfrakcją neutronów przeprowadzili w 1936 r. P. N. Mitchell i D. P. Powells. Te dwa doświadczenia dały początek nowym metodom badania struktury kryształów: elektronografii i neutronografii, a pierwsze z nich doprowadziło ponadto do skonstruowania w 1931 r. przez M. Knolla i E. Ruska, mikroskopu elektronowego.

Krystalografia rentgenowska

Zygmunt Trzaska Durski

Eksperyment, który dał początek krystalografii rentgenowskiej, przeprowadzony w 1912 r. przez Lauego, Friedricha i Knippinga był niesłychanie prosty, zwła-

szcza jeśli weźmie się pod uwagę używane obecnie, nieraz bardzo skomplikowane, przyrządy do badania dyfrakcji promieni rentgenowskich na kryształach. Po prostu — na drodze wąskiej wiązki promieni rentgenowskich ustawione zostały kryształ i klisza fotograficzna. Po naświetleniu i wywołaniu kliszy okazało się, że oprócz śladu padającej wiązki promieniowania na kliszy pojawiły się ślady wiązek odchylonych od kierunku wiązki padającej (tj. ślady wiązek ugiętych).

Odbicie promieni rentgenowskich od płaszczyzn sieciowych kryształu

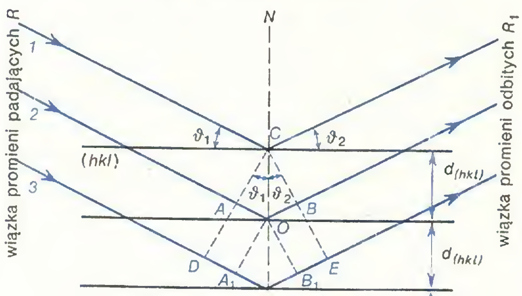
Zastanówmy się nad tym, skąd w doświadczeniu Lauego wzięły się wspomniane wiązki ugięte. Otóż wiązki promieni odchylone od kierunku biegnącej z lampy wiązki padającej powstają w wyniku swoistego odbicia promieni rentgenowskich przez zespoły równoległych do siebie płaszczyzn sieciowych kryształu. Odbicie to ma szczególny charakter i nazywa się odbiciem intereferencyjnym, ponieważ w istocie przyczyną jego jest wtórne promieniowanie atomów (a właściwie elektronów), uporządkowanych w sieć krystaliczną, na które padają promienie rentgenowskie. Mianowicie, każdy z elektronów atomów sieci krystalicznej pod działaniem promieniowania rentgenowskiego wykonuje drgania w takt zmian pola elektromagnetycznego tego promieniowania, stając się w ten sposób źródłem promieniowania wtórnego, o takiej samej długości fali jak promieniowanie wzbudzające, lecz rozchodzącego się w przestrzeni kuliście. Te fale wtórne interferują ze sobą i w pewnych ściśle określonych dla danej sieci przestrzennej i danej długości fali promieniowania kierunkach, ulegają wzmocnieniu, co prowadzi do powstawania tzw. wzmocnionych promieni interferencyjnych, obserwowanych w doświadczeniach jako promienie ugięte czy też promienie odbite.

odbicie interferencyjne

Kierunek rozchodzenia się wzmocnionych promieni interferencyjnych (promieni interferencyjnie odbitych) można określić za pomocą tzw. równania Wulfa–Bragga. Według tego równania interferencyjne odbicie promieni rentgenowskich o długości fali λ od płaszczyzny sieciowej (hkl) (rys. 1) następuje tylko w takim kierunku, określonym tzw. kątem odbłysku $\vartheta_{(hkl)}$, przy którym różnica dróg $\Delta S = 2d_{(hkl)} \sin \vartheta_{(hkl)}$ przebytych przez promienie rentgenowskie odbite od dwóch sąsiednich równoległych płaszczyzn sieciowych jest równa całkowitej wielokrotności długości fali $\Delta S = n\lambda$. Różnica dróg ΔS jest sumą odcinków AO i OB, o które promień *2* ma dłuższą drogę niż promień *1*. Mówiąc inaczej — interferencyjne odbicie nastąpi wtedy, gdy na drodze AOB zmieści się całkowita liczba długości fal, czyli gdy

równanie Wulfa–Bragga

$$n\lambda = 2d_{(hkl)} \sin \vartheta_{(hkl)}.$$



Rys. 1. Ilustracja równania Wulfa–Bragga; promienie rentgenowskie R padają na rodzinę płaszczyzn sieciowych (hkl) pod kątem połysku ϑ_1 i zostają odbite (R_1) pod kątem odbłysku ϑ_2 $(\vartheta_1 = \vartheta_2)$; ($N$ — prosta prostopadła do płaszczyzny (hkl); N, R oraz R_1 leżą w jednej płaszczyźnie)

W równaniu tym liczba $n = 1, 2, 3 \ldots$ określa tzw. rząd odbicia, czyli liczbę długości fali mieszczących się na drodze ABC (rys. 5). Z równania Wulfa–Bragga wynika, że odbicie promieni rentgenowskich przez płaszczyzny sieciowe kryształu ma charakter selektywny, tzn. występuje tylko pod pewnymi kątami ϑ, ściśle określonymi dla danej sieci krystalicznej i dla danej długości fali promieniowania.

rząd odbicia

Każdą wiązkę promieni rentgenowskich odbitą przez płaszczyzny sieciowe można scharakteryzować jej kierunkiem rozchodzenia się i jej intensywnością. O geometrycznym rozkładzie w przestrzeni kierunków odbitych wiązek promieni decydują symetria i rozmiary komórki elementarnej. Liczba, rodzaje i wzajemna konfiguracja atomów znajdujących się w komórce elementarnej wpływają wyłącznie na intensywność wiązek odbitych. W związku z tym całość badań struktury kryształów za pomocą promieni rentgenowskich dzieli się na krystalografię rentgenowską i analizę strukturalną kryształów.

Krystalografia rentgenowska zajmuje się badaniem geometrycznego rozkładu w przestrzeni kierunków wiązek promieni rentgenowskich odbitych od płaszczyzn sieciowych kryształów; na podstawie tego rozkładu wyznacza się symetrię i stałe sieciowe sieci przestrzennych. Badanie intensywności wiązek odbitych, prowadzące do wyznaczania położeń (współrzędnych) atomów w komórce elementarnej, jest podstawą analizy strukturalnej kryształów (zob. rozdział „Strukturalna analiza kryształów").

analiza strukturalna kryształów

Do zakresu krystalografii rentgenowskiej należą takie badania jak: określenie układu krystalograficznego, klasy dyfrakcyjnej, grupy przestrzennej (lub dyfrakcyjnej) kryształu, wyznaczenie stałych sieciowych i objętości komórki elementarnej, wyznaczenie liczby atomów (cząsteczek) znajdujących się w komórce elementarnej.

Doświadczalne metody krystalografii rentgenowskiej

W celu wykonania sformułowanych wyżej zadań oraz przygotowania danych wyjściowych do strukturalnej analizy kryształów stosowane są różne metody doświadczalne. W metodach tych tzw. obrazy dyfrakcyjne monokryształów lub ciał polikrystalicznych, uzyskane przy użyciu promieniowania rentgenowskiego zawierającego fale różnych długości (promieniowanie polichromatyczne) lub promieniowania o jednej długości fali (promieniowanie monochromatyczne), rejestrowane są na błonie fotograficznej bądź też za pomocą liczników Geigera–Müllera, proporcjonalnych lub scyntylacyjnych.

Przyrządy umożliwiające rejestrowanie obrazów dyfrakcyjnych na błonie fotograficznej nazywają się kamerami rentgenowskimi. W użyciu są różne typy kamer. Istnieją kamery uniwersalne, pozwalające na wykonywanie rentgenogramów różnymi metodami, oraz kamery przeznaczone do wykonywania rentgenogramów tylko jedną metodą. Istnieją kamery służące do badania monokryształów oraz kamery do badania ciał polikrystalicznych, przy czym zarówno jedne, jak i drugie konstruowane są z przeznaczeniem do badań w warunkach normalnej temperatury i ciśnienia oraz do badań w warunkach specjalnych, np. w wysokiej i niskiej temperaturze, pod wysokim ciśnieniem, w atmosferze ochronnej (w atmosferze helu, azotu) itp.

kamery rentgenowskie

Aparaty, w których położenie i intensywność wiązek odbitych rejestruje się za pomocą liczników Geigera–Müllera lub liczników scyntylacyjnych, nazywa się dyfraktometrami rentgenowskimi (il. 77, 78, tabl. 20).

dyfraktometry rentgenowskie

Uzyskiwane różnymi metodami i zarejestrowane na błonie fotograficznej lub za pomocą liczników obrazy dyfrakcyjne ciał krystalicznych nazywa się rentgenogramami lub dyfraktogramami. Zaczernione

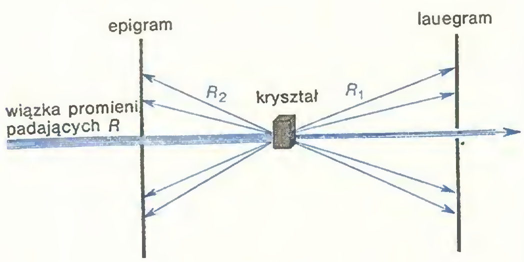
rentgenogramy (dyfraktogramy)

plamki na rentgenogramach, będące śladami wiązek promieni rentgenowskich odbitych od płaszczyzn sieciowych, nazywa się zwykle refleksami. **refleksy**

Przystępując do badania struktury kryształu, przede wszystkim należy ustalić kształt (tj. symetrię), typ i wymiary komórki elementarnej, a także liczbę atomów czy cząsteczek, które w tej komórce należy rozmieścić. Jeżeli na podstawie zewnętrznego wyglądu kryształu nie można wyciągnąć wniosków o jego symetrii i przynależności do układu krystalograficznego, wówczas wykorzystuje się zjawisko dyfrakcji promieni rentgenowskich, stosując tzw. metodę Lauego.

metoda Lauego

Metodą Lauego bada się monokryształ grubości 0,05–0,3 mm. Badany monokryształ przykleja się (np. woskiem, plasteliną) do tzw. główki goniometrycznej (il. 75, tabl. 20) będącej urządzeniem umożliwiającym zmianę nachylenia monokryształu względem padającej wiązki promieni oraz przesuwanie go w kamerze w kierunkach do siebie prostopadłych. Na umieszczony w kamerze Lauego (il. 73, tabl. 20) nieruchomy monokryształ pada wiązka polichromatycznych, równoległych promieni rentgenowskich, wydzielona za pomocą tzw. kolimatora z szerokiej, rozbieżnej wiązki promieni rozchodzących się od anody lampy rentgenowskiej. Wiązka padająca po przejściu przez kolimator ma średnicę ok. 0,5–1,0 mm (rys. 2).



Rys. 2. Metoda Lauego; schemat otrzymywania rentgenogramów za pomocą promieni przechodzących R_1 i promieni wstecznych R_2

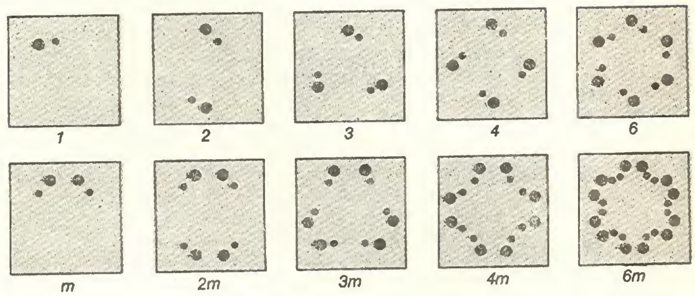
Otrzymany obraz dyfrakcyjny kryształu rejestruje się zwykle na płaskiej błonie fotograficznej ustawionej za monokryształem, prostopadle do kierunku wiązki padającej. Przy takim ustawieniu błony fotograficznej mamy do czynienia z metodą promieni przechodzących. Otrzymane tą metodą rentgenogramy nazywa się lauegramami (il. 68, tabl. 18). **lauegram**

W metodzie Lauego błonę fotograficzną można ustawić również między monokryształem a lampą rentgenowską. Jest to wtedy metoda promieni wstecznych, którą można wykonywać rentgenogramy (tzw. epigramy) monokryształów zupełnie nieprzenikliwych dla stosowanej wiązki promieni, np. monokryształów bardzo dużych lub bardzo silnie pochłaniających promieniowanie. **epigram**

Metoda Lauego służy głównie do badania symetrii makroskopowej monokryształów. Jest to możliwe dlatego, że gdy w monokrysztale promienie biegną równolegle do dwu-, trój-, cztero- czy sześciokrotnej osi symetrii lub równolegle do płaszczyzny symetrii, to ślady wiązek odbitych od ułożonych zgodnie z symetrią płaszczyzn sieciowych będą na rentgenogramie odpowiednio symetrycznie rozmieszczone wokół śladu wiązki padającej. Jeżeli wiązka padająca przechodzi przez kryształ równolegle do n-krotnej osi symetrii, wzdłuż której przecina się n płaszczyzn symetrii, to na rentgenogramie wokół plamki pierwotnej wystąpi symetryczne ułożenie refleksów równocześnie względem n-krotnej osi symetrii i n płaszczyzn symetrii.

obrazy Lauego

Istnieje 10 różnych typów obrazów Lauego, tzn. różnych symetrycznych rozmieszczeń refleksów na rentgenogramach. Obrazy te mają symetrię: 1, 2, 3, 4, 6, m, $2m$, $3m$, $4m$, $6m$ (rys. 3). Obraz oznaczony symbolem 1 jest obrazem asymetrycznym i powstaje zawsze wtedy, gdy w monokrysztale, równolegle do wiązki padającej, nie znajduje się żaden z wymienionych wyżej elementów symetrii.



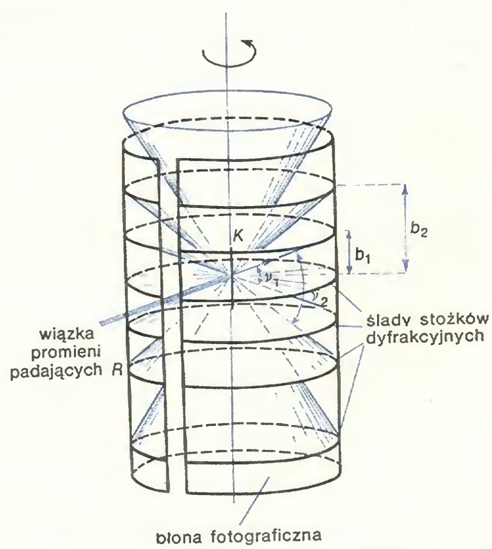
Rys. 3. 10 typów symetrii lauegramów

Na podstawie rentgenogramów wykonanych metodą Lauego nie można jednak stwierdzić obecności w monokrysztale środka symetrii. W związku z tym, stosując metodę Lauego, nie można danego monokryształu zaszeregować do jednej z 32 klas krystalograficznych (→ Budowa kryształów), tylko do jednej z 11 klas dyfrakcyjnych (klas Lauego), obejmujących po kilka klas krystalograficznych (różniących się między sobą jedynie obecnością lub brakiem środka symetrii). Jedną klasę Lauego o symbolu $4/mmm$ tworzą np. cztery klasy tetragonalne: $4/mmm$, 422, $4mm$, $\bar{4}2m$. Klasy dyfrakcyjne oznaczone są symbolami: $\bar{1}$, $2/m$, mmm, $4/m$, $4/mmm$, $\bar{3}$, $\bar{3}m$, $6/m$, $6/mmm$, $m3$, $m3m$. **klasy Lauego**

Opisana niżej metoda obracanego kryształu i jej różne odmiany wymagają obracania badanego monokryształu wokół pewnej prostej sieciowej (np. osi krystalograficznej X, Y lub Z). Odpowiednie ustawienie monokryształu przeprowadza się stosując metodę Lauego. Po wyznaczeniu w monokrysztale kierunków osi krystalograficznych można już stosunkowo łatwo wykonać pomiary długości krawędzi komórki elementarnej za pomocą metody obracanego kryształu.

metoda obracanego kryształu

W metodzie obracanego kryształu wąska wiązka (o średnicy 0,5–1,0 mm) równoległych monochromatycznych promieni rentgenowskich pada na monokryształ, któremu zwykle nadaje się kształt walca (o średnicy 0,1–0,3 mm i długości kilku mm) lub kulki



Rys. 4a. Schemat metody obracanego kryształu; K — prosta sieciowa kryształu, wokół której kryształ jest obracany

(mającej średnicę kilku dziesiątych mm). Główkę goniometryczną z przyklejonym monokryształem umieszcza się w odpowiedniej kamerze. Monokryształ podczas wykonywania rentgenogramu jest obracany wokół jednej z osi krystalograficznych lub wokół innej prostej sieciowej o niskich wskaźnikach. Oś obrotu monokryształu jest prostopadła do wiązki promieni padających. W tych warunkach wzmocnione promienie interferencyjne układają się na powierzchniach stożków kołowych (rys. 4a), zwanych stożkami dyfrakcyjnymi, o wspólnej osi będącej prostą sieciową, wokół której obracany jest monokryształ. Monokryształ obracany jest w tym celu, aby coraz to inne jego płaszczyzny sieciowe mogły się znaleźć w pozycji odbijającej promienie rentgenowskie.

stożki dyfrakcyjne

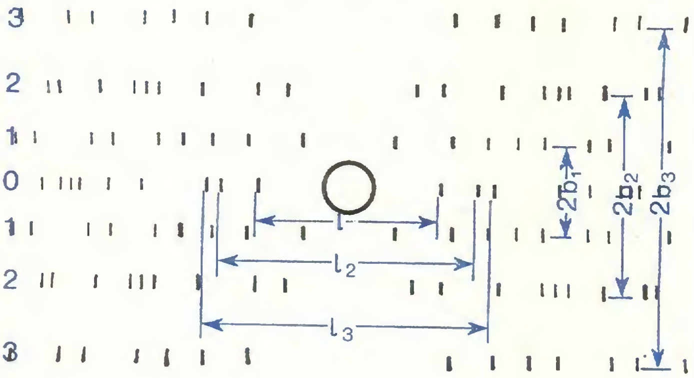
Zbiór wiązek promieni odbitych od płaszczyzn sieciowych danego monokryształu, czyli jego tzw. obraz dyfrakcyjny, rejestruje się najczęściej na błonie fotograficznej zwiniętej w walec, umieszczony współosiowo z osią obrotu kryształu. Na tak wykonanym rentgenogramie, po jego rozprostowaniu, refleksy ułożone są wzdłuż równoległych do siebie prostych, które nazywane są warstwicami (ilustracja 70, tabl. 19). Warstwice są oczywiście śladami stożków dyfrakcyjnych.

obraz dyfrakcyjny

warstwice

wyznaczanie periodu identyczności

Przy danej długości fali λ promieniowania odległość między warstwicami b_n zależy od periodu identyczności J tej prostej sieciowej, wokół której kryształ



był obracany. J można obliczyć stosując wzór $n\lambda = J \sin \nu_n$, gdzie n jest rzędem ugięcia i numerem kolejnej warstwicy, ν_n — kątem warstwicowym dla n-tej warstwicy. Kąt ν_n wyznacza się z zależności $\operatorname{tg} \nu_n = b_n/r$, gdzie r jest promieniem cylindrycznie zwiniętej błony fotograficznej, a b_n — odległością (w mm) między warstwicą n-tą a zerową.

Mierząc wzdłuż jednej warstwicy odległość l (w mm) między każdymi dwoma refleksami symetrycznie położonymi względem śladu wiązki padającej (rys. 4b), można wyznaczyć braggowski kąt ϑ, co z kolei pozwala na obliczenie — z równania Wulfa–Bragga — odległości międzypłaszczyznowej $d_{(hkl)}$ tej rodziny płaszczyzn sieciowych, które odbijając promienie rentgenowskie utworzyły dane refleksy. Dla warstwicy zerowej zachodzi np. zależność:

wyznaczanie $d_{(hkl)}$

$$\frac{l}{2\pi r} = \frac{4\vartheta}{360°}.$$

metoda kołysanego kryształu

Odmianami metody obracanego kryształu są: metoda kołysanego kryształu (w której monokryształ jest obracany nie stale o 360°, lecz jedynie tam i z powrotem, w pewnym niewielkim zakresie kątów) oraz metody goniometryczne.

Ponieważ na podstawie wielkości kąta warstwicowego ν można z łatwością bezpośrednio wyznaczać periody identyczności prostych sieciowych, metody obracanego i kołysanego kryształu są podstawowymi metodami określania rozmiarów komórek elementarnych. W tym celu wykonuje się rentgenogramy przy obracaniu monokryształu kolejno wokół osi krystalograficznych X, Y, Z.

wyznaczanie liczby Z

Po wyznaczeniu kształtu i rozmiarów komórki elementarnej można wyznaczyć liczbę Z atomów (cząsteczek) znajdujących się w komórce elementarnej. Wykonuje się to w następujący sposób: Należy masę jednej komórki elementarnej podzielić przez wyrażoną w gramach masę jednego atomu (czy jednej cząsteczki). Masę jednej komórki elementarnej znajduje się mnożąc jej objętość V przez gęstość monokryształu D. Masę jednego atomu (cząsteczki) uzyskuje się przez pomnożenie masy atomowej (cząsteczkowej) przez $^1/_{12}$ masy atomu węgla ^{12}C, wynoszącą $1{,}660435 \cdot 10^{-24}$ g. Ostatecznie liczbę Z oblicza się ze wzoru

$$Z = \frac{V \cdot D}{M \cdot 1{,}660435 \cdot 10^{-24}},$$

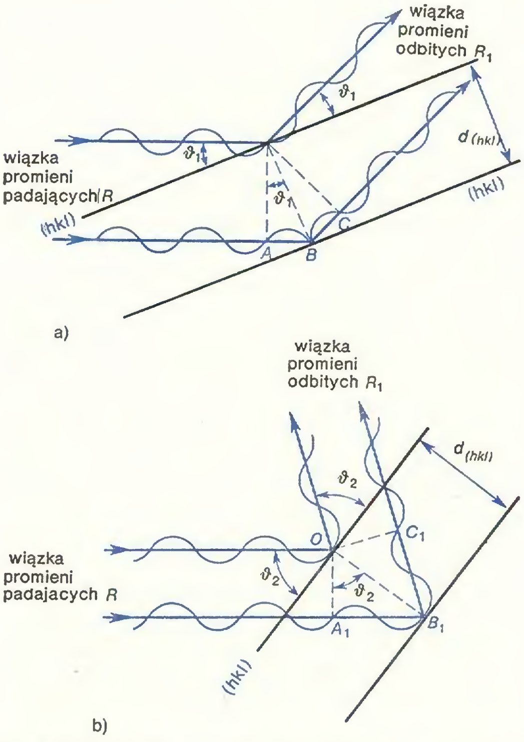
gdzie M oznacza masę atomową (cząsteczkową), V jest wyrażone w cm^3, a D w g/cm^3.

wskaźnikowanie refleksów

W krystalografii rentgenowskiej jednym z ważniejszych, ale i trudniejszych zadań jest tzw. wskaźnikowanie refleksów, tj. przypisywanie poszczególnym refleksom wskaźników płaszczyzn sieciowych odbijających promienie rentgenowskie pod danym kątem ϑ, połączone z ustalaniem rzędu odbicia.

Znajomość wskaźników refleksów jest niezbędna np. przy określaniu grupy przestrzennej kryształów czy wyznaczaniu położeń atomów w komórce elementarnej.

Wskaźnikując refleksy uwzględnia się fakt, że tylko ich część powstaje w wyniku odbicia promieni przez płaszczyzny sieciowe o różnych odległościach międzypłaszczyznowych. Wiele refleksów powstaje bowiem w wyniku odbicia promieni przez jedną i tę samą płaszczyznę sieciową pod różnymi kątami ϑ — spełniającymi jednak zawsze równanie Wulfa–Bragga. Są to refleksy otrzymywane w różnych rzędach odbicia, dla $n = 1, 2, 3...$ Różnica dróg promieni odbitych w rzędzie pierwszym ($n = 1$) od dwoch sąsied-



Rys. 5. Różnica dróg promieni odbitych: a) w rzędzie pierwszym pod kątem ϑ_1, różnica dróg $\Delta S_1 = ABC = 1\lambda$, b) w rzędzie drugim pod kątem ϑ_2, różnica dróg $\Delta S_2 = A_1B_1C_1 = 2\lambda$

nich płaszczyzn sieciowych jest równa 1λ, w rzędzie drugim ($n = 2$) — równa jest 2λ, w rzędzie trzecim ($n = 3$) — 3λ itd. (rys. 5).

Wskaźnikowane refleksy oznacza się millerowskimi wskaźnikami *h*, *k*, *l* tych płaszczyzn sieciowych, które odbijały promieniowanie, przemnożonymi przez rząd odbicia. Otrzymane w ten sposób liczby całkowite *h*, *k*, *l* nazywają się wskaźnikami refleksów. Nie muszą być one oczywiście liczbami pierwszymi względem siebie, a dla odróżnienia od symboli płaszczyzn sieciowych zapisuje się je bez nawiasu okrągłego. Tak więc np. symbol 444 oznacza refleks powstały w wyniku odbicia promieni w rzędzie czwartym od płaszczyzny sieciowej (111). Przy wyznaczaniu grupy przestrzennej kryształu czy do celu analizy strukturalnej kryształów wskaźnikuje się zwykle refleksy na rentgenogramach wykonanych metodami goniometrycznymi. Można jednakże wskaźnikować refleksy i na rentgenogramach otrzymywanych metodą Lauego lub metodami proszkowymi.

symbole refleksów

Wyobraźmy sobie kryształ, w którym istnieją różne rodziny płaszczyzn sieciowych o jednakowych lub prawie jednakowych odległościach międzypłaszczyznowych. Jak się okazuje, tego rodzaju sytuacja ma duży wpływ na wygląd rentgenogramu wykonanego metodą obracanego kryształu, gdyż leżące na jednym stożku dyfrakcyjnym promienie rentgenowskie, odbite przez płaszczyzny sieciowe o takich samych lub bardzo bliskich wartościach $d_{(hkl)}$, utworzą jeden refleks (w obrębie jednej warstwicy kierunki promieni odbitych zależą tylko od $d_{(hkl)}$). W rezultacie takiego nakładania się na siebie refleksów niemożliwe się staje przeprowadzenie jednoznacznego ich wskaźnikowania. Oznacza to, że jednemu refleksowi można przypisać wskaźniki dwu lub nawet kilku płaszczyzn odbijających. Nie można również określić intensywności nakładających się na siebie refleksów. Niejednoznaczne wskaźnikowanie refleksów może np. uniemożliwić prawidłowe wyznaczenie grupy przestrzennej badanego kryształu.

niejednoznaczność wskaźnikowania

Nakładania się na siebie refleksów można jednak uniknąć wykonując rentgenogramy w specjalny sposób. Wykorzystuje się przy tym fakt, że np. płaszczyzny sieciowe (*hkl*) i (*h'k'l'*) o bliskich lub takich samych odległościach międzypłaszczyznowych nie znajdą się podczas wykonywania rentgenogramu metodą obracanego kryształu w pozycji odbijającej równocześnie, lecz znajdą się w niej kolejno, po obrocie kryształu o pewien kąt φ, tj. kąt między normalnymi płaszczyzn sieciowych (*hkl*) i (*h'k'l'*). Jeżeli podczas obracania kryształu o kąt φ przesunięta zostanie błona fotograficzna, to promienie rentgenowskie odbite od płaszczyzn (*hkl*) i (*h'k'l'*) nie padną na to samo miejsce błony.

metody goniometryczne

Metody goniometryczne polegają na zsynchronizowanym z obrotem kryształu przesuwaniu błony fotograficznej. W metodach tych na jednej błonie fotograficznej rejestruje się refleksy należące tylko do jednej warstwicy; pozostałe warstwice zostają usunięte przez specjalne przysłony (tzw. blendy warstwicowe). Tak więc w metodach goniometrycznych refleksy jednej warstwicy zostają rozłożone na całej powierzchni rentgenogramu. Kamery rentgenowskie, za pomocą których uzyskuje się tego rodzaju dwuwymiarowe „rozwinięcia" poszczególnych warstwic, nazywają się goniometrami rentgenowskimi. Najczęściej stosowanymi metodami goniometrycznymi są metoda Weissenberga i metoda de Jonga–Boumana.

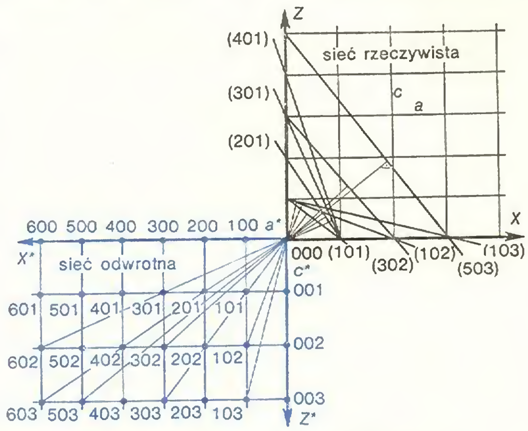
goniometry rentgenowskie

metoda Weissenberga

W metodzie Weissenberga skolimowana wiązka monochromatycznych promieni rentgenowskich pada na obracający się wokół wybranej prostej sieciowej monokryształ. Cylindrycznie zwinięta błona fotograficzna, umieszczona współosiowo z osią obrotu monokryształu, przesuwa się — tam i z powrotem — równolegle do osi obrotu monokryształu. Pomiędzy kryształem a błoną fotograficzną znajduje się cylindryczna blenda warstwicowa ze szczeliną. Przez tę szczelinę promienie odbite od płaszczyzn sieciowych dochodzą do błony fotograficznej. W kamerze Weissenberga rozwinięcie dowolnej warstwicy można uzyskać zmieniając w określony sposób kąt między kierunkiem wiązki padającej promieni a osią obrotu kryształu i ustawiając odpowiednio szczelinę blendy warstwicowej. Wskaźnikowanie refleksów na rentgenogramach wykonanych metodą Weissenberga jest jednak dość skomplikowane, ponieważ metodą tą otrzymuje się na rentgenogramach zniekształcone obrazy płaszczyzn sieci odwrotnej (il. 69, tabl. 18 i il. 76, tabl. 20).

sieć odwrotna

Sieć odwrotna jest konstrukcją ściśle związaną z siecią przestrzenną kryształu i służy do rozwiązywania szeregu zagadnień krystalografii rentgenowskiej. Sieć odwrotną konstruuje się prowadząc z dowolnego węzła „rzeczywistej" sieci przestrzennej — przyjętego za początek układu osi współrzędnych — normalne do wszystkich rodzin płaszczyzn sieciowych (*hkl*) (rys. 6). Wzdłuż tych normalnych zaznacza się punkty (węzły) znajdujące się w odległościach $d^* = n/d_{(hkl)}$ od początku układu współrzędnych (*n* jest liczbą całkowitą). Otrzymane punkty ułożone są w przestrzeni



Rys. 6. Konstrukcja płaszczyzny X^*Z^* sieci odwrotnej

periodycznie, tworząc trójwymiarową sieć nazywaną siecią odwrotną. Jak wynika z samej konstrukcji, płaszczyznom sieci rzeczywistej odpowiadają węzły sieci odwrotnej. Wskaźniki węzłów sieci odwrotnej są wskaźnikami płaszczyzn sieciowych pomnożonymi przez liczbę *n*. Osie sieci odwrotnej oznacza się symbolami X^*, Y^*, Z^*, a jej parametry symbolami a^*, b^*, c^*, α^*, β^*, γ^*. Z konstrukcji sieci odwrotnej wynika również, że

$$a^* = \frac{1}{d_{(100)}}, \qquad b^* = \frac{1}{d_{(010)}}, \qquad c^* = \frac{1}{d_{(001)}}.$$

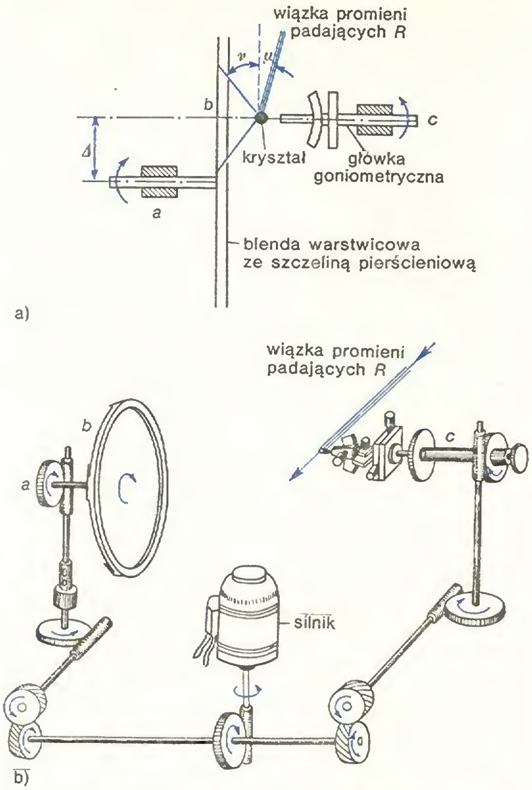
Sieć odwrotna przedstawia wszystkie teoretycznie możliwe odbicia promieni rentgenowskich od płaszczyzn sieciowych kryształu. Warstwica na rentgenogramie otrzymanym metodą obracanego kryształu odpowiada jednej płaszczyźnie sieci odwrotnej.

metoda de Jonga–Boumana

retigraf

Niezniekształcone obrazy płaszczyzn sieci odwrotnej otrzymuje się goniometryczną metodą de Jonga–Boumana, w kamerze nazywanej retigrafem (il. 74, tabl. 20). Skolimowana wiązka monochromatycznych promieni rentgenowskich pada na obracający się monokryształ. Wiązka padająca tworzy z płaszczyzną prostopadłą do osi obrotu monokryształu kąt μ (rys. 7); zmieniając odpowiednio ten kąt można uzyskać rozwinięcie dowolnej warstwicy, tzn. można otrzymać obraz dowolnej płaszczyzny sieci odwrotnej, prostopadłej do osi obrotu. Promienie odbite od płaszczyzn sieciowych kryształu, leżące na powierzchni jednego stożka dyfrakcyjnego, docierają do błony fotograficznej przez pierścieniową szczelinę w płaskiej blendzie warstwicowej. Płaska błona fotograficzna *b*, ustawiona prostopadle do osi obrotu kryształu, obraca się z tą samą prędkością kątową i w tym samym kie-

runku co monokryształ. Osie obrotu monokryształu *a* i błony fotograficznej są do siebie równoległe i przesunięte względem siebie na odległość Δ, zależną od



Rys. 7. Metoda de Jonga–Boumana: a) zasada metody, b) kinematyczny schemat kamery

retigram

numeru rozwijanej warstwicy. Otrzymywane rentgenogramy nazywa się retigramami. Refleksy na retigramach mają kształt litery x.

Wskaźnikowanie refleksów na retigramach jest bardzo łatwe, polega na określeniu współrzędnych refleksów względem osi sieci odwrotnej (il. 72, tabl. 19).

metoda precesyjna

Na podobnej zasadzie oparta jest metoda precesyjna, w której monokryształ nie wykonuje ruchu obrotowego, lecz ruch precesyjny, tak że np. jedna z jego osi (X, Y lub Z) obraca się wokół wiązki promieni padających. tworząc z nią stały kąt.

Wyznaczanie grupy przestrzennej kryształu

Może się zdarzyć, że na rentgenogramach po prostu brakuje niektórych refleksów i na retigramach widoczne są jakby puste miejsca. Mówimy wtedy, że pewne refleksy zostały wygaszone. Jeżeli brakuje refleksów w niektórych, ściśle określonych dla danej płaszczyzny sieciowej rzędach odbicia, to mamy do czynienia z wygaszaniami systematycznymi. Wygaszania systematyczne refleksów są powodowane przez centrowane komórki elementarne oraz przez strukturalne elementy symetrii. Z wygaszaniami systematycznymi refleksów mamy do czynienia wówczas, gdy w wyniku odbicia promieni rentgenowskich, np. od płaszczyzny (110), powstają tylko refleksy 220, 440, 660, 880, ..., a brak refleksów 110, 330, 550, 770 ... itp.

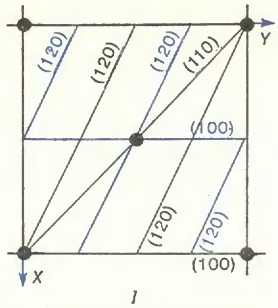
wygaszania systematyczne

Badanie wygaszań systematycznych, które wymaga uprzedniego wywskaźnikowania refleksów, prowadzi więc do wykrywania w krysztale osi śrubowych i płaszczyzn poślizgu oraz ustalania typu sieci Bravais'go.

Wygaszania systematyczne powodowane przez centrowane komórki elementarne nazywają się wygaszaniami integralnymi. Powstają one dlatego, że wskutek centrowania komórki pojawiają się pomiędzy niektórymi płaszczyznami sieciowymi „dodatkowe"

wygaszania integralne

Rys. 8. Płaszczyzny sieciowe (100), (110), (120), w komórce elementarnej sieci typu *I* (komórkę elementarną i płaszczyzny sieciowe przedstawiono w rzucie prostokątnym na płaszczyznę XY). Liniami niebieskimi zaznaczono płaszczyzny „dodatkowe"



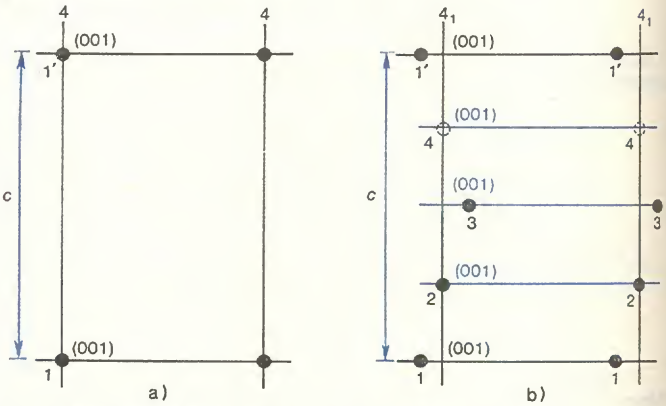
płaszczyzny (w porównaniu z komórką prymitywną), tak że dla pewnych rzędów odbicia różnice dróg stają się równe nie $n\lambda$, lecz $(n/2)\lambda$. W sieci typu *I* dodatkowe płaszczyzny sieciowe znajdują się np. pomiędzy płaszczyznami (100) i (120), a nie ma ich między płaszczyznami (110) (rys. 8). Na rentgenogramach kryształów mających sieci typu *I* wystąpią więc np. refleksy 002, 004, 006 ..., 240, 480, 6·12·0 ..., 110, 220, 330, 440 .... Ogólnie: dla sieci typu *I* występują tylko te refleksy, dla których suma wskaźników jest liczbą parzystą: $h+k+l = 2n$. Powyższe uogólnienie jest tzw. regułą wygaszań dla sieci typu *I*.

reguła wygaszań

Typ sieci Bravais'go w krysztale można również wyznaczyć bez wskaźnikowania refleksów i badania reguł wygaszań, po prostu mierząc oprócz stałych sieciowych *a*, *b*, *c* periody identyczności prostych sieciowych będących przekątnymi ścian komórki elementarnej oraz period identyczności jednej z prostych sieciowych będących przekątnymi przestrzennymi komórki elementarnej.

wygaszania seryjne

Śrubowe osie symetrii powodują powstawanie wygaszań systematycznych zwanych wygaszaniami seryjnymi. Są to wygaszania dotyczące refleksów powstających w wyniku odbicia promieni rentgenowskich od płaszczyzny sieciowej prostopadłej do osi śrubowej. Wygaszania te pojawiają się dlatego, że w wyniku działania *n*-krotnej osi śrubowej następuje jakby rozszczepienie płaszczyzny sieciowej prostopadłej do niej np. na *n* płaszczyzn (w porównaniu z działaniem zwykłej *n*-krotnej osi symetrii; rys. 9). Jeżeli w sieci przestrzennej istnieje np. oś 4_1 o kierunku [001], to refleksy o wskaźnikach typu 00*l* wystąpią na rentgenogramie tylko wtedy, gdy $l = 4n$ (tzn. refleksy 004, 008, 00·12 ...).

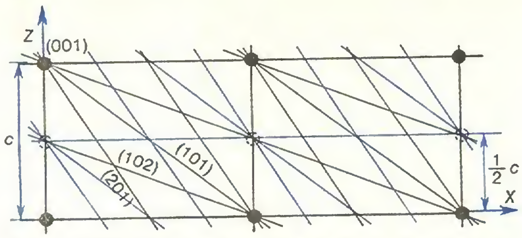


Rys. 9. Płaszczyzna sieciowa (001) w komórkach elementarnych sieci przestrzennych z czterokrotną osią symetrii: a) zwykłą, b) śrubową (rzut prostokątny na płaszczyznę YZ); liniami niebieskimi zaznaczono „dodatkowe" płaszczyzny (001)

Trzeci rodzaj wygaszań systematycznych — wygaszania pasowe — powodują płaszczyzny poślizgu, w wyniku bowiem działania tych ostatnich także po-

wygaszania pasowe

wstają dodatkowe płaszczyzny sieciowe (rys. 10). Jeżeli np. w krysztale występuje płaszczyzna poślizgu typu *c* równoległa do (010), to refleksy o wskaźnikach



Rys. 10. Płaszczyzna poślizgu *c* równoległa do płaszczyzny rysunku powoduje pojawienie się dodatkowych (linie niebieskie) płaszczyzn sieciowych (*h0l*), w stosunku do sieci przestrzennej bez płaszczyzny poślizgu (rzut prostokątny na płaszczyznę *XZ*)

typu *h0l* wystąpią na rentgenogramach tylko wtedy, gdy $l = 2n$.

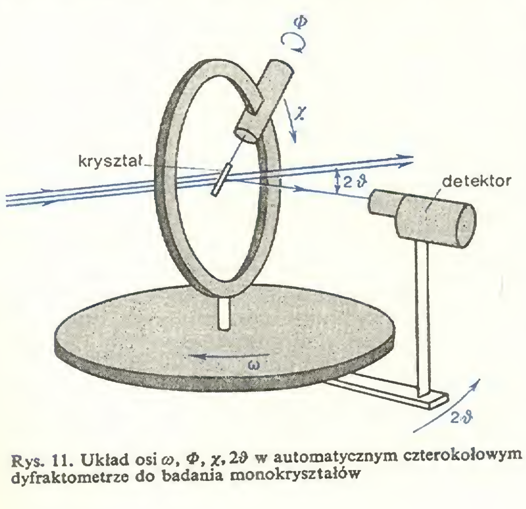
Dla każdego rodzaju sieci Bravais'go, osi śrubowej, płaszczyzny poślizgu istnieje ściśle określona reguła wygaszań refleksów. Kombinacje tych reguł dają prawa wygaszań dla poszczególnych grup przestrzennych. Prawa te, zebrane w specjalnych tabelach (np. International Tables for *X*-Ray Crystallography, Birmingham 1952), umożliwiają identyfikację grup przestrzennych (dyfrakcyjnych) po przeprowadzeniu analizy wygaszań refleksów. Ponieważ na podstawie wygaszań nie można stwierdzić obecności lub braku środka symetrii w krysztale, a pozostałe makroskopowe elementy symetrii również nie wpływają na wygaszania, to niektórych grup przestrzennych nie można od siebie odróżnić, gdyż grupy te dają jednakowe wygaszania. W związku z tym dany kryształ można zakwalifikować jedynie do jednej ze 122 grup dyfrakcyjnych, a nie do jednej z 230 grup przestrzennych. Ostatecznego wyboru grupy przestrzennej w obrębie grupy dyfrakcyjnej można dokonać badając anizotropię niektórych fizycznych własności kryształów (np. występowanie zjawiska piezoelektrycznego lub piroelektrycznego) lub analizując liczbę maksimów na wykresie funkcji Pattersona (zob. rozdział „Strukturalna analiza kryształów").

tablice międzynarodowe

grupy dyfrakcyjne

Czterokołowy dyfraktometr do monokryształów

Najdoskonalszym obecnie przyrządem do badania struktury kryształów jest automatyczny czterokołowy dyfraktometr rentgenowski, określający z wielką precyzją położenie odbitych wiązek promieni oraz ich natężenie. Dyfraktometr taki składa się ze źródła promieniowania rentgenowskiego, czterokołowego



Rys. 11. Układ osi ω, Φ, χ, 2ϑ w automatycznym czterokołowym dyfraktometrze do badania monokryształów

goniometru, detektora, urządzenia pomiarowo-rejestrującego i komputera.

Monokryształ przymocowany do główki goniometrycznej można obracać niezależnie wokół czterech osi ω, Φ, χ, 2ϑ (rys. 11) za pomocą 4 kół (łuków) urządzenia goniometrycznego. Kryształ znajduje się zawsze w punkcie przecięcia się osi wszystkich czterech kół. Obroty kryształu wokół osi ω, Φ, χ pozwalają ustawić dowolną jego płaszczyznę sieciową w pozycji odbijającej promieniowanie. Koło 2ϑ umożliwia ustawienie detektora w takiej pozycji, przy której dany promień odbity może zostać zarejestrowany (detektorem jest zwykle licznik proporcjonalny lub scyntylacyjny). Dzięki licznikom można zmierzyć intensywność refleksu z bardzo dużą dokładnością — ok. 1%.

osie ω, Φ, χ i ϑ

Zadaniem komputera jest sterowanie procesem automatycznego poszukiwania refleksu oraz pomiaru jego położenia i natężenia, a także rejestrowanie mierzonych wielkości.

Niektóre typy dyfraktometrów wymagają uprzedniego wyznaczenia stałych sieciowych i grupy przestrzennej badanego kryształu (metodami fotograficznymi) oraz wstępnego zorientowania kryształu na goniometrze. Inne typy — nie potrzebują żadnych wstępnych wiadomości o badanym krysztale oraz nie wymagają orientowania kryształu: same, sterowane swoim minikomputerem określają i mierzą komórkę elementarną. Istnieją też dyfraktometry, które ze zmierzonych intensywności od razu przygotowują dane (zbiory wartości $|F_{hkl}|$) do analizy strukturalnej.

Automatyczne dyfraktometry do monokryształów (il. 78, tabl. 20), mimo że zapewniają znacznie większą precyzję i szybkość pomiarów niż metody fotograficzne, nie są jeszcze przyrządami idealnymi, zwłaszcza gdy chodzi o badanie struktur kryształów niestabilnych (np. kryształów białek) czy kryształów o bardzo dużych krawędziach komórki elementarnej. Zbyt długi jest bowiem czas pomiaru każdego refleksu (wynosi 1,5–3 min) i czas poszukiwania refleksu, tzn. odpowiedniego orientowania czterech osi goniometru. Trzeba wziąć pod uwagę, że mierzy się nieraz po kilka tysięcy refleksów (np. A. Korczyński, M. Nardelli, M. A. Pellinghelli zmierzyli w r. 1973 przy badaniu struktury kryształów $HgCl_2 \cdot 4SC(NH_2)_2$ za pomocą dyfraktometru AED Siemensa intensywność 5785 refleksów, z szybkością 1 pomiar na 90 s, nie licząc czasu ustawienia przez przyrząd płaszczyzn sieciowych w pozycji odbijającej). Przy badaniu struktury kryształów białek — oprócz konieczności zmierzenia wielu tysięcy refleksów — występuje dodatkowe utrudnienie, związane z rozkładem chemicznym tych kryształów podczas pomiarów; w rezultacie wielokrotna wymiana kryształów na goniometrze wydłuża czas badania.

Trwające obecnie na świecie poszukiwania nowych rozwiązań konstrukcyjnych dyfraktometrów, umożliwiających skrócenie czasu przeprowadzania badań jednego kryształu, idą w kilku kierunkach: w kierunku przyspieszania pracy istniejących goniometrów, np. przez równoczesne ustawianie wszystkich czterech osi goniometru czy zwiększanie szybkości obrotu monokryształu (do 1500°/min) wokół poszczególnych osi; w kierunku stosowania źródeł promieniowania rentgenowskiego o dużej mocy (lampy rentgenowskie z wirującą anodą), co skraca czas ustawiania kryształu i detektora we właściwej pozycji dzięki powstawaniu intensywniejszych wiązek odbitych, oraz kierunku zastosowania zamiast jednego licznika kilku liczników, mierzących równocześnie kilka wiązek, tzw. „jednoczesnych" lub „prawie jednoczesnych".

dyfraktometry szybko pracujące

Szybkie pomiary przeprowadza się również w dyfraktometrach pozycyjnych, w których obraz dyfrakcyjny jest rejestrowany przez mozaikę liczników, albo w tzw. dyfraktometrach energodyspersyjnych, w których się stosuje promieniowanie polichromatyczne i rejestruje się równocześnie wszystkie rzędy odbicia od jednej płaszczyzny sieciowej. W pierwszym wypadku miniaturowe liczniki półprzewodnikowe lub proporcjonalne układa się w jedno- lub dwuwymiaro-

dyfraktometry pozycyjne

we mozaiki. Kryształ wykonuje tylko jeden obrót o 360°, a podczas tego obrotu mierzy się wszystkie wiązki odbite odpowiadające jednej płaszczyźnie sieci odwrotnej. W jednym z dyfraktometrów zastosowano np. jednowymiarową mozaikę liczników półprzewodnikowych rozmieszczoną na łuku 180°. W tej mozaice znajdowało się 128 liczników o wymiarach 3×4 mm. W takim dyfraktometrze uzyskano prawie dziesięciokrotne skrócenie czasu przeprowadzania pomiarów jednego kryształu. W innym dyfraktometrze zastosowano 4 segmenty łukowe mozaiki liczników, a każdy z nich zawierał 128 liczników proporcjonalnych. Wydajność takiego dyfraktometru zależy od stałych sieciowych badanego kryształu i wygaszań systematycznych i jest 20–60 razy większa niż wydajność dyfraktometru z jednym licznikiem.

mozaiki liczników

W dyfraktometrach do monokryształów stosuje się również tzw. wieloniciową kamerę proporcjonalną, będącą dwuwymiarowym detektorem „czułym na pozycję", oraz pozycyjno czułe detektory scyntylacyjne, przetwarzające (za pomocą ekranu — konwertora polikrystalicznego ZnS lub monokrystalicznego NaI) rentgenowski obraz dyfrakcyjny na obraz w świetle widzialnym. Jasność takiego obrazu przetwornik elektronooptyczny wzmacnia 10^5 razy, a po wzmocnieniu obraz ten jest zapamiętywany w lampie telewizyjnej (ikonoskopie) i dalej odczytywany przez elektroniczną maszynę cyfrową. Te ostatnie dyfraktometry imitują w swej pracy serię rentgenogramów wykonywanych metodą kołysanego kryształu.

kamera wieloniciowa

detektory czułe na pozycję

Metody proszkowe

Większość ciał stałych występujących w przyrodzie i wytwarzanych przez człowieka (np. skały, metale, materiały ceramiczne i budowlane) to konglomeraty drobnych, mikroskopowej wielkości kryształów. Nie występują tu tak duże kryształy, by do ich badania można było zastosować metodę obracanego kryształu czy metodę Lauego.

Rentgenostrukturalne badania ciał polikrystalicznych przeprowadza się za pomocą metody opracowanej przez Debye'a, Sherrera i Hulla oraz jej różnych odmian (takich jak metody ogniskujące czy metoda dyfraktometryczna).

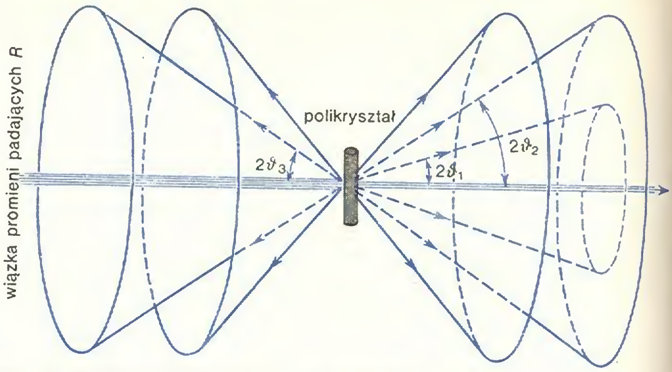
metoda Debye'a–Scherrera–Hulla

W metodzie Debye'a–Scherrera–Hulla wąska wiązka równoległych, monochromatycznych promieni rentgenowskich pada na preparat sporządzony z krystalicznego proszku. Preparat może mieć kształt pręcika o średnicy ok. 0,3 mm i długości 2–3 mm. Wielkość ziaren proszku powinna być rzędu 1 µm, co oznacza, że w celu otrzymania takiego proszku należałoby np. sześcienny monokryształ o krawędzi długości 1 mm podzielić na ok. 10^9 drobnych krystalitów. W takim preparacie w odbijaniu promieni rentgenowskich biorą udział te płaszczyzny sieciowe poszczególnych krystalitów, które zupełnie przypadkowo znajdują się w pozycji spełniającej równanie Wulfa–Bragga. W każdym krystalicie odbija promienie rentgenowskie tylko jedna płaszczyzna sieciowa, ale teoretycznie w każdym krystalicie może to być inna płaszczyzna. Ponieważ w preparacie jest praktycznie nieskończona liczba krystalitów, a liczba płaszczyzn sieciowych odbijających promieniowanie jest ograniczona, to w wielu krystalitach promienie rentgenowskie będą odbijane przez tę samą płaszczyznę sieciową.

Płaszczyzny sieciowe o tej samej odległości międzypłaszczyznowej, ułożone w preparacie chaotycznie, odbijają promienie rentgenowskie w różnych kierunkach przestrzeni, ale zawsze pod tym samym kątem $2\vartheta_{hkl}$ w stosunku do wiązki padającej. W rezultacie promienie odbite w różnych krystalitach od tej samej rodziny płaszczyzn sieciowych (*hkl*) leżą na powierzchni stożka (tzw. stożka Debye'a), którego osią jest promień padający. Dla każdej rodziny płaszczyzn sieciowych powstaje odrębny stożek promieni odbitych (rys. 12).

stożek Debye'a

Obraz dyfrakcyjny preparatu proszkowego rejestruje się albo na płaskiej błonie fotograficznej, albo znacznie częściej — na błonie fotograficznej zwiniętej



Rys. 12. Stożki promieni odbitych w metodzie Debye'a–Scherrera–Hulla

w walec. W pierwszym wypadku refleksy na rentgenogramie mają kształt pierścieni o różnej intensywności, w drugim — linii krzywych wyższego rzędu (il. 67, tabl. 18). Dla danego refleksu można zawsze wyznaczyć kąt połysku ϑ, obliczyć wartość d_{hkl} oraz zmierzyć intensywność. Obraz dyfrakcyjny rejestruje się również za pomocą licznika Geigera–Müllera lub scyntylacyjnego w dyfraktometrze rentgenowskim.

Metodami proszkowymi można rozwiązywać niektóre zadania krystalografii rentgenowskiej, jak np. wyznaczanie stałych sieciowych czy grup przestrzennych. Jednak możliwości stosowania tych metod w krystalografii rentgenowskiej czy w analizie strukturalnej kryształów są bardzo ograniczone, przede wszystkim ze względu na trudności przy wskaźnikowaniu refleksów. Wskaźnikowanie to jest stosunkowo proste jedynie na rentgenogramach proszkowych kryształów regularnych, heksagonalnych i tetragonalnych; natomiast dla kryształów trójskośnych, jednoskośnych i rombowych wskaźnikowanie jest bardzo trudne, a w dodatku często niejednoznaczne ze względu na nakładanie się refleksów.

zastosowanie metod proszkowych

Rentgenografia stosowana

Metody proszkowe znajdują natomiast szereg zastosowań praktycznych w chemii, fizyce, metalurgii czy geologii i są podstawą tzw. rentgenografii stosowanej. Do takich praktycznych zastosowań należą: jakościowe i ilościowe określanie składu fazowego substancji (tzw. analiza fazowa), wyznaczanie wielkości krystalitów, precyzyjne pomiary stałych sieciowych (dokładność ich określenia może osiągnąć 10^{-5}–10^{-6}!), wyznaczanie współczynników rozszerzalności, badanie tekstur, badanie naprężeń wewnętrznych.

analiza fazowa

Każda substancja krystaliczna oddziałując z padającymi na nią promieniami rentgenowskimi tworzy charakteryzujący ją jednoznacznie obraz dyfrakcyjny. Tak więc i rentgenogramy proszkowe (bo tylko na nich zarejestrowany jest cały obraz dyfrakcyjny) jednoznacznie charakteryzują daną substancję i to bez względu na to, czy występuje ona jako czysta faza, czy jako składnik mieszaniny. Jeśli badana substancja stanowi mieszaninę dwóch lub więcej związków chemicznych (faz), to każda z tych faz daje na rentgenogramie proszkowym swoje własne refleksy. Analiza fazowa pozwala wykrywać substancje krystaliczne w takiej postaci, w jakiej występują one w danym preparacie, a nie w postaci jonów czy pierwiastków chemicznych wchodzących w ich skład. Tak więc np. dyfrakcyjna analiza fazowa pozwala stwierdzić, że w mieszaninie NaCl i KBr występują chlorek sodowy i bromek potasowy, a nie chlorek potasowy i bromek sodowy (analiza chemiczna wykazałaby w takiej mie-

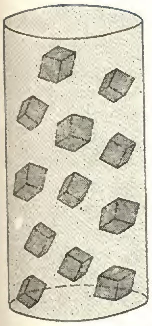
szaninie obecność jonów Na^+, K^+, Cl^-, Br^-, lecz nie wskazałaby, jak te jony połączone są ze sobą w ciele stałym). Dyfrakcyjna analiza fazowa pozwala rozróżniać w prosty sposób odmiany polimorficzne ciał krystalicznych. Można odróżnić np. kalcyt $CaCO_3$ od aragonitu $CaCO_3$ i vaterytu $CaCO_3$ lub odróżnić od siebie jedenaście odmian SiO_2.

wzorzec liczbowy

Jakościową analizę fazową przeprowadza się porównując wartości d_{hkl} i intensywności I poszczególnych refleksów wyznaczone z rentgenogramu proszkowego badanej substancji z wartościami d_{hkl} i I podanymi w tzw. rentgenograficznym wzorcu liczbowym. Wzorzec ten jest tabelką, w której się znajdują wartości d_{hkl} i I otrzymane z rentgenogramu proszkowego substancji wzorcowej. Rentgenograficzne wzorce liczbowe gromadzi się w specjalnych kartotekach, jak np. kartoteka Powder Diffraction Data, wydawana w USA przez JC PDS — International Centre for Diffraction Data (dawniej tzw. kartoteka ASTM). Odpowiedni wzorzec w takiej kartotece (liczącej kilkadziesiąt tysięcy kart) wynajduje się za pomocą specjalnego klucza. Na podstawie rentgenogramu proszkowego można jednoznacznie zidentyfikować substancję — również w mieszaninie — znajdując w kartotece odpowiedni wzorzec.

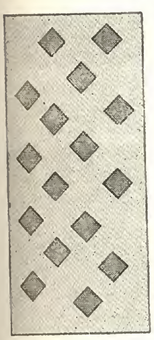
tekstura

Teksturą, albo teksturą krystaliczną, nazywa się uprzywilejowaną wzajemną orientację przestrzenną krystalitów w materiale polikrystalicznym. Tekstury powstają m.in. podczas krystalizacji, odlewania, ciągnienia i zgniatania materiałów. Występują np. we włóknach azbestu, w drutach i blachach. W drutach ciągnionych tekstura polega na ułożeniu krystalitów pewnym kierunkiem w przybliżeniu równolegle do osi drutu (rys. 13). W blachach walcowanych krystality układają się pewnym kierunkiem w przybliżeniu równolegle do kierunku walcowania (oznaczonego strzałką na rys. 14), a pewną płaszczyzną sieciową równolegle do powierzchni blachy. Występowanie tekstury ma wpływ — czasem dodatni, czasem ujemny — na wiele właściwości materiałów, m.in. na wytrzymałość i na twardość metali. Stąd duże znaczenie ma badanie tekstur rentgenostrukturalną metodą proszkową. Występowanie tekstury w materiale polikrystalicznym wpływa na wygląd rentgenogramu proszkowego, albowiem intensywność jednych refleksów wzrasta, a innych maleje lub pozostaje bez zmiany. Refleksy przybierają kształt łuków i układają się wzdłuż warstwic (il. 71b, c, tabl. 19).



Rys. 13. Tekstura drutu aluminiowego; do osi drutu równoległy jest w przybliżeniu kierunek [111] krystalitów

Za pomocą metod proszkowych można uzyskać wiele interesujących danych o polimerach organicznych. Istnienie w strukturze polimerów obszarów bezpostaciowych i krystalicznych powoduje występowanie na rentgenogramie proszkowym rozmytego szerokiego refleksu od fazy bezpostaciowej oraz refleksów charakterystycznych dla fazy krystalicznej (il. 71a, tabl. 19). Pod wpływem np. rozciągania próbki polimeru obszary krystaliczne bezładnie w nim rozłożone mogą uzyskać orientację zależną od kierunku przykładanych sił, czyli w polimerze może zostać wytworzona tekstura. Rentgenogram proszkowy stekturowanego polimeru (również i metalu) jest podobny do rentgenogramu wykonanego metodą obracanego kryształu (il. 71, tabl. 19), a z takiego rentgenogramu, z odległości między warstwicami, można wyznaczyć jedną z krawędzi komórki elementarnej kryształu polimeru. Często jest to jedyna metoda uzyskania stałych sieciowych polimerów, gdyż nie zawsze można uzyskać monokryształy danego polimeru o wielkości odpowiedniej do badań dyfrakcyjnych.



Rys. 14. Tekstura walcowanej blachy stalowej; do powierzchni blachy jest równoległa płaszczyzna sieciowa (100) krystalitów, a do kierunku walcowania — kierunek [110]

G. B. Bokij, M. A. Poraj-Koszic *Rentgenostrukturnyj analiz* t. 1, Moskwa 1964; M. J. Buerger *X-Ray Crystallography*, New York 1942; J. Chojnacki *Elementy krystalografii chemicznej i fizycznej*, Warszawa 1971; W. C. Hamilton *Science* 169, 133 (1970); A. Korczyński, M. Nardelli, M. A. Pellinghelli *Roczniki Chem.* 47, 905 (1973); G.H.W. Milburn *X-Ray Crystallography*, London 1973; K. Łukaszewicz, W. Trzebiatowski *Zarys rentgenograficznej analizy strukturalnej*, Katowice 1960; Z. Trzaska Durski *Podstawy krystalografii rentgenowskiej i analizy strukturalnej kryształów*, Warszawa 1973; W. Trzebiatowski *Zarys rentgenograficznej analizy strukturalnej*, Katowice 1960; M. M. Umanskij *Apparatura rentgenostrukturnych issledowanij*, Moskwa 1960.

Elektronografia

Janusz Leciejewicz

Odkrycie dyfrakcji wiązki elektronów w folii niklowej przez C. J. Davissona i L. H. Germera (1927 r.) było doświadczalnym potwierdzeniem hipotezy L. de Broglie'a o falowym charakterze ruchu cząstek (1924 r.). Długość fali elektronów λ jest związana z pędem elektronu p zależnością de Broglie'a

$$\lambda = h/p = h/mv,$$

gdzie: h — stała Plancka, m — masa spoczynkowa elektronu, v — prędkość elektronu. Wyrażając prędkość elektronów przez przyspieszającą je różnicę potencjałów U i podstawiając wartości liczbowe na h i m, otrzymuje się zależność

$$\lambda = 12{,}26/\sqrt{U},$$

w której λ wyraża się w Å a U w V. W zależności tej nie uwzględnia się relatywistycznego wzrostu masy elektronu ze zwiększaniem się jego prędkości. Błąd w określeniu λ, związany z pominięciem poprawki relatywistycznej, wynosi 2–3% dla U rzędu 20–40 kV, wzrasta jednak do 10% przy różnicy potencjałów 200 kV.

Spójne i sprężyste rozpraszanie wiązki elektronów przez sieć krystaliczną prowadzi do powstawania maksimów interferencyjnych (refleksów) występujących pod kątem ugięcia ϑ określonym prawem Wulfa–Bragga (zob. rozdział „Krystalografia rentgenowska"). Natężenie ugiętych wiązek elektronów zależy od struktury krystalicznej badanej substancji.

ugięcie wiązki elektronowej

Wiązka elektronów padając na atom ulega ugięciu wskutek oddziaływania kulombowskiego z polem elektrostatycznym atomu. Pole to pochodzi zarówno od dodatniego jądra atomu jak i ujemnej powłoki elektronowej. Dodatni potencjał jądra (proporcjonalny do liczby atomowej Z) jest skupiony na małej przestrzeni wokół jądra, natomiast potencjał ujemny pochodzący od powłoki elektronowej, bardziej „rozmyty", jest funkcją zarówno liczby elektronów jak i ich rozkładu w powłoce atomu. Zatem wielkość charakteryzująca ugięcie elektronów na atomie — atomowy czynnik rozpraszania elektronów f_e — zależy od ładunku jądra i od liczby elektronów. Zależność od liczby elektronów w powłoce atomu opisuje atomowy czynnik rozpraszania promieni rentgenowskich f_r, będący funkcją kąta ugięcia ϑ. Ogólnie można to zapisać:

$$f_e(\vartheta) = \mathrm{const}[Z - f_r(\vartheta)].$$

Taka zależność f_e od kąta ugięcia ϑ powoduje, że atomy o małym Z (atomy lekkie) są stosunkowo silnymi ośrodkami rozpraszającymi elektrony. Jest więc możliwe określenie położeń atomów lekkich w obecności atomów z dużym Z (atomów ciężkich) ze znacznie większą dokładnością niż na to pozwala metoda dyfrakcji promieni rentgenowskich.

Elektrony jako cząstki obdarzone ładunkiem elektrycznym, oddziałują silnie z atomami i są absorbowane. Dlatego wnikają do wnętrza kryształu na niewielkie głębokości (przyspieszone różnicą potencjałów 200 kV wnikają na głębokość 1000 Å). Zjawisko to ogranicza zakres stosowalności metody dyfrakcji elektronów do preparatów cienkowarstwowych. Z drugiej strony jednak umożliwia badanie struktury materiałów występujących w postaci cienkich warstw, nawet wtedy gdy grubości ich wynoszą tylko 500 Å.

badanie cienkich warstw

Badanie materiałów w postaci warstw grubości rzędu mikrometrów stwarza konieczność stosowania elektronów przyspieszanych różnicą potencjału 50–100 kV. Elektrony te mają długość fali 0,055–0,039 Å, wtedy braggowskie kąty ugięcia dla większości materiałów nieorganicznych nie przekraczają 3–4°.

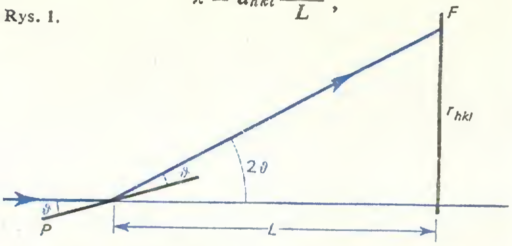
Do określania odległości międzypłaszczyznowych d_{hkl} można użyć równania Wulfa–Bragga ($n\lambda = 2d_{hkl}\sin\vartheta$) w postaci przybliżonej. Ponieważ kąt

określanie d_{hkl}

2ϑ (rys. 1) jest bardzo mały, można napisać $2\sin\vartheta \approx \sin 2\vartheta \approx \mathrm{tg}\, 2\vartheta = r_{hkl}/L$, a więc

$$\lambda = d_{hkl}\frac{r_{hkl}}{L},$$

Rys. 1.



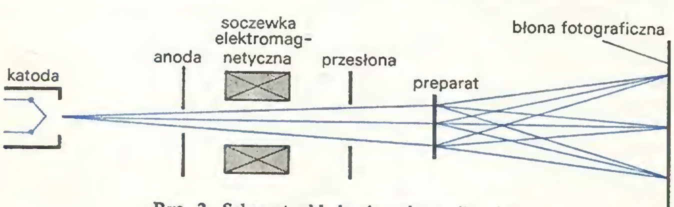
gdzie: r_{hkl} — odległość między śladem ugiętej wiązki a śladem wiązki pierwotnej na błonie fotograficznej, L — odległość między preparatem a błoną fotograficzną. Przybliżenie to wprowadza błąd nie większy niż 1%. Mierząc różnicę potencjału przyspieszającego, wyznacza się długość fali elektronów λ; L jest stałą aparaturową, r_{hkl} natomiast mierzy się na tzw. elektronogramie — obrazie dyfrakcyjnym zarejestrowanym na błonie fotograficznej.

elektronogramy

Można wyróżnić 3 grupy elektronogramów ciał stałych: 1) Elektronogramy punktowe z refleksami w postaci plamek (il. 81a, tabl. 21). Otrzymuje się je, gdy badana substancja jest pojedynczym kryształem lub składa się z oddzielnych bloków krystalicznych o osiach w przybliżeniu równoległych do siebie. Elektronogramy tego rodzaju są głównie wykorzystywane do oznaczenia struktury krystalicznej badanej substancji. 2) Elektronogramy tekstur — maksima dyfrakcyjne (refleksy) mają postać łuków (il. 81b, tabl. 21). Powstają one, gdy badany preparat jest zbiorem bloków krystalicznych rozłożonych w sposób statystycznie przypadkowy, jednakże mających określone płaszczyzny równoległe do siebie. Na jednym takim elektronogramie występują prawie wszystkie refleksy możliwe dla danego typu struktury. Refleksy te układają się wg określonych reguł, co ułatwia ich wskaźnikowanie. Na elektronogramach tekstur można mierzyć z dużą dokładnością natężenia refleksów. Z ich wartości wyznacza się rozkład potencjału elektrycznego w komórce elementarnej, a zatem pozycje atomów w komórce (zob. rozdział „Strukturalna analiza kryształów"). 3) Elektronogramy ciał polikrystalicznych (badany preparat zbudowany jest z chaotycznie rozłożonych bloków krystalicznych), zawierają maksima interferencyjne w postaci pierścieni (il. 81c, tabl. 21). Rozkład pierścieni określa zbiór odległości międzypłaszczyznowych d_{hkl}. Takie elektronogramy wykorzystywane są w badaniach struktury, zwłaszcza struktury substancji o wysokiej symetrii krystalograficznej, do wyznaczenia parametrów komórki elementarnej, a także do wyznaczenia składu fazowego preparatu.

elektronograf

Elektronogramy otrzymuje się bądź przepuszczając wiązkę elektronów przez odpowiednio cienką folię (transmisja), bądź przez odbicie od powierzchni materiału. Schemat układu do rejestracji ugiętych promieni elektronowych (tzw. elektronografu) przedstawia rys. 2. Rozbieżna wiązka elektronów z żarzącej się katody zostaje przyspieszona w polu elektrycznym między katodą a anodą. Wiązka ta zostaje następnie zogniskowana w płaszczyźnie błony fotograficznej



Rys. 2. Schemat układu do rejestracji ugiętych promieni elektronowych

za pomocą soczewki elektromagnetycznej. Preparat substancji badanej (w położeniu transmisji) jest umieszczony między soczewką a błoną fotograficzną. Cały układ znajduje się w wysokiej próżni (10^{-3}–10^{-7} Pa).

W zależności od różnicy potencjału przyspieszającego wyróżnia się dwie techniki dyfrakcyjne. W pierwszej metodzie — metodzie dyfrakcji elektronów niskoenergetycznych, tzw. LEED (Low Energy Electron Diffraction) — elektrony są przyspieszane różnicą potencjałów od 30 do 600 V. Mała przenikliwość powolnych elektronów ogranicza stosowalność tej metody. Przy tym jest konieczne stosowanie preparatów cienkowarstwowych o nadzwyczaj starannie przygotowywanej powierzchni — bez zanieczyszczeń i zniekształceń, wymagana próżnia — rzędu 10^{-7} Pa. Ugięte elektrony są przyspieszane dodatnim potencjałem 5 kV, co ułatwia znacznie rejestrację ugiętych wiązek. Obraz dyfrakcyjny otrzymuje się w bardzo krótkim czasie, co pozwala na śledzenie procesów zachodzących w trakcie naświetlania. Metodę LEED wykorzystuje się w badaniach rozkładu jonów metali na powierzchni katalizatora, mechanizmu absorpcji gazów na powierzchniach, mechanizmu powstawania cienkich warstw tlenkowych na powierzchniach półprzewodników itd. Druga metoda — to metoda dyfrakcji elektronów wysokoenergetycznych, tzw. HEED (High Energy Electron Diffraction). Jest ona stosowana w typowych elektronografach. Elektrony są przyspieszane różnicą potencjałów 50 kV i wnikają do badanego materiału na głębokość do 1000 Å. Większa przenikliwość powoduje, że nie jest konieczne tak bardzo staranne przygotowanie powierzchni, jak przy dyfrakcji elektronów niskoenergetycznych.

metoda LEED

metoda HEED

Metody elektronografii pozwalają na uzyskanie danych szczególnie ważnych do poznania własności materiałów. Umożliwiają określenie struktury wewnętrznej kryształów, tzn. określenie rozmiarów komórki elementarnej kryształu oraz ustalenie pozycji atomów w komórce. Elektronografia nadaje się szczególnie dobrze do badania struktury warstw powierzchniowych i ich zmian wywołanych obróbką technologiczną, procesami utleniania i azotowania, korozją. Struktura warstw powierzchniowych jest na ogół różna od struktury wnętrza materiału (→ Stany powierzchniowe ciał stałych). Bardzo ważne wyniki uzyskuje się z elektronograficznych badań warstw epitaksjalnych oraz badań naprężeń wywołanych deformacją plastyczną. Metody dyfrakcji elektronów są wykorzystywane powszechnie w dziedzinie krystalografii submikroskopowej — do wyznaczania wielkości i kształtu mikroziaren krystalicznych w stanie dużego rozdrobnienia, do określania orientacji mikroziaren na granicy faz, do badania mechanizmu powstawania struktur przejściowych na powierzchniach. W dziedzinie krystalografii submikroskopowej znaczne usługi oddaje łączenie metody dyfrakcji elektronów z mikroskopią elektronową (zob. rozdział „Mikroskopia elektronowa").

J. Chojnacki *Metalografia strukturalna*, Katowice 1966; B. K. Wajnsztejn *Strukturnaja Elektronografija*, Moskwa 1956; B. K. Wajnsztejn, *Sowremiennaja Krystallografia*, Moskwa 1979.

Neutronografia strukturalna i magnetyczna

Janusz Leciejewicz

Rozpraszanie neutronów

Wkrótce po odkryciu neutronu (J. Chadwick, 1932 r.) Mitchell i Powells (1936 r.) wykonali doświadczenie mające na celu sprawdzenie, czy ruch neutronów ma charakter falowy, tzn. czy jest spełnione równanie falowe de Broglie'a

$$\lambda = h/mv,$$

gdzie: λ jest długością fali neutronu, h — stałą Plancka, m — masą neutronu, v — jego prędkością.

dyfrakcja neutronów

Stwierdzono, że neutrony ze źródła radowo-berylowego po przepuszczeniu przez warstwę parafiny uległy odbiciu od płaszczyzny sieciowej (111) kryształu tlenku magnezu pod kątem padania i odbicia wynikającym z równania Wulfa-Bragga dla długości fali neutronów ok. 1,6 Å. Doświadczenie to miało charakter pokazowy. Wskutek bardzo małego natężenia wiązek neutronów ze źródeł radowo-berylowych, zjawisko dyfrakcji neutronów nie mogło być wykorzystane do badania struktury kryształów. Dopiero rozpowszechnienie reaktorów jądrowych po 1945 r. umożliwiło praktyczne zastosowanie dyfrakcji neutronów termicznych do badań strukturalnych.

W oddziaływaniu z kryształem swój charakter falowy szczególnie silnie przejawiają neutrony termiczne. Ulegają one dyfrakcji zgodnie z prawem Wulfa-Bragga podobnie jak promienie rentgenowskie o zbliżonej długości fali (zob. rozdział „Krystalografia rentgenowska") i elektrony (zob. rozdział „Elektronografia"). Zjawisko dyfrakcji neutronów może być więc wykorzystane do określenia położeń atomów w krysztale, czyli do oznaczania struktury kryształu.

Mechanizm oddziaływania neutronów z atomami jest jednak inny niż w wypadku promieni rentgenowskich i elektronów, które są rozpraszane przez powłokę elektronową atomu. Neutrony są rozpraszane przez jądra atomowe. W procesie rozpraszania siły działające między jądrem a neutronem są bardzo krótkiego zasięgu — rzędu promienia jądra, tj. 10^{-14} m (→ Siły jądrowe). Jądra atomowe są więc dla neutronów o długości fali rzędu 10^{-10} m (1 Å) punktowymi ośrodkami rozpraszania.

Długość rozpraszania niektórych nuklidów

Nuklid	Z	b, 10^{-14} m	Nuklid	Z	b, 10^{-14} m
^{1}H	1	−0,378	Fe	26	0,96
^{2}H	1	0,65	Co	27	0,25
^{12}C	6	0,661	Ni	28	1,03
N	7	0,940	W	74	0,466
O	8	0,577	Pb	82	0,96
Ti	22	−0,34	^{238}U	92	0,85

Wg Acta Cryst. A28, 357 (1972)

długość rozpraszania

Zdolność jądra do rozpraszania neutronów termicznych charakteryzuje wielkość zwana długością rozpraszania b, zależna od struktury danego jądra. Tak więc izotopy tego samego pierwiastka mają różne wartości b, a ponadto zależność b od liczby atomowej jest nieregularna (tabela). Dla niektórych nuklidów (np. jąder wodoru — protonów, jąder tytanu) przesunięcie fazowe między rozproszoną falą neutronową a falą padającą wynosi 0, a nie π, jak to jest przy rozpraszaniu promieni rentgenowskich i elektronów. Umownie przyjmuje się więc, że przy wartości zero, długość rozpraszania ma znak ujemny. Niezależność b od liczby atomowej ma istotne znaczenie przy wykorzystaniu dyfrakcji neutronów w analizie strukturalnej kryształów. Przede wszystkim możliwe jest wyznaczenie z jednakową dokładnością położeń atomów lekkich i ciężkich (o małym i dużym Z) występujących obok siebie w komórce elementarnej kryształu. Przy stosowaniu metod krystalografii rentgenowskiej atomy ciężkie odgrywają dominującą rolę w procesie rozpraszania. Stąd oznaczenie położeń atomów lekkich, gdy w komórce elementarnej występują również atomy ciężkie, jest zwykle obarczone dużym błędem.

wyznaczanie położeń atomów lekkich

Wykorzystując metody neutronograficzne można więc wyznaczyć z dużą dokładnością położenia atomów (jąder) wodoru w substancjach organicznych, wodorkach metali i innych związkach zawierających wodór, np. w hydratach, wodorotlenkach itd. Ponieważ jądra wodoru rozpraszają neutrony w znacznej mierze niespójnie (powstaje duże tło), w neutronograficznej analizie strukturalnej korzystniejsze jest stosowanie związków chemicznych, w których atomy wodoru zastąpione są atomami deuteru.

Neutronografia strukturalna

Metody neutronograficzne były i są stosowane przy weryfikacji struktur kryształów tlenków metali ciężkich (np. tlenków ołowiu, tlenków uranu), węglików ciężkich metali (np. węglików wolframu).

rozróżnianie atomów o zbliżonych Z

Metodami tymi można również rozróżniać atomy o zbliżonych liczbach atomowych, np. węgla, azotu i tlenu. Jest to szczególnie ważne przy analizie struktury substancji organicznych. Atomy te są nierozróżnialne dla promieni rentgenowskich, podobnie jak atomy żelaza, niklu i kobaltu występujące w ważnych technicznie stopach i niektórych materiałach magnetycznych (np. w ferrytach).

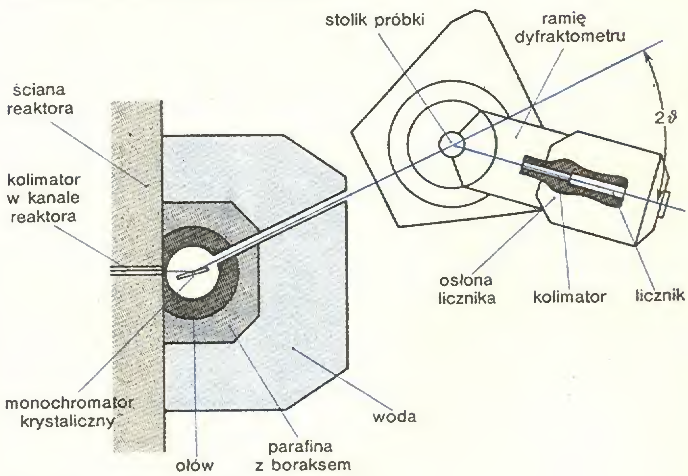
natężenie refleksów

Natężenie neutronowych maksimów interferencyjnych (refleksów), podobnie jak w dyfrakcji promieni rentgenowskich (zob. rozdział „Strukturalna analiza kryształów") jest proporcjonalne do kwadratu modułu amplitudy struktury F_{hkl} (czynnika struktury):

$$I_{hkl} \sim |F_{hkl}|^2,$$

$$F_{hkl} = \sum_{j=1}^{N} b_j \, e^{2\pi i(hx_j+ky_j+lz_j)};$$

N — liczba jąder (atomów) w komórce elementarnej kryształu, x_j, y_j, z_j — współrzędne j-ego jądra w komórce, hkl — wskaźnik refleksu, b_j — długość rozpraszania jądra j. Monochromatyzacji wiązki, czyli wyodrębnienia wiązki o bardzo wąskim pasmie długości fali, dokonuje się zwykle przez odbicie wiązki od określonej płaszczyzny kryształu ustawionej pod braggowskim kątem w stosunku do skolimowanej wiązki polichromatycznej. Typowymi monochromatorami neutronów są monokryształy miedzi, cynku lub glinu. Monochromator jest umieszczony w osłonie z ołowiu lub żeliwa otoczonej następnie grubymi osłonami sporządzonymi z mieszaniny boraksu z parafiną oraz zewnętrzną osłoną wodną (rys. 1). Osłony pochłaniają promieniowanie γ oraz rozproszone przez monochromator neutrony termiczne i prędkie.



Rys. 1. Schemat dyfraktometru neutronów

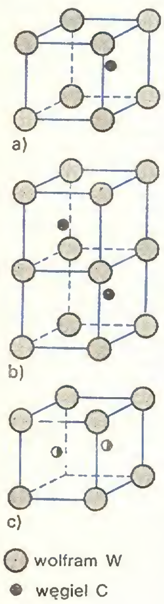
dyfraktometr neutronów

Przyrząd do otrzymywania widma neutronów ugiętych przez kryształ nosi nazwę dyfraktometru neutronów (rys. 1 i il. 83, tabl. 21). Jest on skonstruowany na tej samej zasadzie co dyfraktometr rentgenowski. Ze względu na duże rozmiary licznika neutronów oraz znaczny ciężar jego osłony, dyfraktometr neutronów ma większe rozmiary i jest cięższy niż dyfraktometr rentgenowski. Obrót stolika próbki dyfrakto-

metru jest sprzężony w stosunku 1:2 z ruchem ramienia, na którym znajduje się licznik. Zapewnia to zachowanie równości kąta padania i odbicia neutronów na próbkę badaną. Ruch stolika i ramienia odbywa się zwykle skokowo, np. co 5′ kąta rozproszenia.

Do badań strukturalnych wykorzystuje się z reguły monochromatyczną wiązkę neutronów. Neutrony polichromatyczne wychodzące na zewnątrz reaktora przechodzą zwykle przez kolimator umieszczony w kanale reaktora. Zadaniem kolimatora jest otrzymanie wiązki o określonym przekroju i rozbieżności. Kolimator ten powoduje również obniżenie tła instrumentalnego wywołanego neutronami prędkimi i promieniowaniem γ. Strumień neutronów monochromatycznych ze standardowych reaktorów doświadczalnych ma natężenie o kilka rzędów wielkości mniejsze niż strumień fotonów promieni rentgenowskich (z lampy rentgenowskiej), do badań neutronograficznych konieczne jest więc stosowanie próbek stosunkowo dużych: rzędu kilkunastu gramów (próbki polikrystaliczne) i kilkunastu miligramów (monokryształy).

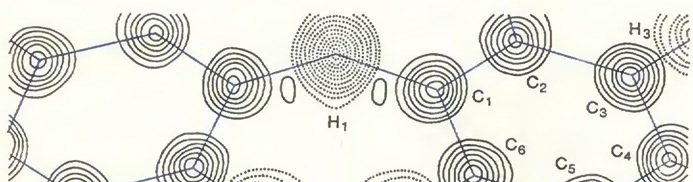
Stolik typowego dyfraktometru neutronów jest tak skonstruowany, aby mógł udźwignąć ciężar kilkuset kG — kriostatu, elektromagnesu itd. Ruchy zespołów dyfraktometru są zdalnie sterowane a pomiary całkowicie zautomatyzowane.



Rys. 2.

Bardzo prostym przykładem wykorzystania metod dyfrakcji neutronów do oznaczania struktury krystalicznej związku zawierającego obok siebie atomy o dużej i małej liczbie atomowej jest monowęglik wolframu (WC). Badania rentgenograficzne pozwoliły ustalić tylko położenie atomów wolframu. Są jednak możliwe aż 3 modele struktury WC różniące się położeniami atomów węgla, ale dające prawie jednakowe natężenia refleksów rentgenowskich. W procesie rozpraszania promieni rentgenowskich atomy wolframu odgrywają bowiem dominującą rolę, gdyż mają 74 elektrony a atomy węgla tylko 6 elektronów.

Jak wynika z przytoczonej poprzednio tabeli, w procesie rozpraszania neutronów termicznych jądra atomów węgla są silniejszymi ośrodkami rozpraszania niż jądra atomu wolframu. Wyniki badań neutronograficznych jednoznacznie pokazały, że monowęglik wolframu ma strukturę taką jak przedstawiona na rys. 2a z uporządkowanym rozkładem atomów węgla. Proponowane modele o strukturze typu arsenku niklu (rys. 2b), czy też ze statystycznie nieuporządkowanym rozkładem atomów węgla (rys. 2c) są nie do przyjęcia ze względu na całkowity brak zgodności między obliczonymi na ich podstawie natężeniami refleksów neutronowych a natężeniami obserwowanymi.



Rys. 3. Rozkład gęstości jądrowej w komórce elementarnej benzenu C_6H_6 wyznaczony metodą dyfrakcji neutronów

struktura benzenu

Dobrym przykładem ilustrującym wykorzystanie dyfrakcji neutronów jest oznaczenie struktury związku organicznego benzenu. Jądra wodoru mają ujemną amplitudę rozpraszania b. Strukturę C_6H_6 wyznaczono wykorzystując dane otrzymane dla próbki monokrystalicznej. Na rys. 3 jest pokazana fourierowska mapa gęstości jądrowej, analogiczna do mapy gęstości elektronowej omówionej w rozdziale „Strukturalna analiza kryształów". Kontury zaznaczone linią przerywaną dotyczą „ujemnej" gęstości jądrowej wynikającej z faktu, że b dla protonów ma znak ujemny. Tak więc, dopiero badanie neutronograficzne dostarczyło po raz pierwszy bezpośredniej doświadczalnej weryfikacji modelu Kekulégo budowy cząsteczki benzenu.

Metoda dyfrakcji neutronów wykorzystywana jest powszechnie do badania struktury krystalicznej zarówno związków nieorganicznych jak i organicznych i biologicznie czynnych, a także badania przemian fazowych zachodzących w ferroelektrykach, przemian porządek-nieporządek w metalach, a ponadto do identyfikacji składu fazowego materiałów.

Neutronografia magnetyczna

Oddziaływanie momentu magnetycznego neutronu z momentami magnetycznymi jonów w sieci kryształu prowadzi do zjawiska rozpraszania neutronów. W substancjach paramagnetycznych rozproszenie to jest niespójne i maleje ze wzrostem kąta rozpraszania, wskutek zależności od magnetycznego czynnika atomowego, określającego rozkład niesparowanych elektronów warunkujących moment magnetyczny jonów. W substancjach odznaczających się uporządkowaniem momentów magnetycznych, tzn. o określonej orientacji momentów magnetycznych w stosunku do elementów symetrii kryształu, czyli mających strukturę magnetyczną, rozpraszanie magnetyczne neutronów jest spójne i sprężyste. Na neutronogramach, łącznie z refleksami pochodzącymi od neutronów rozproszonych przez wszystkie jądra w komórce elementarnej kryształu, występują magnetyczne maksima interferencyjne (refleksy magnetyczne).

natężenie refleksu magnetycznego

Wykorzystując magnetyczne rozpraszanie neutronów uzyskuje się nieosiągalny innymi metodami bezpośredni wgląd w konfigurację momentów magnetycznych poniżej temperatury Curie lub Neéla (→ Teoria magnetyzmu). Śledząc zmianę natężenia refleksów magnetycznych w funkcji temperatury, można wyznaczyć krzywą namagnesowania, bowiem natężenie refleksu magnetycznego I_{hkl}^{M} jest proporcjonalne do kwadratu momentu magnetycznego jonu:

$$I_{hkl}^{M} \sim |\vec{q}|^2 |\vec{F}_{hkl}^{M}|^2.$$

Magnetyczna amplituda struktury

$$\vec{F}_{hkl}^{M} = \sum_{j=1}^{M} \vec{q}_j p_j e^{2\pi i (hx_j + ky_j + lz_j)},$$

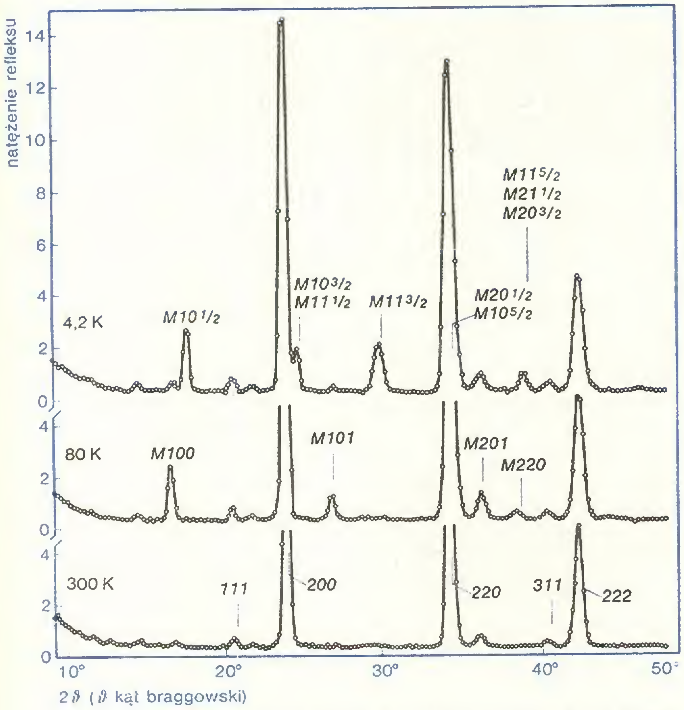
gdzie: $p_j = \dfrac{\gamma e^2}{2mc^2} f_j \mu_j = 0{,}2695 \cdot 10^{-12} f_j \mu_j$; e i m — ładunek i masa spoczynkowa elektronu, γ — moment magnetyczny neutronu, c — prędkość światła, f_j — atomowy czynnik magnetyczny jonu, μ_j — moment magnetyczny jonu (w magnetonach Bohra), $\vec{q}_j$ — wektor oddziaływania magnetycznego wiążący kierunek momentu magnetycznego jonu z wektorem prostopadłym do płaszczyzny odbijającej (hkl) ($|\vec{q}|^2 = \sin^2\alpha$, gdzie α jest kątem między tymi kierunkami) x_j, y_j, z_j — współrzędne jonu; sumowanie obejmuje jony paramagnetyczne w magnetycznej komórce elementarnej, których jest M. Gdy do badania wykorzystywane są niespolaryzowane neutrony, tj. gdy wszystkie kierunki ustawienia momentów magnetycznych neutronów są jednakowo prawdopodobne, wtedy rozpraszanie magnetyczne i jądrowe są względem siebie niespójne. Natężenie refleksu spowodowanego rozpra-

szaniem neutronów przez jądra i rozpraszaniem magnetycznym wyraża się zależnością:

$$I_{hkl} \sim |F_{hkl}|^2 + |\vec{q}|^2 |F^M_{hkl}|^2.$$

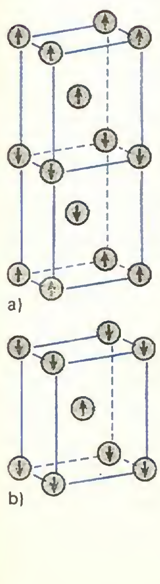
wyznaczanie struktury magnetycznej

W substancjach antyferromagnetycznych momenty magnetyczne jonów w sąsiednich komórkach elementarnych skierowane są antyrównolegle. Oznacza to, że co druga komórka elementarna jest identyczna. Elementarna komórka magnetyczna powstaje przez podwojenie w jednym kierunku komórki krystalograficznej. Na neutronogramach pojawiają się wtedy refleksy od „nadstruktury magnetycznej". Na rys. 4 przedstawione są neutronogramy monoarsenku uranu

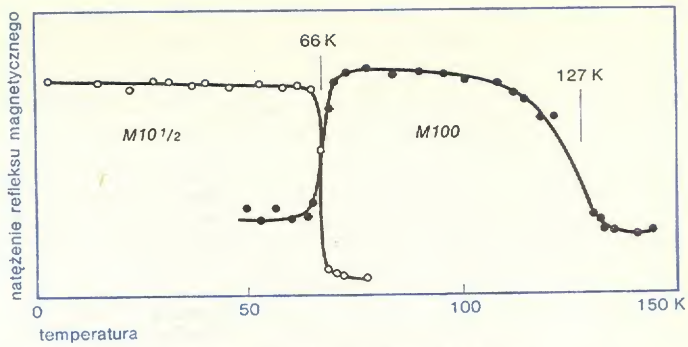


Rys. 4. Neutronogramy monoarsenku uranu otrzymane w temperaturze 4,2 K, 80 K i 300 K. Refleksy magnetyczne oznaczone zostały literą *M*. Wskaźniki połówkowe oznaczają, że magnetyczna komórka jest podwojona w jednym kierunku w stosunku do komórki krystalograficznej (struktura magnetyczna *A*)

(UAs) wykonane w temperaturze 4,2 K, 80 K, 300 K. Położenia refleksów magnetycznych wskazują, że w 4,2 K występuje w UAs uporządkowanie antyferromagnetyczne prowadzące do podwojenia komórki krystalograficznej w jednym tylko kierunku. Symetria komórki magnetycznej jest więc w temperaturze 4,2 K tetragonalna (struktura *A* na rys. 5). W temperaturze 80 K stwierdzono występowanie innego typu uporządkowania antyferromagnetycznego (struktura *B* na rys. 5) — również o symetrii tetragonalnej, chociaż komórka krystalograficzna i magnetyczna ma wówczas jednakowe rozmiary. Mierząc w funkcji temperatury natężenie refleksu magnetycznego oznaczonego wskaźnikiem $M10^1/_2$, charakterystycznego dla struktury typu *A* stwierdzono, że zanika on nagle w temperaturze 66 K (rys. 6). W tej samej temperaturze pojawia się refleks magnetyczny *M100* związany ze strukturą magnetyczną typu *B*. Zanika on w temperaturze 127 K (temperatura Neéla) — następuje przejście do stanu paramagnetycznego. Można więc neutronograficznie wyznaczyć nie tylko strukturę magnetyczną



Rys. 5. Struktura magnetyczna monoarsenku uranu w temperaturze 4,2 K (struktura magnetyczna *A*) i 80 K (struktura magnetyczna *B*). Kierunki momentów magnetycznych w obu wypadkach są równoległe do osi czterokrotnej — komórka magnetyczna ma symetrię tetragonalną



Rys. 6. Zależność od temperatury natężeń refleksów magnetycznych charakterystycznych dla struktury magnetycznej typu *A* ($M10^1/_2$) i typu *B* (*M100*). W temperaturze 66 K ma miejsce magnetyczna przemiana fazowa. Przejście do stanu paramagnetycznego zachodzi przy 127 K (punkt Neéla)

wyznaczanie punktu Neéla i Curie

w danej temperaturze, lecz również określić zakres temperatury stabilności poszczególnych faz magnetycznych. Prowadząc pomiary neutronograficzne otrzymuje się krzywe namagnesowania bez obecności zewnętrznego pola magnetycznego. Wyznaczyć można również wartości momentu magnetycznego w stanie uporządkowania antyferromagnetycznego.

Zastosowanie zjawiska dyfrakcji magnetycznej neutronów umożliwiło w 1946 r. pełne potwierdzenie modelu uporządkowania antyferromagnetycznego postulowanego przez Neéla jeszcze w 1934 r. Doprowadziło następnie do odkrycia bardziej złożonych typów uporządkowania magnetycznego, w tym helikoidalnego.

wyznaczanie struktury krystalicznej

Natężenie refleksów magnetycznych jest zależne od kierunku zewnętrznego pola magnetycznego przyłożonego do badanej próbki; przyłożenie pola magnetycznego o kierunku prostopadłym do płaszczyzny odbijającej prowadzi do całkowitego wygaszenia rozpraszania magnetycznego. Pozwala to na wyodrębnienie rozpraszania przez jądra od rozpraszania magnetycznego, a tym samym na wyznaczenie struktury krystalicznej substancji magnetycznej. Przykładając zewnętrzne pole magnetyczne równolegle do płaszczyzny rozpraszającej uzyskujemy maksymalne rozpraszanie magnetyczne ($|\vec{q}|^2 = 1$).

metoda spolaryzowanej wiązki

Przy użyciu wiązki spolaryzowanych neutronów (tzn. mających jednakowy kierunek spinów) rozproszenie przez jądra i rozpraszanie magnetyczne są spójne względem siebie. Metoda spolaryzowanej wiązki umożliwia wyznaczenie z dużą dokładnością udziału magnetycznego rozpraszania w natężeniu refleksu. Jest więc wykorzystywana głównie do wyznaczenia wartości magnetycznego czynnika atomowego i wartości momentu magnetycznego jonów paramagnetycznych tworzących substancje ferromagnetyczne.

G. E. Bacon *Neutron Diffraction*, Oxford 1975; J. Chojnacki *Metalografia strukturalna*, Katowice 1966; J. A. Izjumow, R. P. Ozjerow *Magnitnaja Niejtronografija*, Moskwa 1966. J. Leciejewicz *Zarys neutronografii kryształów*, Warszawa 1980; J. Leciejewicz *Wstęp do dyfraktometrii neutronów*, Warszawa 1979; J. Z. Nozik, R. P. Ozjerow, K. Hennig *Strukturnaja Niejtronografija*, Moskwa 1979.

Strukturalna analiza kryształów

Zofia Kosturkiewicz

Strukturalną analizą kryształów (rentgenowską analizą strukturalną kryształów) nazywa się zespół metod obliczeniowych prowadzących do uzyskania szczegółowego obrazu rozmieszczenia atomów w przestrzeni, na podstawie analizy zmierzonych intensywności wiązek promieni rentgenowskich odbitych przez płaszczyzny sieciowe kryształu. W celu określenia struktury całego kryształu wystarczy określić wymiary i symetrię jego komórki elementarnej i wzajemne rozmieszczenie atomów (cząsteczek) wewnątrz komórki.

Kierunki rozchodzenia się promieni ugiętych zależą od wielkości parametrów sieciowych i od symetrii kryształu (zob. rozdział „Krystalografia rentgenowska"). Natomiast natężenia promieni ugiętych zależą od tego, jakie atomy znajdują się w poszczególnych pozycjach komórki elementarnej, a więc od tego jak obsadzone są atomami płaszczyzny sieciowe (*hkl*).

pomiar intensywności refleksów

Dane o natężeniach ugiętych wiązek promieni rentgenowskich uzyskuje się metodami fotograficznymi, stosując odpowiednie kamery rentgenowskie, bądź też metodami licznikowymi w dyfraktometrach rentgenowskich. Przy stosowaniu fotograficznych metod rejestracji obrazu dyfrakcyjnego zaczernienie refleksów na rentgenogramach ocenia się (mierzy) następującymi sposobami: 1) wizualnie, porównując zaczernienie refleksów z kalibrowaną, wielostopniową skalą; wówczas błąd pomiaru jest jednak nie mniejszy niż 10%; 2) za pomocą mikrofotometru, w którym zaczernienie refleksów określa fotokomórka; 3) automatycznym mikrodensytometrem, rejestrującym zmierzone przez fotokomórkę intensywności refleksów na taśmie magnetycznej w minikomputerze, z którego otrzymuje się wydruk intensywności refleksów wraz z ich wskaźnikami; precyzja pomiarów przeprowadzonych tą metodą dorównuje precyzji metod licznikowych. Metody licznikowe pomiaru natężeń polegają na zliczaniu fotonów w ugiętej wiązce promieni za pomocą licznika Geigera–Müllera lub licznika proporcjonalnego czy scyntylacyjnego w dyfraktometrze rentgenowskim. Liczniki umożliwiają wielką precyzję pomiarów intensywności, błąd pomiaru bowiem nie przekracza 1%. Coraz większą popularnością cieszy się czterokołowy dyfraktometr do monokryształów.

Według kinematycznej teorii dyfrakcji promieni rentgenowskich na kryształach o budowie mozaikowej natężenie I ugiętej wiązki promieni jest proporcjonalne do kwadratu czynnika struktury F:

$$I_{hkl} = k|F_{hkl}|^2$$

(współczynnik proporcjonalności k jest związany m.in. z absorpcją promieniowania przez kryształ, drganiami cieplnymi atomów w krysztale i polaryzacją promieniowania rentgenowskiego).

czynnik struktury

Czynnik struktury zależy z kolei od rozmieszczenia atomów w komórce elementarnej i jest dany wzorem

$$F_{hkl} = \sum f_j e^{2\pi i (hx_j+ky_j+lz_j)},$$

gdzie x_j, y_j, z_j są współrzędnymi j-ego atomu w komórce elementarnej, wyrażonymi w ułamkach długości jej krawędzi, zaś czynnik atomowy f_j określa stosunek amplitudy fali ugiętej na j-ym atomie do amplitudy fali ugiętej przez pojedynczy elektron; jego wartość jest zależna od liczby atomowej j-ego atomu oraz od braggowskiego kąta ϑ.

czynnik atomowy

Czynnik struktury jest wynikiem sumowania j amplitud fal rentgenowskich ugiętych na j atomach, znajdujących się w komórce elementarnej, w trakcie odbijania tych promieni przez rodzinę płaszczyzn sieciowych (*hkl*) (rys. 1).

kąt przesunięcia fazowego

Każda z fal ugiętych na j-ym atomie ma amplitudę proporcjonalną do czynnika atomowego f_j i kąt przesunięcia fazowego (fazę) α w stosunku do fali ugiętej na atomie znajdującym się w początku układu współrzędnych. Ponieważ różnica faz między promieniami rentgenowskimi odbitymi od kolejnych równoległych płaszczyzn sieciowych (*hkl*) wynosi 2π (warunek Wulfa–Bragga), to różnica faz dla translacji jednostkowych (parametrów a, b, c) wzdłuż osi X, Y, Z wynosi $2\pi h$, $2\pi k$, $2\pi l$. Dla rodziny płaszczyzn sieciowych (*hkl*) różnica faz (przesunięcie fazowe) między punktem początkowym układu osi współrzędnych 000 i punktem x, y, z (każdy z tych dwóch punktów leży w płaszczyźnie sieciowej należącej do rodziny (*hkl*), rys. 2) równa się sumie różnic faz pomiędzy końcami wektorów równoległych do osi i łączących te dwa punkty:

$$\alpha = 2\pi(hx+ky+lz).$$

Czynnik struktury można przedstawić za pomocą jego modułu i fazy

$$F_{hkl} = |F_{hkl}|\exp(i\alpha_{hkl})$$

lub jako liczbę zespoloną

$$F_{hkl} = A_{hkl}+iB_{hkl},$$

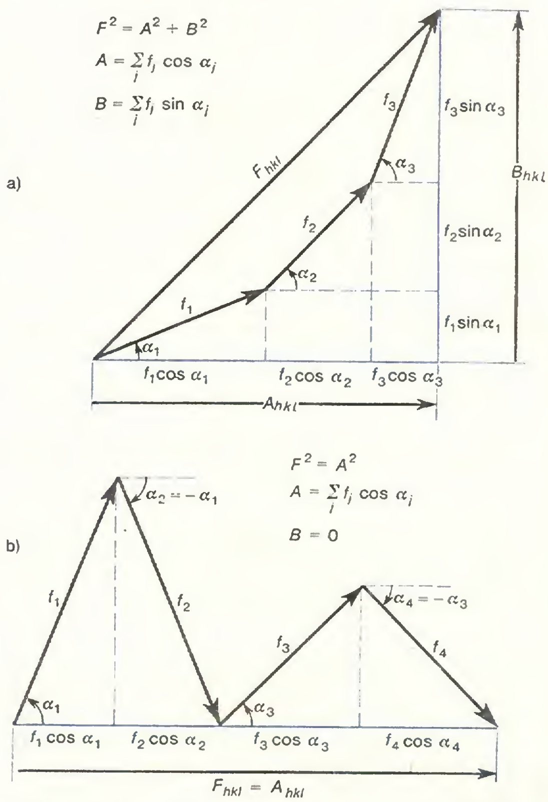
przy czym A i B można przedstawić (rys. 1) w postaci trygonometrycznej:

$$A_{hkl} = \sum_j f_j \cos 2\pi(hx_j+ky_j+lz_j),$$

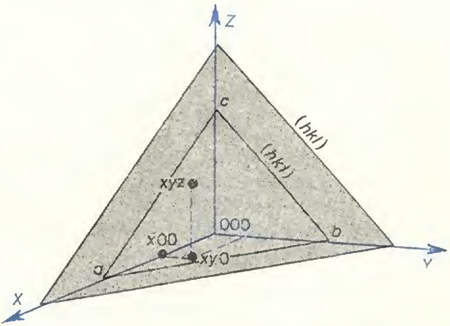
$$B_{hkl} = \sum_j f_j \sin 2\pi(hx_j+ky_j+lz_j).$$

Jeśli w kryształach znajduje się środek symetrii, to część urojona liczby zespolonej staje się równa zeru.

Gdyby była znana struktura kryształu, można by było obliczyć czynniki struktury, a więc i intensywność wiązek promieni rentgenowskich odbitych od wszystkich płaszczyzn sieciowych kryształu. W rent-



Rys. 1. Czynnik struktury F jako suma amplitud atomowych f_j: a) dla kryształu niecentrosymetrycznego, b) dla kryształu centrosymetrycznego



Rys. 2. Różnica faz pomiędzy punktami 000 i $x00$ wynosi $2\pi hx$, pomiędzy $x00$ i $xy0$ — $2\pi ky$ i pomiędzy $xy0$ i xyz — $2\pi lz$

cel analizy strukturalnej

genowskiej analizie strukturalnej kryształów zadanie jest jednak odwrotne: zmierzone są intensywności wiązek odbitych od różnych płaszczyzn sieciowych, czyli znane są moduły czynników struktury, a celem analizy strukturalnej jest określenie struktury kryształu.

Kryształ można traktować jako obszar wypełniony elektronami w sposób okresowy, przy czym okresami są parametry komórki elementarnej a, b, c. Zagęszczenie elektronów określa miejsce znajdowania się atomu. Jądra atomów są „niewidoczne" dla promieni rentgenowskich. Za pomocą dyfrakcji promieni rentgenowskich bada się rozkład chmur elektronowych. Liczba elektronów przypadająca na jednostkę objętości w punkcie komórki elementarnej o współrzędnych x, y, z jest dana przez następujące wyrażenie:

gęstość elektronowa

$$\varrho(xyz) = \frac{1}{V}\sum\sum_{h,k,l=-\infty}^{+\infty}\sum F_{hkl}e^{-2\pi i\,(hx+ky+lz)},$$

gdzie $\varrho(xyz)$ jest gęstością elektronową w punkcie x, y, z, a V — objętością komórki elementarnej.

trójwymiarowa synteza Fouriera

Obliczenie rozkładu gęstości elektronowej przez sumowanie podanego wyżej szeregu Fouriera nazywa się w analizie strukturalnej trójwymiarową syntezą Fouriera. Przeprowadzając syntezę Fouriera komórkę elementarną dzieli się na takie części (rys. 3), jakie pozwalają uniknąć przeoczenia jakichkolwiek atomów. Np. komórkę o wymiarach 10,5 × 10, 4 × 8,0 Å dzieli się odpowiednio na 40 × 40 × 30 części, tak aby gęstość elektronową można było obliczać w punktach odległych od siebie o ok. 0,25 Å. Taki podział komórki elementarnej stwarza konieczność sumowania szeregu Fouriera w 48 tys. punktów. Jeśli dla badanego kryształu zarejestrowanych zostało tylko 2 tys. natężeń refleksów (a rejestruje się i 4–6 tys.), to obliczenie jednej syntezy wymaga 96 mln (!) sumowań. Zwykle prowadzi się obliczenia tylko dla połowy lub czwartej części komórki elementarnej, w pozostałej bowiem części komórki rezultaty obliczeń są powtarzalne w wyniku operacji symetrii, niemniej ogrom obliczeń spowodował, że badania tego typu mogły się rozwinąć na szerszą skalę dopiero w erze komputerów.

Niekiedy można zmniejszyć ilość obliczeń w ten sposób, że na podstawie danych o natężeniach refleksów z jednej warstwicy (zawierającej np. refleksy $hk0$) oblicza się rzut gęstości elektronowej wzdłuż jednej z osi krystalograficznych na płaszczyznę prostopadłą do tej osi (rys. 4 i 5). Jest to dwuwymiarowa synteza Fouriera

dwuwymiarowa synteza Fouriera

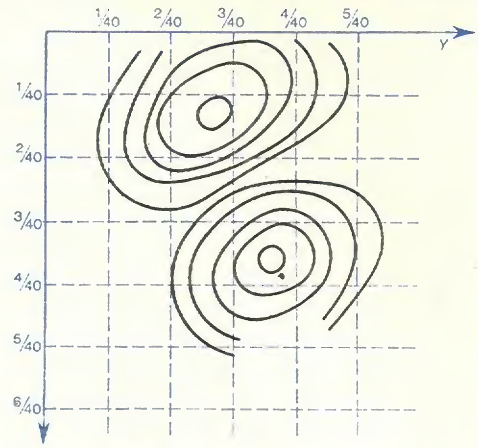
$$\varrho(xy) = \frac{1}{A}\sum_{h,k=-\infty}^{\infty+}\sum F_{hk0}\,e^{-2\pi i\,(hx+ky)}$$

(A jest powierzchnią płaszczyzny rzutu prostopadłej do osi Z). W dwuwymiarowej syntezie Fouriera oblicza się gęstość elektronową, np. w 1600 punktach (40 × 40) na podstawie ok. 200 refleksów z jednej warstwicy zerowej. Zmniejsza to ilość obliczeń do 320 tys. sumowań dla jednej syntezy. Wykonane w ten sposób dwa rzuty wzdłuż dwóch osi dają obraz trójwymiarowy. Wynik jest jednak mniej dokładny niż w syntezie trójwymiarowej. Syntezy Fouriera przedstawia się graficznie jako mapy gęstości elektronowej, na których punkty o jednakowej gęstości elektronowej łączy się liniami ciągłymi w warstwice (rys. 3 i 5).

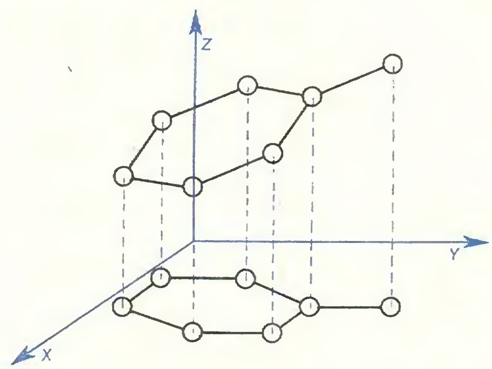
mapy gęstości elektronowej

Jednak nie pracochłonność obliczeń stanowi największą trudność w badaniach strukturalnych. Podstawową trudnością jest problem fazowy. W doświadczeniach otrzymuje się intensywności refleksów I_{hkl}, z których wyznacza się czynniki struktury, lecz nie można określić kąta fazowego α. Podkreślmy więc, że doświadczalnie uzyskuje się tylko $|F_{hkl}|$ — tj. wartości bezwzględne czynników struktury, a nie poznaje się ich kątów fazowych. Problem fazowy musi być rozwiązywany w badaniach strukturalnych na podstawie dość złożonych analiz. Obecnie stosuje się głównie dwa sposoby rozwiązywania problemu fazowego. Są nimi analiza funkcji Pattersona (nazywanej też funkcją wektorów międzyatomowych) oraz metody bezpośrednie, wykorzystujące statystyczne zależności między fazami czynników strukturalnych. Dawniej

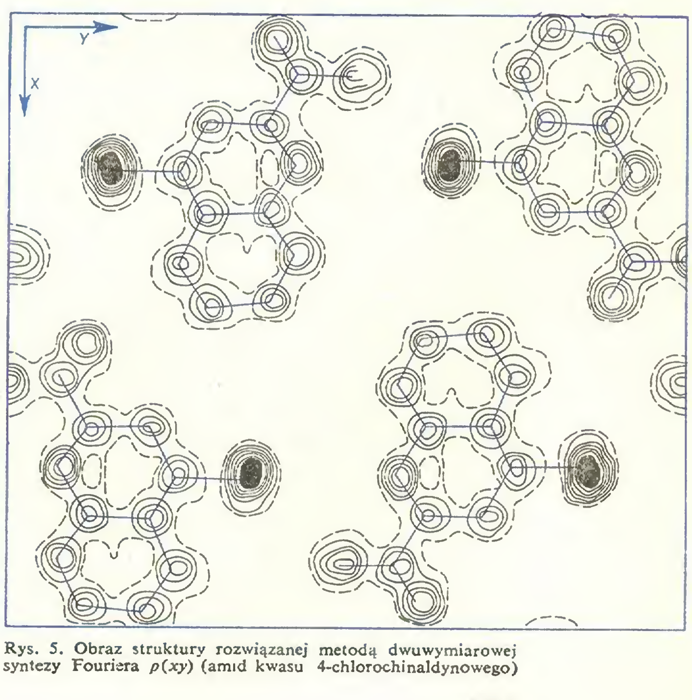
problem fazowy



Rys. 3. Fragment komórki elementarnej z uwidocznionym podziałem krawędzi komórki elementarnej na części umożliwiające lokalizację atomów



Rys. 4. Metoda rzutu dwuwymiarowego — synteza $\rho(xy)$



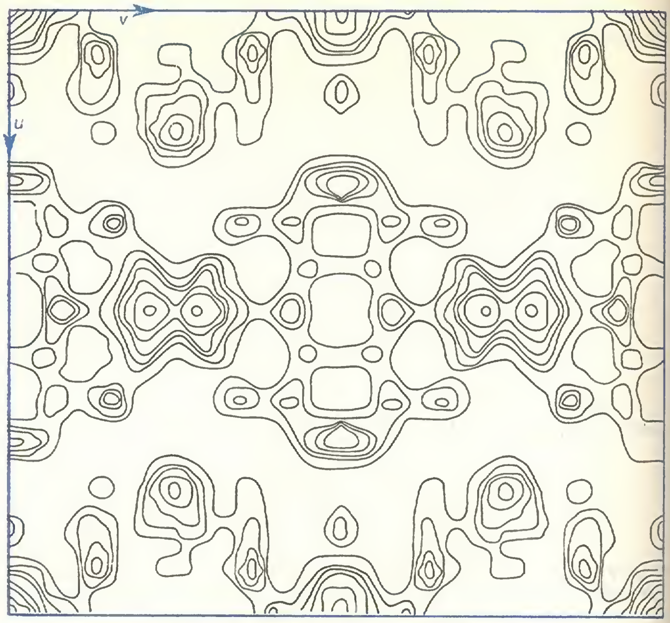
Rys. 5. Obraz struktury rozwiązanej metodą dwuwymiarowej syntezy Fouriera $\rho(xy)$ (amid kwasu 4-chlorochinaldynowego)

metoda prób i błędów

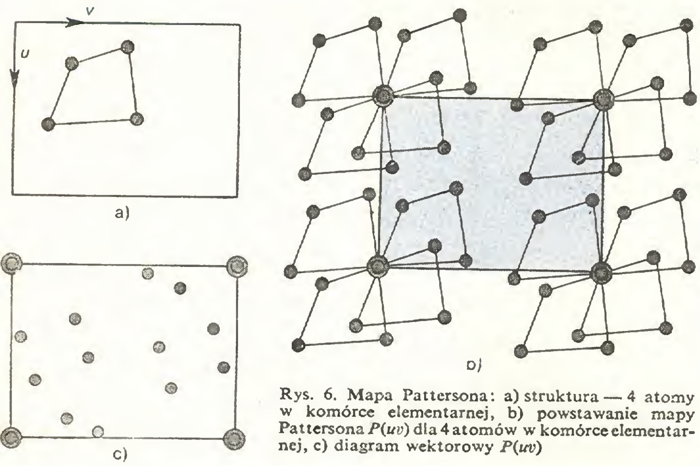
stosowano jeszcze metodę prób i błędów, polegającą na tym, że odgadywano przybliżoną strukturę, obliczano dla niej F_{hkl} i porównywano wartości obliczonych czynników struktury F_c ze zmierzonymi doświadczalnie czynnikami F_o. Jeśli czynniki F_c i F_o były w miarę zgodne, uznawano strukturę za rozwiązaną. Metoda ta mogła być jednak stosowana tylko do najprostszych struktur kryształów, w których atomy lub jony zajmują położenia szczególne — w narożach lub na środkach ścian czy krawędzi komórki elementarnej (jak np. w diamencie i NaCl). O takim rozmieszczeniu wnioskuje się łatwo np. na podstawie liczby atomów znajdujących się w komórce elementarnej, wielokrotnie mniejszej niż liczba zdeterminowana przez symetrię kryształu. Wówczas gdy atomy znajdują się w położeniu dowolnym, odgadnięcie prawidłowej struktury jest niemożliwe, gdyż zmiana o ułamek angstrema położenia w komórce elementarnej cząsteczki nawet o znanym kształcie (a często i pojedynczego atomu) zmienia drastycznie wartości czynników struktury, dając negatywny wynik porównania F_o i F_c.

synteza Pattersona

Najczęściej stosowaną metodą rozwiązywania problemu fazowego jest synteza Pattersona. Polega ona na tym, że do wzoru na gęstość elektronową ϱ zamiast czynnika struktury F_{hkl} o nieznanym znaku lub



Rys. 7. Przykład dwuwymiarowego rozkładu funkcji Pattersona $P(uv)$ (amid kwasu 4-chlorochinaldynowego)



Rys. 6. Mapa Pattersona: a) struktura — 4 atomy w komórce elementarnej, b) powstawanie mapy Pattersona $P(uv)$ dla 4 atomów w komórce elementarnej, c) diagram wektorowy $P(uv)$

fazie wstawia się uzyskaną bezpośrednio w doświadczeniu wartość F^2_{hkl}, której nie potrzeba przypisywać znaku lub fazy. Funkcję Pattersona w punkcie uvw komórki elementarnej określa więc wyrażenie

$$P(uvw) = \frac{1}{V} \sum_{h,k,l=-\infty}^{+\infty}\sum\sum |F_{hkl}|^2 e^{-2\pi i\,(hu+kv+lw)}$$

lub w postaci trygonometrycznej

$$P(uvw) = \sum_{h,k,l=-\infty}^{+\infty}\sum\sum |F_{hkl}|^2 \cos 2\pi(hu+kv+lw).$$

Rozkład funkcji Pattersona, obliczony w sposób analogiczny do opisanego dla rozkładu gęstości elektronowej, prezentuje nie strukturę, lecz położenia wektorów łączących każde dwa atomy w strukturze (wektorów międzyatomowych), poprowadzonych z początku układu współrzędnych. Każdemu maksimum na rozkładzie funkcji Pattersona odpowiadają w komórce elementarnej kryształu dwa atomy o współrzędnych x_i, y_i, z_i oraz x_j, y_j, z_j. Tak więc $u = x_i - x_j$, $v = y_i - y_j$, $w = z_i - z_j$. Jeżeli w komórce elementarnej znajduje się N atomów, to na rozkładzie funkcji Pattersona wystąpi $N(N-1)$ maksimów, często nakładających się na siebie.

wektory międzyatomowe

Niezależnie od symetrii kryształu, rozkład funkcji Pattersona jest zawsze centrosymetryczny, ponieważ dowolna para atomów może być wiązana dwoma równoległymi, lecz przeciwnie skierowanymi wektorami. Syntezę Pattersona przedstawia się graficznie jako mapę (diagram wektorowy) z przeprowadzonymi na niej warstwicami (rys. 7). W punkcie początkowym układu osi współrzędnych znajduje się zawsze silne maksimum, którego wielkość jest proporcjonalna do sumy kwadratów liczb atomowych $Z_1, Z_2, ..., Z_i, Z_j$ atomów tworzących daną strukturę:

diagram wektorowy

$$H \sim \sum Z_j^2.$$

Jeśli wektor uvw w tzw. przestrzeni Pattersona łączy dwa atomy o dużych liczbach porządkowych Z_{j1}, Z_{j2} silnie rozpraszające promieniowanie (tzw. atomy ciężkie), to funkcja Pattersona ma silne maksimum w punkcie u, v, w. Wielkość tego maksimum jest proporcjonalna do iloczynu $Z_{j1} \cdot Z_{j2}$. Orientację i długość takiego wektora można z łatwością odczytać z mapy (diagramu) Pattersona, co pozwala zlokalizować położenia najcięższych atomów w komórce elementarnej.

Rozszyfrowanie syntezy Pattersona polega na znalezieniu struktury, z której ta synteza powstała. Wśród wielu maksimów pattersonowskich należy więc znaleźć te, które tworzą motyw struktury. Do rozszyfrowania syntez Pattersona wykorzystuje się obecnie głównie metody superpozycyjne, a spośród nich metodę minimalizacji, która w korzystnej sytuacji wydziela z syntezy Pattersona pełny obraz struktury.

metoda minimalizacji

Funkcję Pattersona można zastosować do określenia symetrii kryształów. Analizując liczbę maksimów na mapie Pattersona, można stwierdzić obecność lub brak środka symetrii w krysztale. Badając wielkość pattersonowskich maksimów oraz skupienia maksimów na pewnych liniach lub płaszczyznach funkcji wektorów międzyatomowych, zamiast 122 grup dyfrakcyjnych (zob. rozdział „Krystalografia rentgenowska") można jednoznacznie wyznaczyć aż 208 (!) grup przestrzennych.

Jeżeli w komórce elementarnej kryształu występują równocześnie atomy ciężkie i lekkie, to z rozkładu funkcji Pattersona odnajduje się zwykle tylko współrzędne x, y, z atomów ciężkich. Współrzędne te pozwalają na obliczenie wartości czynników F_{hkl} dla struktury złożonej tylko z tych atomów. Tak obliczony zbiór wartości F_c różni się oczywiście od wartości F_o określonych doświadczalnie dla badanej struktury, ale niemniej umożliwia dalsze postępowanie polega-

jące na tym, że doświadczalnie uzyskanym wartościom F_o przypisuje się kąty fazowe α, obliczone dla pozycji x, y, z atomów ciężkich i za ich pomocą oblicza się pierwsze przybliżenie rozkładu gęstości elektronowej $\varrho(xyz)$ (synteza Fouriera). Jeśli w krysztale występuje środek symetrii, to część urojona wyrażenia $F_{hkl} = A_{hkl}+iB_{hkl}$ równa się 0 (wtedy bowiem kąt fazowy wynosi 0° lub 180°, a jego cosinus równa się +1 lub −1) i wówczas problem fazowy sprowadza się do przypisania każdej wartości $|F_{hkl}|$ znaku + lub −.

metoda kolejnych przybliżeń

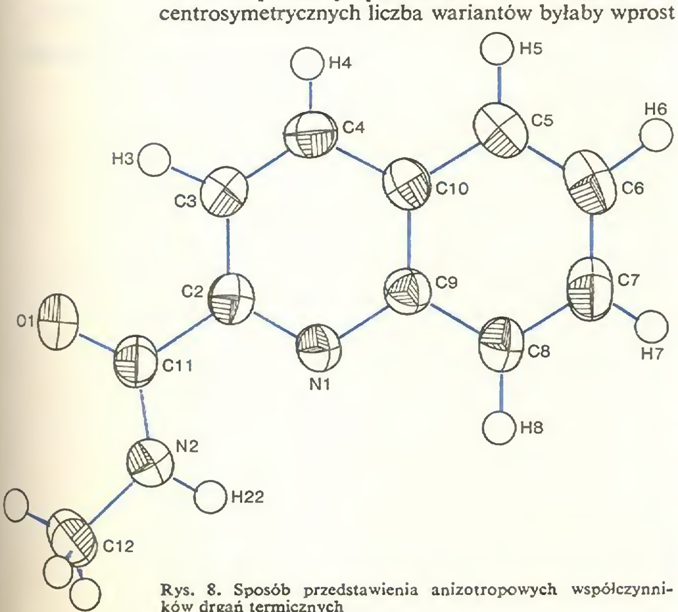
Gdy atom ciężki ma odpowiednio dużą liczbę atomową w stosunku do sumy liczb porządkowych atomów lekkich, to pierwsze przybliżenie syntezy Fouriera wyjaśnia niemal całą strukturę. Strukturę tę udokładnia się dalej metodą kolejnych przybliżeń szeregów Fouriera, polegającą na powtarzaniu opisanej wyżej procedury.

Najlepszą metodą uzyskiwania bezpośrednich informacji o kątach fazowych czynników struktury jest badanie par kryształów izomorficznych. W jednym z tych kryształów podstawia się na miejsce określonego atomu — atom ciężki. Takie dwa kryształy, o zbliżonym składzie chemicznym, krystalizujące w tej samej grupie przestrzennej i mające zbliżone rozmiary komórek elementarnych oraz takie same współrzędne odpowiadających sobie atomów — nazywają się kryształami izomorficznymi. Różnica intensywności tych samych refleksów (tj. refleksów o tych samych wskaźnikach) dla obydwu kryształów pozwala oznaczyć kąty fazowe (znaki) czynników struktury F_{hkl}, co umożliwia obliczenie $\varrho(xyz)$.

metody bezpośrednie

Metody bezpośrednie oznaczania kątów fazowych (znaków) są oparte na założeniu, że gęstość elektronowa w krysztale nie jest nigdy ujemna i że jest bliska zeru poza miejscami znajdowania się atomów. W metodach tych, za pomocą równań statystycznie prawdopodobnych ustala się zależności pomiędzy fazami czy znakami czynników struktury.

Problemu fazowego nie można rozwiązać metodą prób i błędów. Jeżeli bowiem dla danej struktury zmierzono intensywności N refleksów, to zmieniając znak każdego kolejnego czynnika struktury, korzystając ze wzoru na gęstość elektronową $\varrho(xyz)$ można obliczyć 2^N wariantów rozkładów gęstości elektronowej, przy czym tylko jeden z nich będzie przedstawiał prawdziwą strukturę. Postępując w ten sposób, dla 10 refleksów należałoby obliczyć 1024 rozkładów, dla 20 czynników struktury 1 048 576 (!) rozkładów, a badając struktury kryształów mierzy się intensywności dla ponad tysiąca refleksów. Dla struktur niecentrosymetrycznych liczba wariantów byłaby wprost



Rys. 8. Sposób przedstawienia anizotropowych współczynników drgań termicznych

nieskonczona, a ich mechaniczne obliczanie byłoby pozbawione sensu.

Jaki jest więc tok postępowania? Na początku wybiera się tylko najbardziej intensywne refleksy (dwukrotnie silniejsze od średniej arytmetycznej wszystkich intensywności) i przez obliczenia arytmetyczne wyprowadza się zależności między fazami lub znakami; słuszność tych zależności jest tym prawdopodobniejsza, im wyższe są wartości bezwzględne czynników struktury wchodzących w relacje. Dla struktur niecentrosymetrycznych najczęściej jest stosowana następująca zależność między kątami fazowymi:

$$\alpha F_{hkl} \approx \alpha F_{h'k'l'}+\alpha F_{h-h',\,k-k',\,l-l'}$$

a dla struktur centrosymetrycznych — zależność między znakami s:

$$sF_{hkl} \approx sF_{h'k'l'}\cdot sF_{h-h',\,k-k',\,l-l'}.$$

Wynika stąd, że np. $sF_{2k00} = sF_{100}\cdot sF_{100}$, a więc warunek, by czynniki struktury o wskaźnikach parzystych — były dodatnie. Następny krok, po znalezieniu modelu struktury na podstawie najsilniejszych refleksów, polega na zastosowaniu metody kolejnych przybliżeń obliczania gęstości elektronowej przy korzystaniu już z pełnego zbioru F_{hkl}.

Metody bezpośrednie rozwiązywania problemu fazowego stają się coraz bardziej popularne, gdyż służą do badania struktur tych kryształów, w których wszystkie atomy mają zbliżone liczby atomowe, co uniemożliwia stosowanie metody Pattersona.

wskaźnik rozbieżności

Kryterium poprawności określenia struktury stanowi wskaźnik rozbieżności

$$R = \sum\big||F_o|-|F_c|\big|\Big/\sum|F_o|.$$

Im mniejszy jest wskaźnik rozbieżności, tym większa jest dokładność rozwiązania struktury. Przy wyznaczeniu intensywności refleksów metodami fotograficznymi wskaźnik rozbieżności wynosi ok. 0,1 (10%), a przy stosowaniu licznikowych metod pomiaru natężeń refleksów wartość jego zmniejsza się do 0,07–0,03 (tj. 7–3%). Po ustaleniu na podstawie syntez Fouriera przybliżonych współrzędnych wszystkich atomów, dalsze udokładnienie parametrów pozycyjnych atomów oraz wielkości drgań termicznych atomów przeprowadza się metodą najmniejszych kwadratów. Funkcja, która podlega minimalizacji, ma postać $\sum_{r=1}^{m} w_r(F_{or}-F_{cr})^2$, gdzie w jest wagą statystyczną przypisywaną dokładności pomiaru natężenia poszczególnych refleksów F_o, a F_c — czynnikiem struktury F_{hkl} obliczonym dla udokładnianego modelu struktury.

uwzględnianie drgań termicznych

W najprostszym modelu struktury przyjmuje się występowanie izotropowych (tzn. jednakowych we wszystkich kierunkach) drgań termicznych atomów i dla każdego atomu oblicza się tzw. izotropowe indywidualne czynniki temperaturowe. Drgania termiczne atomów wzrastają szybko, gdy temperatura kryształu jest bliska temperaturze topnienia i powodują zmniejszenie intensywności refleksów ze wzrostem kąta odbłysku ϑ. Zależność czynnika atomowego f od izotropowego czynnika temperaturowego B wyraża wzór $f = f_0 e^{-B(\sin\vartheta/\lambda)^2}$ (λ — długość fali promieniowania rentgenowskiego).

izotropowy czynnik temperaturowy

Ruchy termiczne atomów można analizować dokładniej, opisując je za pomocą elipsoidy trójosiowej, której osie określają wielkości drgań atomu w trzech prostopadłych do siebie kierunkach. Zależność czynnika atomowego f od anizotropowego czynnika temperaturowego wyraża wzór

anizotropowy czynnik temperaturowy

$$f = f_0 e^{-(b_{11}h^2+b_{22}k^2+b_{33}l^2+b_{12}hk+b_{23}kl+b_{31}hl)}.$$

Sześć składowych tensora drgań anizotropowych b_{ij} określa wymiary półosi elipsoidy oraz orientację półosi względem przyjętego układu osi współrzędnych. Ten sposób przedstawiania ruchów termicznych atomów ilustruje rys. 8.

Ostatnie stadium analizy stanowi obliczenie odchyleń standardowych dla współrzędnych *x*, *y*, *z* atomów, ocena błędów przypadkowych i systematycznych popełnionych przy pomiarze natężeń refleksów. Na podstawie współrzędnych wszystkich atomów oblicza się długości wiązań, kąty walencyjne, określa wzajemne ułożenie cząsteczek lub innych elementów strukturalnych w krysztale.

Analiza strukturalna daje obraz struktury uśrednionej w czasie. Wielkości drgań termicznych „widzimy" jako wielkość rozmycia położenia atomu. Metodami analizy strukturalnej określono struktury zarówno o budowie najprostszej (np. metali), jak i najbardziej złożonej (białka i tRNA).

G. H. Stout, L. H. Jensen, *X-ray Structure Determination*, Toronto 1969.

Mikroskopia elektronowa

Tadeusz Warmiński

Mikroskopia elektronowa jest skrótowym określeniem szerokiego zespołu zagadnień badawczych i konstrukcyjnych związanych z coraz powszechniejszym wykorzystaniem w naukach przyrodniczych, medycznych i technicznych grupy przyrządów elektronowo-optycznych zwanych mikroskopami elektronowymi. Ich wspólną cechą charakterystyczną jest zdolność do projekcji na ekranie i rejestracji powiększonych obrazów badanych przedmiotów (próbek lub preparatów) przy wykorzystaniu jako źródła oświetlenia skolimowanej wiązki elektronów o energii od kilku set eV do kilku MeV. Skonstruowanie pierwszego mikroskopu elektronowego typu transmisyjnego (M. Knoll i E. Ruska, 1931 r.), który powstał przez analogię z prześwietleniowym mikroskopem optycznym, było logiczną konsekwencją doświadczalnego potwierdzenia natury falowej elektronu (C. J. Davisson, L. H. Germer, G. P. Thomson, 1927 r.). Zgodnie z teorią Abbego należało oczekiwać wysokiej zdolności rozdzielczej takiego mikroskopu, ponieważ

Prędkości, masy i długości fali elektronu w zależności od jego energii (m_0 — masa spoczynkowa)

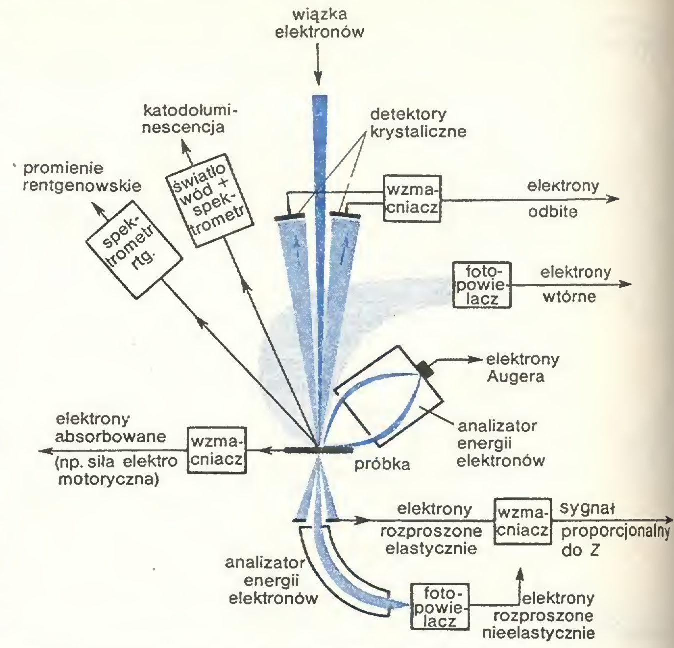
W, eV	v, m/s	v/c	$m\times10^{31}$,kg	m/m_0	λ, Å
1	$5{,}93\cdot10^5$	0,002	9,11	$1+2\cdot10^{-6}$	12,3
10	$1{,}87\cdot10^6$	0,006	9,11	$1+2\cdot10^{-5}$	3,87
10^2	$5{,}93\cdot10^6$	0,02	9,11	1,0002	1,32
10^3	$1{,}86\cdot10^7$	0,06	9,13	1,0020	0,386
10^4	$5{,}83\cdot10^7$	0,19	9,29	1,0196	0,122
10^5	$1{,}64\cdot10^8$	0,55	10,9	1,196	0,0369
10^6	$2{,}83\cdot10^8$	0,94	26,9	2,96	0,00868
10^7	$2{,}99\cdot10^8$	0,999	187,8	20,6	0,00118

wysokoenergetycznym elektronom odpowiadają długości fali znacznie krótsze aniżeli w wypadku światła.

zdolność rozdzielcza

Czynnikiem znacznie ograniczającym zdolność rozdzielczą soczewek elektromagnetycznych (niezależnie od zjawiska dyfrakcji elektronów) jest zła zdolność ogniskowania wiązek elektronów biegnących pod kątem α do osi optycznej. Dopuszczalne wartości tego kąta nie przekraczają zazwyczaj 10^{-3}–10^{-4} rad. Osiągalne zdolności rozdzielcze współczesnych mikroskopów elektronowych wynoszą 2–3 Å i są porównywalne z graniczną rozdzielczością soczewki magnetycznej wynikającą z rozważań teoretycznych ($\delta = 0{,}4c_s\alpha^3+0{,}61\lambda/\sin\alpha$; c_s — współczynnik aberracji sferycznej). Jest to niewątpliwa przewaga mikroskopu elektronowego nad optycznym, którego zdolność rozdzielcza nie przekracza 2000 Å.

W ostatnim dziesięcioleciu okazało się jednak, że nie tylko możliwość uzyskiwania bardzo wysokich powiększeń rozstrzyga o atrakcyjności mikroskopów elektronowych. Przyjrzyjmy się efektom oddziaływania skolimowanej wiązki elektronów z próbką materii, zazwyczaj ciała stałego. Przestrzeń wokół próbki wypełnia się wówczas sygnałami elektronowymi i elektromagnetycznymi, zaznaczonymi schematycznie na rys. 1. Każdy z nich niesie określonego typu informację o strukturze materiału próbki (w sensie przestrzennym lub energetycznym), dotyczącą



Rys. 1. Sygnały wzbudzane w próbce wiązką elektronów

obszaru oddziaływań. Wybór rejestrowanego sygnału określa zakres dostępnych informacji (tabela); biolog może być zainteresowany innego typu obrazem mikroskopowym aniżeli fizyk czy metalurg.

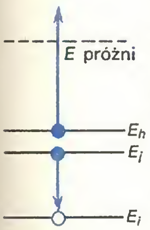
Zależność informacji od wyboru sygnału

Treść informacji	Rodzaj sygnału
Morfologia	wszystkie sygnały z wyjątkiem promieni rentgenowskich i elektronów Augera
Skład chemiczny (kompozycja)	promienie rentgenowskie, katodoluminescencja, elektrony odbite i elektrony Augera
Krystalografia	elektrony odbite, elektrony przechodzące (transmitowane) przez próbkę, elektrony wtórne, promienie rentgenowskie
Wiązania chemiczne	elektrony Augera, promienie rentgenowskie
Własności elektromagnetyczne	elektrony wtórne, elektrony absorbowane w próbce, siła elektromotoryczna

Stosując odpowiedni, różny dla różnych sygnałów, system generacji i detekcji wybranego sygnału otrzymuje się obraz mikroskopowy lub, inaczej, mapę stanowiącą czasoprzestrzennie uporządkowany zapis jego intensywności. Podstawowym warunkiem właściwego odczytu obrazu mikroskopowego jest znajomość procesów fizycznych zachodzących w próbce podczas bombardowania jej wysokoenergetycznymi elektronami. Na szybki rozwój mikroskopii elektronowej miał więc niewątpliwy wpływ znaczny postęp w zakresie fizyki ciała stałego, gazów, elektronów i promieniowania rentgenowskiego. Trzeba pamiętać, że z punktu widzenia mikroskopii elektronowej są ważne nie tylko procesy elastycznych rozproszeń elektronów w próbce, ale również rozproszeń nieelastycznych, prowadzących do zmniejszenia ich pierwotnej energii. Straty energii dzielą się statystycznie w następujący sposób: w energię termiczną zostaje przekształcone średnio 95% utraconej energii, w energię promieniowania rentgenowskiego (ciągłego i charakterystycznego) — ok. 1%, elektrony reemitowane (wtórne, odbite i Augera) unoszą 4% energii.

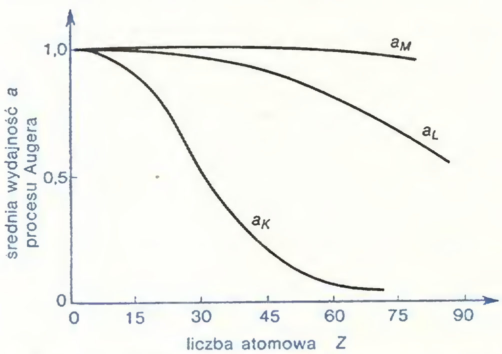
emisja elektronu Augera

Jonizacja atomów elektronami lub promieniami rentgenowskimi powoduje powstawanie luk na wewnętrznych powłokach elektronowych. Elementarnym procesem relaksacji jest zapełnienie luki na poziomie energetycznym E_i przez elektron z niższej energetycznie powłoki E_j. Różnica energii E_{ij} może opuścić atom unoszona przez foton charakterystycznego promieniowania rentgenowskiego, ale także w postaci bezpromienistej (emisja elektronu Augera). Elektron Augera wyrzucony z powłoki (lub podpowłoki) E_h ma energię $E_i - E_j - E_h$, która podobnie jak energia fotonu jest wielkością charakteryzującą atom danego pierwiastka chemicznego. Relaksacja atomu przez emisję elektronu Augera jest pośrednim, ale nie końcowym, procesem prowadzącym do stanu równowagi elektrycznej, ponieważ pozostawia atom podwójnie zjonizowany z lukami na poziomach E_j oraz E_h. Wydajność każdego z dwóch wzajemnie konkurencyjnych procesów określają liczbowo współczynniki prawdopodobieństwa zwane wydajnością fluorescencyjną i wydajnością procesu Augera, a w sumie równe jedności. Wydajność fluorescencyjną można zapisać w postaci ilorazu zawierającego w liczniku sumę fotonów wszystkich linii promieniowania rentgenowskiego danej serii, emitowanych statystycznie w jednostce czasu, a w mianowniku liczbę luk na poziomie E_i generowanych w tym samym czasie. Na przykład dla serii K iloraz ten ma postać:



$$\omega_K = \sum_j \frac{(n_K)_j}{N_K} = \frac{n_{K\alpha} + n_{K\beta} + \ldots}{N_K}.$$

Istnieje wyraźna zależność wydajności od liczby atomowej pierwiastka Z. Stwierdzono, że pod względem wydajności proces



Zależność wydajności procesu Augera od liczby atomowej Z

Augera dominuje w pierwiastkach lekkich ($Z < 15$) zjonizowanych na powłoce K oraz we wszystkich pierwiastkach mających luki na dalszych powłokach. Czynnikiem ograniczającym dotychczasowy rozwój spektroskopii elektronów Augera w wypadku ciała stałego jest konieczność stosowania wysokiej próżni. Dla większości pierwiastków energie elektronów Augera mają wartości od 50 do 2000 eV, co określa maksymalną grubość warstwy powierzchniowej próbki poddawanej analizie na ok. 10–15 Å. Czystość powierzchni jest więc w spektroskopii Augera sprawą bardzo istotną. Taka czystość jest nieosiągalna w otoczeniu gazów resztkowych o ciśnieniu wyższym niż 10^{-6} Pa.

tworzenie obrazu mikroskopowego

Obecnie istnieją dwa systemy tworzenia obrazu mikroskopowego, niezależnie od wyboru sygnału. W systemie konwencjonalnym zakłada się równoczesne oświetlenie względnie szeroką wiązką elektronów (kilka μm) całego pola obserwacji, przedstawianego w obrazie mikroskopowym. Tym samym wszystkie elementy obrazu mikroskopowego powstają również równocześnie i, pomijając możliwe zmiany próbki w trakcie obserwacji, proces tworzenia obrazu nie jest w tym systemie funkcją czasu. W systemie skaningowym wiązka elektronów „omiata" obszar obserwacji w sposób dokładnie taki sam, jak w lampie kineskopowej, która zazwyczaj spełnia funkcję czytnika w układzie detekcyjnym mikroskopu elektronowego. Ruch wiązki elektronowej po próbce jest zsynchronizowany z ruchem plamki świetlnej poruszającej się po ekranie lampy kineskopowej. Zasada systemu skaningowego polega na rozbiciu obszaru obserwacji na drobne części składowe i kolejnym ich odwzorowywaniu w obrazie mikroskopowym, przy czym każdej części odpowiada tylko jeden element obrazu. System zakłada czasowoprzestrzenną korelację przekazu informacji między próbką i czytnikiem obrazu (ekranem lampy kineskopowej, samopisem X–Y itp.). Zdolność rozdzielcza w systemie skaningowym zależy od stopnia kolimacji wiązki elektronowej oświetlającej próbkę, ale także od geometrycznych rozmiarów mikroobszaru, wewnątrz któ-

zdolność rozdzieloza w systemie skaningowym

rego jest generowany określony sygnał (tabela). Wysoką rozdzielczość uzyskuje się, gdy wiązka elektronów jest możliwie silnie skolimowana. Przy zasto-

Zestawienie osiągalnych wartości zdolności rozdzielczej w systemie skaningowym

δ, Å	Rodzaj sygnału
1–10	elektrony transmitowane przez próbkę
$5 \cdot 10^2$	elektrony Augera
10^2	elektrony wtórne
$5 \cdot 10^2$–10^4	elektrony odbite
10^4–10^5	elektrony absorbowane w próbce
10^3–$5 \cdot 10^4$	promieniowanie rentgenowskie
10^3–10^5	katodoluminescencja

sowaniu termoemisyjnych katod wolframowych i dwu- lub trójstopniowego układu ogniskowania przekrój wiązki elektronów o natężeniu 10^{-12} A wynosi 80–100 Å, natomiast przy zastosowaniu bardzo jasnych, samoogniskujących wyrzutni polowoemisyjnych (autoemisja z katody ostrzowej) można go zmniejszyć do kilkunastu Å.

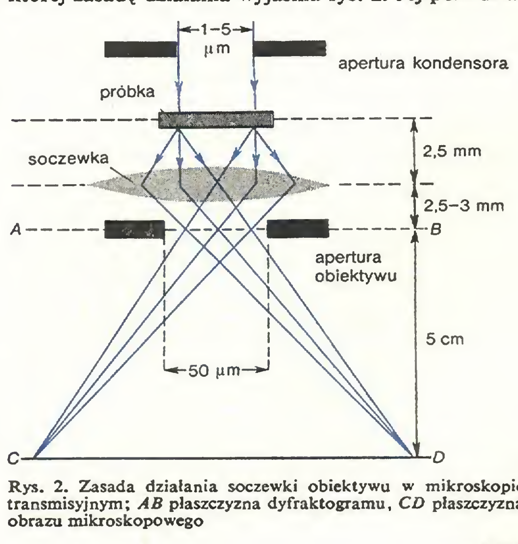
W określonego rodzaju detektorach istnieje ścisła współzależność pomiędzy liczbą różnych elementów obrazu mikroskopowego, czasem jego rejestracji i najmniejszą dopuszczalną wartością prądu sygnału elektronowego. Stosując jako detektor fotopowielacz można np. zarejestrować obraz złożony z 10^6 elementów w czasie 100 s przy prądzie sygnału $2{,}6 \cdot 10^{-12}$ A ($i_{min} = 2{,}6 \cdot 10^{-18} n/t$, gdzie n — liczba elementów, t — czas rejestracji).

Obydwa wymienione systemy są stosowane w kontrukcji różnego typu mikroskopów elektronowych, których nazwy zawierają specyfikację systemu oraz ewentualnie podstawowego rejestrowanego sygnału. W ostatnich latach produkowane są coraz bardziej skomplikowane układy, swego rodzaju kombajny elektronowo-optyczne, z przeznaczeniem do kompleksowych badań mikroskopowo-spektroskopowych, którym trudno znaleźć właściwą nazwę i dlatego są one oznaczane wyłącznie symbolami (np. HP-5, HP-51, BS-350).

Transmisyjny mikroskop elektronowy (TEM)

Jest to historycznie pierwszy i jednocześnie najbardziej rozpowszechniony typ mikroskopu elektronowego (il. 88, tabl. 22). TEM pracujący w systemie konwencjonalnym (CTEM) jest najbardziej zbliżony do mikroskopu optycznego. Zasadniczym elementem CTEM jest soczewka elektromagnetyczna obiektywu, której zasadę działania wyjaśnia rys. 2. Jej pole dzia-

TEM w systemie konwencjonalnym

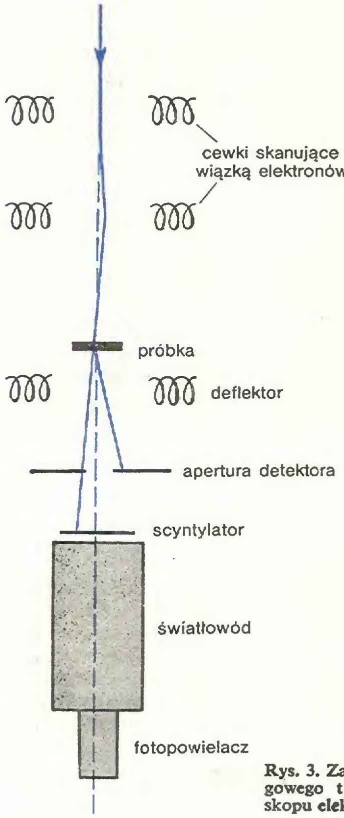


Rys. 2. Zasada działania soczewki obiektywu w mikroskopie transmisyjnym; AB płaszczyzna dyfraktogramu, CD płaszczyzna obrazu mikroskopowego

łania to obszar pomiędzy próbką i diafragmą, zwaną aperturą obiektywu. Wiązka elektronów oświetlająca próbkę jest ograniczona przez aperturę kondensora do średnicy 1–5 μm. Powiększenie soczewki obiektywu wynosi kilkaset. Po przejściu przez tę soczewkę elektrony tworzą dwa obrazy: wzór dyfrakcyjny (elektronogram) w płaszczyźnie ogniskowania *AB* oraz obraz mikroskopowy w płaszczyźnie *CD* (odległość pomiędzy płaszczyznami *AB* i *CD* wynosi 5 cm). Wzór dyfrakcyjny i obraz mikroskopowy sprzężone są transformatą Fouriera (przy założeniu braku aberracji). Po powiększeniu, zazwyczaj dwustopniowym (soczewka zw. pośrednią i soczewka projektora tworzą układ odpowiadający okularowi w mikroskopie optycznym), każdy z tych obrazów może być zarejestrowany na ekranie fluorescencyjnym (pokrytym luminoforem) albo na kliszy fotograficznej.

TEM w systemie skaningowym

TEM pracujący w systemie skaningowym (STEM) nie zawiera soczewek tworzących obraz między próbką i detektorem obrazu — zazwyczaj fotopowielaczem (rys. 3). Obniża to negatywny wpływ strat



Rys. 3. Zasada działania skaningowego transmisyjnego mikroskopu elektronowego

energii elektronów, przez co z równą wyrazistością w STEM można oglądać próbki 2–5 razy grubsze aniżeli w CTEM. Teoretycznie STEM i CTEM powinny dawać identyczne obrazy mikroskopowe, w praktyce jednak zdolność rozdzielcza STEM — ograniczona średnicą sondy (wiązki elektronowej) — jest nie mniejsza niż 5 Å. Zaletą STEM, szczególnie istotną w zastosowaniach w biologii i medycynie, jest elektroniczna regulacja kontrastowości obrazu i możliwość pracy przy bardzo małym prądzie wiązki (10^{-10}–10^{-12} A, w porównaniu z 10^{-5}–10^{-7} A dla CTEM). Wielkość prądu wpływa z kolei na osiągalną w TEM zdolność rozdzielczą (tabela).

obrazy w polu jasnym i ciemnym

Obydwa systemy TEM umożliwiają tworzenie obrazów mikroskopowych w polu jasnym i w polu ciemnym. Otrzymuje się je odpowiednio przez wyodrębnienie wiązki centralnej lub jednego z bocznych maksimów dyfrakcyjnych (refleksów). Obserwacja w polu jasnym odpowiada patrzeniu na przedmiot pod światło, co nie zawsze sprzyja dobrej widoczności. Z kolei obserwacja przedmiotu w świetle rozproszonym, czyli w polu ciemnym, przy jednoczesnej możliwości zmian konfiguracji oko–przedmiot–źródło światła, pozwala na znacznie bardziej precyzyjną analizę jego struktury (w wypadku mikroskopii elektronowej bardziej istotna jest struktura niejednorodności wnętrza aniżeli powierzchni próbki).

Zależność zdolności rozdzielczej od rodzaju próbki (TEM)

Rodzaj próbki	Doza krytyczna m_c		Graniczna zdolność rozdzielcza $\delta_{min} = 50/\sqrt{m_c}$
	A/cm²	elektronów/Å²	
Kryształy nieorganiczne	0,1–10	10^{2}–10^{4}	5–0,5 Å
Substancje organiczne	10^{-3}–1	1–10^{3}	10–1 Å
Żywa materia	10^{-9}–10^{-6}	10^{-6}–10^{-3}	1–0,1 μm

preparatyka próbek

W mikroskopii elektronowej duże znaczenie ma odpowiednie przygotowanie próbki. Przy zazwyczaj stosowanych wiązkach elektronowych o energii 100 keV próbki muszą mieć grubość nie większą niż 1000 Å. Preparatyka próbek jest tematem, który stanowi samoistną dziedzinę wiedzy fizykochemicznej. Próbki dzielą się na dwie zasadnicze grupy: repliki powierzchni i cienkie folie materiału stanowiącego przedmiot badań mikroskopowych.

repliki

Repliki, otrzymywane przez pokrycie próbki cienką warstewką dobrze przylegającej substancji (np. naparowanie węglem), po oddzieleniu od podłoża i odpowiedniej dalszej obróbce (cieniowanie, czyli naparowanie pod kątem ciężkim metalem) umożliwiają otrzymanie obrazu mikroskopowego topografii powierzchni próbki z rozdzielczością do 20 Å. Pewne rodzaje replik, zwane ekstrakcyjnymi, umożliwiają bezpośrednią obserwację drobnych wydzieleń (mikroobszarów o innych własnościach fizykochemicznych) pochodzących z próbki.

W ostatnim dziesięcioleciu znaczenie replik zmalało, w wyniku upowszechnienia skaningowych metod przetwarzających na obraz mikroskopowy sygnały elektronów odbitych i wtórnych emitowanych z powierzchni próbek litych.

cienkie folie

Wzrosła natomiast liczebność i wyrafinowanie technik służących otrzymywaniu cienkich folii. Próbki do badań fizycznych czy inżynierii materiałowej otrzymuje się przez polerowanie chemiczne, elektrochemiczne, trawienie jonowe lub walcowanie. Preparaty biologiczne tnie się na cienkie próbki za pomocą ultramikrotomu. Problemem jest w tym wypadku zabezpieczenie delikatnej tkanki preparatu przed uszkodzeniem w wyniku cięcia lub odwodnienia. Podstawowymi składnikami próbek organicznych są pierwiastki lekkie, dające znikomy kontrast. Kontrast od różnych strukturalnie części próbki można polepszyć nasycając próbkę substancjami zawierającymi pierwiastki ciężkie (np. Os). Nasycenie ciężkimi pierwiastkami stanowi analogię do barwienia preparatów organicznych przy obserwacji w mikroskopie optycznym.

badanie obiektów biologicznych

W dziedzinie badań biologicznych i medycznych wykorzystuje się TEM do poznania biologii komórek, w szczególności budowy jądra komórkowego, cytoplazmy i błony komórkowej (→ Błony komórkowe, Molekularne podstawy skurczu mięśnia). Ustalenie ilości i rodzajów występujących w nich organelli w powiązaniu z możliwością określenia ich funkcji chemicznych dało podstawę pogłębienia wiedzy o funkcjonowaniu tkanek. Wiele obiektów biologicznych ma dostatecznie małe rozmiary, aby je bezpośrednio obserwować w TEM, np. bakterie i wirusy. Ich cechą specyficzną jest zachowanie określonego typu symetrii przestrzennej. Doskonalenie jakości zdjęć właśnie tego typu obiektów oraz zastosowanie komputera i holografii umożliwiło rozwój mikroskopowej metody badań zw. trójwymiarową rekonstrukcją. Trójwymiarowa rekonstrukcja cząsteczek dostarcza

dziś pełnej, przestrzennej informacji o budowie wewnętrznej i kształcie, na przykład wirusa z rozdzielczością 10 Å.

badania materiałowe

W dziedzinie inżynierii materiałowej za pomocą TEM uzyskano bezpośredni dowód istnienia defektów sieci krystalicznej, dyslokacji i błędów ułożenia (il. 87, tabl. 22). Zrozumienie znaczenia dyslokacji w procesie deformacji plastycznej materiałów, szczególnie blokowania ich ruchu przez drobne wydzielenia stało się podstawą oceny własności sprężystych wielu metali i stopów. Możliwość łatwego aparaturowego przejścia od wzoru dyfrakcyjnego do obrazu mikroskopowego została wykorzystana do identyfikacji struktury wydzieleń w materiałach wieloskładnikowych. Obecnie jest możliwe przy zastosowaniu mikro-mikrodyfrakcji określenie struktury wydzielenia o średnicy 200 Å. W tym kontekście duże znaczenie ma zastosowanie TEM do badania przemian fazowych. W fizyce półprzewodników poznanie geometrii dyslokacji pozwoliło w kilku wypadkach (Si, ZnTe) na pełniejsze wyjaśnienie własności elektronowych i optycznych tych materiałów (→ Dyslokacje w kryształach).

Za pomocą TEM o dużej rozdzielczości i przy zastosowaniu komputera można — z obrazu mikroskopowego płaszczyzn sieciowych — wnioskować o strukturze sieci krystalicznej materiału (defektach itp.). Innym zadaniem mikroskopii wysokorozdzielczej jest obserwacja pojedynczych atomów i cząsteczek. Operując analogią optyczną zauważenie obiektu o rozmiarach atomu (3–4 Å) z odległości 30 cm odpowiada sytuacji, w której chcemy dostrzec kulkę szklaną o średnicy 0,03 mm z odległości 30 km. Obraz atomów ciężkich pierwiastków (U, Th) uzyskano (lata 1969–70) zarówno w CTEM (w ciemnym polu), jak również w STEM.

badania „in situ"

Ważnym kierunkiem zastosowań TEM są badania „in situ", w których próbka jest poddawana programowanym zmianom (przez chłodzenie, grzanie, podświetlanie, rozciąganie) wewnątrz kolumny mikroskopu. Do grupy mikroskopów specjalnie przystosowanych do takich badań należy niewątpliwie zaliczyć wysokonapięciowy TEM (HVEM), pracujący przy napięciu przyspieszającym elektrony od 1 do 3 MV. Elektrony o takich wysokich energiach, poza zwiększoną przenikliwością (próbki mogą mieć grubość ok. 10^4 Å), powodują powstawanie defektów radiacyjnych. Relatywistyczne elektrony o energii większej niż krytyczna wartość W_{min} po przeniknięciu powłoki K atomu w sieci krystalicznej powodują jego wybicie w położenie międzywęzłowe. W ten sposób powstaje para defektów typu Frenkla — luka oraz atom międzywęzłowy (tabela). Defekty radiacyjne bada się dla celów poznawczych i aplikacyjnych w półprzewodnikach i materiałach reaktorowych. Głównym atutem HVEM jest przyspieszenie tempa procesów takich, jakie zachodzą w akceleratorach liniowych i reaktorach, procesów, które prowadzą do tworzenia agregatów defektów punktowych wpływających na zmianę własności materiałów. Przy ustalonej gęstości prądu elektronowego skupienie ich w plamkę o średnicy ok. 10^{-12} m², zamiast na powierzchni 10^{-4} m², powoduje przyspieszenie generacji defektów w proporcji czasu jak 1 min do 200 lat.

HVEM

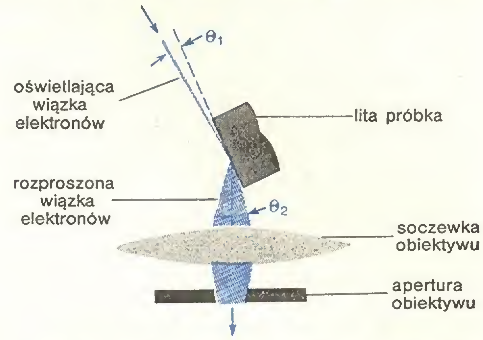
Defekty radiacyjne generowane wysokoenergetycznymi elektronami (HVEM)

Wybity atom	Napięcie krytyczne, kV	Wybity atom	Napięcie krytyczne, kV
H	2–3	O w MgO (kryształ)	330
C w substancjach organicznych	25–30	Mg w MgO (kryształ)	480
Si w krzemianach	40	Metale (kryształy)	100–1500

Odbiciowy mikroskop elektronowy (REM)

REM w systemie konwencjonalnym

Mikroskopy należące do tej i do dwóch dalszych grup stosuje się wyłącznie do obserwacji powierzchni próbek litych. W systemie konwencjonalnym wykorzystywane są elektrony o wysokiej energii, rozproszone na próbce pod małymi kątami (rys. 4). Zasadniczym czynnikiem ograniczającym zdolność rozdzielczą jest

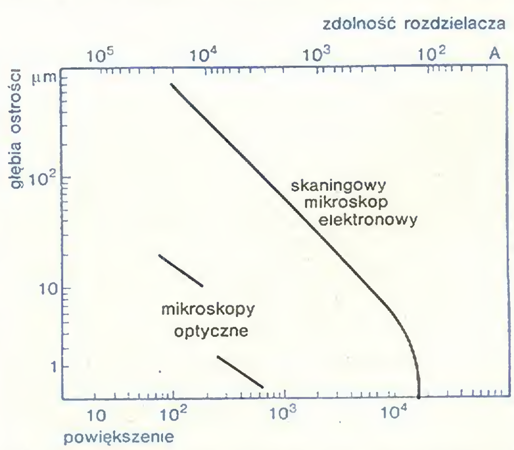


Rys. 4. Zasada działania mikroskopu odbiciowego

aberracja chromatyczna, którą zmniejsza się przez stosowanie małych kątów θ_1 i θ_2. Czynnikiem pogarszającym zdolność rozdzielczą przy małych aperturach jest rozmycie dyfrakcyjne. Zdolność rozdzielcza nie przekracza 300 Å ($\theta_1 = 2°$, $\theta_2 = 23°$). Problemem jest również mała jasność obrazu. Przetworniki elektronooptyczne dają rozwiązanie tylko pozorne, ponieważ ze swej strony pogarszają zdolność rozdzielczą. Konwencjonalny REM wykorzystuje się głównie do badań związanych z dyfrakcją wysokoenergetycznych elektronów (HEED), stanowiących cenne źródło informacji o strukturze krystalicznej cienkiej warstwy powierzchniowej. Inne jego zastosowanie — to stereometria, korzystająca z kontrastowych i długich cieni od nierówności topograficznych.

REM w systemie skaningowym

REM pracujący w systemie skaningowym (SREM, często skrótowo oznaczany symbolem SEM) tworzy obraz powierzchni próbki wykorzystując sygnały elektronów odbitych, wtórnych, absorbowanych i katodoluminescencji, a w szczególnych wypadkach również elektronów Augera (konieczna próżnia co najmniej 10^{-6} Pa). Wiązka elektronów oświetlająca próbkę pada na nią na ogół pod dowolnym (zazwyczaj 90°), ale stałym kątem. Tylko w badaniach orientacji krystalograficznej powierzchni stosuje się zmienny kąt wejścia elektronów do próbki. Ze względu na silną kolimację pierwotnej (oświetlającej) wiązki elektronów SREM ma dużą głębię ostrości (rys. 5). Dlatego nadaje się specjalnie do badania kruchych próbek o złożonym kształcie, często spotykanych



Rys. 5. Zależność głębi ostrości od powiększenia dla mikroskopów optycznych i mikroskopu elektronowego typu SREM

rozdzielanie obrazów topograficznych i kompozycyjnych

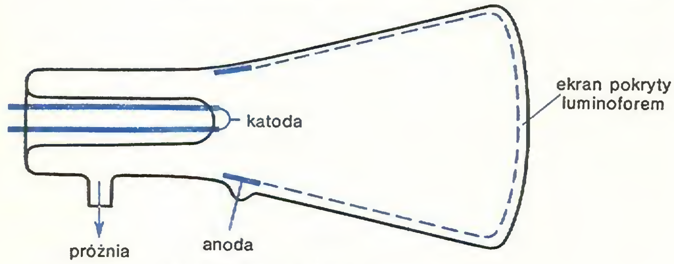
w biologii i w mikroelektronice. Zdjęcia mikroskopowe wykonane w świetle elektronów odbitych (il. 85a, tabl. 22) ujawniają nie tylko nierówności powierzchni (kontrast topograficzny), ale także jej skład chemiczny (kontrast kompozycyjny). Istnieje możliwość rozdzielenia obrazów topograficznych i kompozycyjnych. W badaniu półprzewodników jest istotna możliwość przetwarzania na obraz mikroskopowy prądu wywołanego absorbowaną wiązką elektronów, w w szczególności siły elektromotorycznej (il. 85b). Jeśli próbka ma złącze *p-n*, prąd ten indukuje dodatkowe przewodnictwo elektryczne, którego rozkład przestrzenny dostarcza informacji o własnościach elektronowych półprzewodnika, takich jak droga dyfuzji i czas życia nośników mniejszościowych. Na koniec katodoluminescencja (il. 85c) obserwowana w materiałach organicznych i niektórych półprzewodnikach (np. CdS, ZnS, GaAs, CdTe) jest czuła na zanieczyszczenia i defekty sieci.

Emisyjny mikroskop elektronowy (EEM)

W tej grupie mikroskopów lita próbka spełnia funkcję katody, tzn. emitera elektronów tworzących obraz. Wymuszenie emisji elektronów może nastąpić w wyniku podwyższenia temperatury próbki (termoemisja), bombardowania elektronami lub jonami (emisja wtórna i jonowo-elektronowa), oświetlenia (fotoemisja) lub działania silnego pola elektrycznego (autoemisja). Układ elektronowo-optyczny formowania obrazu w EEM jest podobny jak w CTEM. Zdolność rozdzielcza δ zależy od rozmycia energetycznego $\overline{\Delta W}$ elektronów i od pola elektrostatycznego E przy powierzchni próbki ($\delta = 0{,}6\overline{\Delta W}/E$), a także od rodzaju emisji (200 Å przy termoemisji i $T = 2800$ K, $E = 3\cdot10^6$ V/m; 500–1000 Å przy emisji jonowo-elektronowej; 800 Å przy fotoemisji; 1000 Å przy emisji wtórnej). W warunkach termo- i fotoemisji zasadniczym czynnikiem tworzenia kontrastu jest lokalna wartość pracy wyjścia elektronów z próbki.

zastosowanie EEM

EEM znalazł zastosowanie w fizyce metali przy badaniu zdolności emisyjnej katod, w analizie przemian fazowych w badaniach „in situ” wzajemnego oddziaływania jonów z materią (obserwacja równoczesnego trawienia jonowego i utleniania próbki) itp.



Rys. 6. Projektor elektronowy

projektor elektronowy

Przykładem prostego EEM jest projektor elektronowy pokazany na rys. 6. Zazwyczaj jest to kolba szklana (wewnątrz pokryta w przedniej części luminoforem), w którą są wtopione: metalowe ostrze o promieniu krzywizny 10^3–10^4 Å (katoda) oraz otaczający je metalowy pierścień (anoda). Ciśnienie w kolbie wynosi 10^{-8}–10^{-9} Pa. Przyłożenie do anody napięcia dodatniego rzędu kilku kV powoduje powstanie przy powierzchni ostrza pola elektrostatycznego 10^9–10^{10} V/m, które wystarcza do wywołania autoemisji. Powiększenie $M = R/r\beta$, gdzie R — odległość między katodą i ekranem, r — promień krzywizny ostrza (nagrzewanie ostrza powoduje, że jego czubek ulega zaokrągleniu), β — współczynnik zniekształcenia obrazu. Na ogół $M = 200$ tys. Zdolność rozdzielczą projektora elektronowego (ok. 25 Å) ogranicza występowanie składowej prędkości elektronów stycznej do powierzchni ostrza oraz, w mniejszym stopniu, dyfrakcja elektronów. Mikroskopy tego typu wykorzystuje się do badania autoemisji z metali trudno topliwych i półprzewodników, badania zjawiska adsorpcji i chemisorpcji oraz wyznaczania pracy wyjścia z różnych płaszczyzn sieciowych w monokryształach.

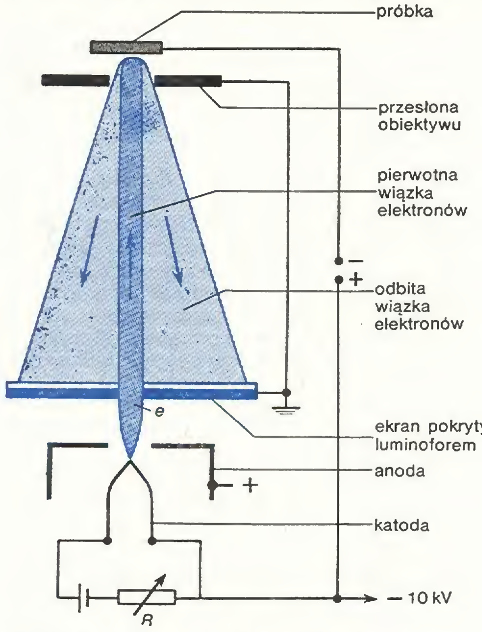
zastosowanie projektora

Zwierciadlany mikroskop elektronowy (MEM)

W mikroskopie zwierciadlanym lub inaczej w zwierciadle elektronowym elektrony oświetlające powierzchnię litej próbki nie stykają się z nią bezpośrednio. Zasadniczym jego elementem jest obiektyw imersyjny, utworzony z próbki i cylindrycznej diafragmy (rys. 7). Pomiędzy próbką (podaje się na nią ujemne napięcie $V_0 = 10$–30 kV) i diafragmą (uziemiona) istnieje silne pole elektrostatyczne E_0, które przekazuje elektronom tworzącym obraz mikroskopowy informacje o próbce. Szczególne znaczenie ma kształt tzw. zerowej powierzchni ekwipotencjalnej, to jest powierzchni, na której następuje zahamowanie i zawrócenie elektronów w kierunku ekranu fluorescencyjnego. Wyboru zerowej powierzchni dokonuje się przez przyłożenie do katody napięcia o kilka do kilkunastu woltów wyższego niż do próbki. Próbka przekazuje zerowej powierzchni ekwipotencjalnej szczegóły topografii swojej powierzchni oraz dodatkowo kształt występujących na niej pól elektrostatycznych i magnetycznych. Poprzeczna (względem osi optycznej) geometryczna zdolność rozdzielcza nie przekracza 1000 Å (przy $E_0 = 5\cdot10^6$ V/m). Najważniejszym parametrem decydującym o atrakcyjności MEM jest jednak nie geometryczna, ale napięciowa zdolność rozdzielcza, określająca wykrywalną składową zmiany pola elektrycznego w kierunku równoległym do powierzchni próbki. Stwierdzono, że nie przekracza ona wartości $6\cdot10^7$ V/m². Dzięki wysokiej czułości napięciowej MEM znalazł zastosowanie w badaniu elektrycznie aktywnych elementów powierzchni półprzewodników (np. złącz *p-n*, il. 86, tabl. 22), ferroelektryków, a także materiałów magnetycznych. Na specjalne podkreślenie zasługuje przydatność MEM do badań fizycznych, czego przykładem jest obserwacja zmian przewodnictwa elektrycznego powierzchni kryształu krzemu w wyniku wymuszonej migracji zawartych w jego sieci krystalicznej domieszek atomów obcych.

zerowa powierzchnia ekwipotencjalna

zastosowanie MEM



Rys. 7. Zasada działania zwierciadlanego mikroskopu elektronowego

Analityczny mikroskop elektronowy (EMA)

przestrzenny rozkład pierwiastka

Mikroskopy tego typu służą do przedstawienia w obrazie mikroskopowym przestrzennego rozkładu wybranego pierwiastka chemicznego, stanowiącego składnik materiału próbki. Zazwyczaj osiąga się taki obraz (il. 84, tabl. 22) przez sprzężenie TEM lub REM z odpowiednim spektrometrem, który analizuje widmo charakterystycznego promieniowania rentgenowskiego, elektronów transmitowanych przez próbkę, ewentualnie elektronów Augera. Procesy emisji kwantu charakterystycznego promieniowania rentgenowskiego i elektronu Augera o określonej energii są wzajemnie konkurencyjne. Pierwszy dominuje w atomach pierwiastków ciężkich, drugi — lekkich. Dobór spektrometru (w wypadku elektronów Augera jest to również specjalnego typu SREM) powinien zależeć od oczekiwanego składu chemicznego próbki.

Z. Bojarski *Mikroanalizator rentgenowski*, Katowice 1971; I. Goldstein H. Yakowitz (red.) *Practica Scanning Electron Microscopy*, New York 1975; P. B. Hirsch i in. *Electron Microscopy of Thin Crystals*, London 1971; D. B. Holt i in. (ed.) *Quantitative Scanning Electron Microscopy*, London 1974; J. Kozubowski *Metody transmisyjnej mikroskopii elektronowej*, Katowice 1975.

Osiągnięcia krystalografii białek

Tadeusz Bartczak

Największe osiągnięcia analizy strukturalnej kryształów i krystalochemii w ostatnich dziesięciu latach wiążą się z tzw. krystalografią białek. W obecnej chwili rentgenowska analiza strukturalna białek stanowi jedyną metodę, za pomocą której można uzyskać precyzyjne dane strukturalne, tzn. przestrzenne współrzędne atomów.

Biopolimery takie jak białka są bardzo delikatnym i trudnym obiektem badań. Po pierwsze — już próby wyhodowania kryształów dostatecznie dużych (wymiary liniowe $\geq$ 0,3 mm) wymagają zastosowania nowoczesnych technik eksperymentalnych rozległej wiedzy biochemicznej, wykonania setek doświadczeń i nierzadko kończą się niepowodzeniem. Po drugie — kryształy biopolimerów są zwykle o wiele bardziej nieuporządkowane niż kryształy zbudowane z małych cząsteczek organicznych. Znaczy to, że cząsteczki, grupy lub łańcuchy atomów mogą być różnie zorientowane w tej samej części komórki elementarnej. Takie nieuporządkowanie nosi nazwę nieuporządkowania statycznego i powoduje duże trudności eksperymentalne. W kryształach biopolimerów obserwuje się ponadto nieuporządkowanie całych cząsteczek spowodowane drganiami termicznymi atomów. Jak wiadomo, ok. 50% objętości kryształów białek zwykle stanowi woda. Te cząsteczki wody mogą być mocno związane z powierzchnią cząsteczki białka albo luźno utrzymywane w przestrzeniach pomiędzy sąsiednimi cząsteczkami białka. Otoczenie przez „poduszkę" wody powoduje, że cząsteczki biopolimeru są luźniej ulokowane w sieci krystalicznej i dlatego zdolne są do wykonywania dużych drgań termicznych.

nieuporządkowanie statyczne i termiczne

Wreszcie wielkie rozmiary kryształów białek (typowy okres identyczności komórki elementarnej 100 Å w porównaniu z przeciętnym okresem identyczności 10 Å dla kryształu małej cząsteczki organicznej) powoduje, że trzeba zmierzyć i przetworzyć ogromną liczbę interferencji.

Te specyficzne cechy kryształów białek spowodowały, że metody ich badania za pomocą promieni rentgenowskich wyodrębniły się w oddzielną dziedzinę badań: krystalografię białek. Opracowanie wielu specyficznych technik badawczych i metod obliczeniowych umożliwiło pomyślne zbadanie wielu struktur białkowych i dlatego stanowi największe osiągnięcie analizy strukturalnej i jedno z największych osiągnięć nauki ostatnich lat. Dyfrakcja promieni rentgenowskich umożliwiła realizację jednego z podstawowych celów biologii molekularnej: interpretację funkcji biologicznej poprzez poznanie własności fizycznych i chemicznych struktur biologicznych. Wstępnym warunkiem fizykochemicznej interpretacji procesów życiowych jest poznanie struktury odpowiednich komponentów życia na poziomie atomowym. Krystalografia rentgenowska jest jedyną znaną metodą dostarczającą tych informacji w sposób kompleksowy.

Rozwój krystalografii białek

Krystalografia białek rozwinęła się raptownie w ostatnich kilkunastu latach, dzięki rozwojowi elektronicznej techniki obliczeniowej i nowoczesnych metod pomiaru intensywności promieniowania rentgenowskiego ugiętego na krysztale (stosuje się tzw. automatyczne, czterokołowe dyfraktometry monokrystaliczne). Pierwsze eksperymenty wykonane były przez J. Bernala i D. Crowfoot w 1934 r. w Cambridge na krysztale pepsyny, metodą rejestracji na błonie fotograficznej. Uświadomiły one badaczom fundamentalny fakt, że wielkie cząsteczki biologiczne są uporządkowane w sieć krystaliczną i że ich struktury można wobec tego określić metodami dyfrakcji rentgenowskiej. Z tego okresu pochodzi anegdota o tym, jak D. Crowfoot — późniejsza laureatka nagrody Nobla, pedałowała na rowerze do laboratorium, w środku nocy, przez uśpiony Oxford, aby jeszcze raz upewnić się, że kryształ insuliny naprawdę ugiął promieniowanie, co poprzedniego dnia zarejestrowała na błonie fotograficznej. Historia ta dobrze ilustruje wielkie podniecenie, jakie opanowywało badaczy w tamtych czasach, po otrzymaniu każdego udanego rentgenogramu białka. Wspomniane kryształy pepsyny otrzymano w Uppsali w sposób zupełnie przypadkowy z roztworu, który stał w spokoju przez kilka tygodni. Ilustruje to sposób, w jaki wybierano obiekty do badań we wczesnych latach rozwoju krystalografii białek. Badano nie te białka, które były najważniejsze z punktu widzenia ich znaczenia biologicznego, ale te, które były łatwo dostępne i które można było łatwo wykrystalizowac. Upłynąć musiało następnych 35 lat, zanim metody krystalografii białek rozwinęły się na tyle, żeby można było ogłosić drukiem opis struktury insuliny.

izomorficzne pochodne białek z atomem ciężkim

W krystalografii białek tzw. „problem fazowy" (→ Strukturalna analiza kryształów) jest szczególnie trudny ze względu na złożoną budowę tych cząsteczek. Trudności te pokonał M. von Perutz wraz ze współpracownikami w laboratorium L. Bragga w Cambridge. Wykazali oni, że tzw. „ciężki atom", np. atom rtęci, można wprowadzić do kryształu białka bez zakłócenia sposobu ułożenia atomów w krysztale, otrzymując kryształ pochodnej izomorficznej, który wykazuje znaczne i możliwe do zarejestrowania zmiany intensywnosci promieni ugiętych. Metoda ta doprowadziła w 1960 r. do ustalenia po raz pierwszy struktury białka — mioglobiny przez J.C. Kendrew i współpracowników.

Ograniczone możliwości krystalografii białek

W krystalografii białek analiza strukturalna uwieńczona sukcesem stanowi dopiero punkt wyjściowy do wielu doświadczeń mających na celu wyświetlenie

ich funkcji biologicznych. Należy tu wspomnieć o trudnościach występujących w badaniu funkcji, jakie spełniają białka. Po pierwsze białko musi być wykrystalizowane, ażeby w ogóle można je było badać za pomocą dyfrakcji promieni rentgenowskich. Powstaje problem, czy krystalizacja białka zmienia jego strukturę? Na szczęście odpowiedź może być przecząca, jeśli przyjmiemy, że struktura w krysztale reprezentuje trwałą konformację równowagową cząsteczki. Siły, które utrzymują cząsteczki białka w sieci krystalicznej, są o wiele słabsze niż te siły, które rządzą strukturą samego białka wskutek mniejszej liczby kontaktów międzyatomowych między cząsteczkami białka niż w obrębie samej cząsteczki. Zatem duże zmiany konformacyjne nie są prawdopodobne. Niekiedy te same białka wyizolowane z różnych organizmów zwierzęcych czy roślinnych lub w ogóle to samo białko wykrystalizowane w bardzo różniących się warunkach miało struktury bardzo podobne. Małych różnic konformacyjnych pomiędzy białkiem krystalicznym a tym samym białkiem w roztworze nie można wykluczyć, chociaż znany jest również fakt, że pewne białka zachowują swoją aktywność biologiczną w krysztale. W sytuacji, kiedy brakuje informacji strukturalnych na temat białek w roztworach, musimy akceptować strukturę krystaliczną jako punkt wyjściowy do interpretacji zachowania się tej cząsteczki w roztworze.

wpływ krystalizacji na konformację

Metoda rentgenograficzna jest w swej istocie statyczna, a procesy biologiczne są w swej istocie dynamiczne. Eksperyment rentgenograficzny, który wymaga określonego czasu na jego przeprowadzenie, zwykle kilku dni, dostarcza obrazu cząsteczki, który jest uśredniony w tym czasie. Nie można za pomocą jedynie metody rentgenograficznej oznaczyć struktur przejściowych przyjmowanych przez cząsteczkę białka w trakcie spełniania jej funkcji biologicznej. Wyciąganie prawidłowych wniosków wymaga ścisłej współpracy z innymi dyscyplinami nauki. Metody magnetycznego rezonansu jądrowego stale ulepszane stwarzają największe szanse badania takich struktur przejściowych.

Etapy badania strukturalnego białek krystalicznych

Badanie struktury białka składa się zwykle z ośmiu etapów: krystalizacja, przygotowanie izomorficznych pochodnych z atomem ciężkim, zmierzenie i zarejestrowanie intensywności promieni rentgenowskich ugiętych na krysztale białka, przetwarzanie tych danych w postać nadającą się do obliczeń, określenie położeń atomów ciężkich, obliczanie faz, interpretacja map gęstości elektronowej i uściślenie danych dotyczących otrzymanych współrzędnych atomowych.

krystalizacja

Krystalograf musi nie tylko wykrystalizować białko, ale także wyhodować kryształ pokaźnych rozmiarów. Jest to konieczne, ponieważ intensywności promieni ugiętych na krysztale są mniej więcej proporcjonalne do objętości kryształu, a odwrotnie proporcjonalne do objętości krystalograficznej komórki elementarnej, która nie może być mniejsza od jednej podjednostki białka. Dla białek o masie cząsteczkowej poniżej 50 000 daltonów, kryształ o wymiarach liniowych 0,1 mm dałby prawdopodobnie obraz dyfrakcyjny pozwalający na wydedukowanie podstawowych danych komórki krystalograficznej. Aby pomyślnie wykonać analizę z wysoką rozdzielczością, konieczny jest kryształ o wymiarach liniowych co najmniej 0,3 mm. Dla białek o większej masie cząsteczkowej wymagane wymiary są większe. Aby zastosować metodę neutronograficzną potrzebne są kryształy o wymiarach liniowych 3 mm nawet dla najmniejszych białek. Jedynie metodami mikroskopii elektronowej można badać bardzo małe kryształy.

Krystalizacja białek jest zadaniem bardzo trudnym, gdyż wiele białek jest dostępnych w bardzo małych ilościach. W ostatnich latach rozwinięto szereg nowych metod krystalizacji, co pozwala na wybór białek do badań na bardziej racjonalnych podstawach.

Krystalizuje się białko oczyszczone, wyizolowane z odpowiedniego materiału biologicznego. Trzeba się starać, aby w miarę możliwości białko pozostawało w postaci biochemicznie aktywnej. Należy zbadać zależność rozpuszczalności od pH, mocy jonowej, rozpuszczalników organicznych i temperatury, z uwzględnieniem zależności już określonych przez badania biochemiczne. Aby zobrazować skalę trudności, z jakimi spotykają się badacze przytoczymy pracę F. Lynena i współpracowników nad krystalizacją syntetazy kwasu tłuszczowego — wieloenzymowego kompleksu składającego się z siedmiu białek o całkowitej masie cząsteczkowej $2,3 \times 10^6$ daltonów. W preparacie, który przechowywano przez 15 miesięcy w temperaturze 4°C, zaobserwowano obecność mikrokryształów. Następnie zmieniano warunki krystalizacji w sposób systematyczny w innych preparatach, do których wprowadzano te mikrokryształy jako zarodki krystalizacji. Ustalono w ten sposób optymalne warunki, w których udało się wyhodować kryształy w postaci słupów heksagonalnych o wymiarach liniowych 0,1 mm w ciągu 2 dni.

Nawet jeżeli udaje się wyhodować kryształy, to często warto kontynuować badania nad innymi warunkami krystalizacji, w których uda się może otrzymać odmianę polimorficzną bardziej odpowiednią do badań rentgenograficznych. Tak np. praca nad otrzymaniem kryształów chymotrypsynogenu trwała przez całe lata, na skutek otrzymywania nieodpowiedniej odmiany kryształów. Otrzymano dziesięć różnych odmian krystalicznych enzymu allosterycznego, asparginianu transkarbamylazy (masa cząsteczkowa 300 000 daltonów). Odmianę najbardziej odpowiednią do badań znaleziono przez przypadek. Również insulina wykazuje polimorfizm postaci krystalicznych. Odmiana rombowa otrzymywana przy $pH < 7$ jest nietrwała w temperaturze pokojowej, trwała jest natomiast odmiana romboedryczna wykrystalizowana przy $pH = 6,3$.

Wspomnieć trzeba nadto, że kryształy białek różnią się jeszcze i tym od kryształów mniejszych cząsteczek organicznych, że zawierają znaczną ilość ciekłego rozpuszczalnika (rys. 1).

metoda izomorficznego podstawienia

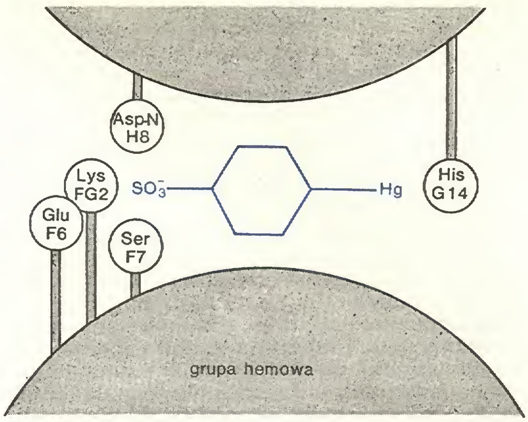
Metoda izomorficznego podstawienia została po raz pierwszy zastosowana przez J.M. Robertsona i I. Woodwarda w 1937 r. w ich analizach ftalocyjanin.

Kryształy białka zawierają duże kanały wypełnione roztworem macierzystym. Często jest możliwe przyłączenie ciężkiego atomu do powierzchni białka w ten sposób, że nie zakłóca on struktury molekularnej ani krystalicznej, ale zajmuje położenie w jednym z tych kanałów. Wzdłuż nich atomy ciężkie dyfundują w głąb natywnych kryształów białka. Jeśli atom ciężki wykazuje wystarczającą „gęstość elektronową", to zaobserwuje się zmiany w obrazie rentgenograficznym. Na przykład wprowadzenie atomu rtęci, który ma 80 elektronów, do białka o masie cząsteczkowej 24 000 daltonów powoduje średnią zmianę intensywności o 40%. Możliwe do interpretacji mapy gęstości elektronowej kompleksu nukleaza–inhibitor uzyskano używając tylko jednej pochodnej izomorficznej, otrzymanej przez wprowadzenie stosunkowo lekkiego atomu jodu — zawierającego 46 elektronów. Większe białka wymagają cięższych atomów i stosuje się jony kompleksowe metali, takie jak $[Ta_6Cl_{12}]^{2+}$, $[W_6Cl_8]^{4+}$ i inne. Każdy z czynników struktury (→ Strukturalna analiza kryształów) pochodnej zawierającej atom ciężki, oznaczanych jako F_H, jest wektorową sumą udziałów białka F i atomu ciężkiego f. Zatem $\vec{F_H} = \vec{F} + \vec{f}$. Pierwszym krokiem w analizie jest określenie położenia atomu ciężkiego. Można to osiągnąć za pomocą syntezy Pattersona,

w której stosuje się jako współczynniki różnice $(|F_H|-|F|)^2$. Cząsteczki białek krystalizują w niecentrosymetrycznych grupach przestrzennych (→ Kryształy), jednak pewne dwuwymiarowe rzuty funkcji Pattersona są zwykle centrosymetryczne. Ułatwia to bardzo określenie współrzędnych atomu ciężkiego na takich rzutach, ponieważ fazy są równe 0 lub 180° a wobec tego czynniki struktury przyjmują wartości $+|F|$ lub $-|F|$.

metoda anomalnego rozpraszania

Zwykle nie wystarcza analizowanie jednej pochodnej z atomem ciężkim i wymagane jest otrzymanie dwu lub więcej takich pochodnych. Alternatywną metodą określania faz, która wymaga zastosowania tylko jednej pochodnej z atomem ciężkim, jest wykorzystanie zjawiska anomalnego rozpraszania. W normalnych warunkach obowiązuje prawo Friedela: $F(hkl) = F(\bar{h}\bar{k}\bar{l})$, nawet jeżeli struktura jest niecentrosymetryczna. Oznacza to, że niemożliwe jest rozróżnienie pomiędzy dwiema strukturami enancjomerycznymi (tzn. mającymi się do siebie tak jak przedmiot i jego odbicie w lustrze, albo jak dłoń prawa w stosunku do lewej). Jeśli jednak długość fali promieniowania pierwotnego jest bardzo zbliżona do tzw. progu absorpcji któregoś atomu w strukturze, wtedy fala ugięta wykazuje anomalne przesunięcie fazy, które nieco wyprzedza fazę normalną. Powoduje to różnicę intensywności pomiędzy refleksami



Rys. 2. Schemat sposobu wiązania pochodnej rtęci z mioglobiną. Dolna część okręgu wyobraża grupę hemową. Symbole w kółkach oznaczają aminokwasy i ich położenie w łańcuchu polipeptydowym (wg H. C. Watsona i in.)

(hkl) i $(\bar{h}\bar{k}\bar{l})$. D.M. Blow i M.G. Rossmann wykorzystali tę metodę do analizy hemoglobiny i jej pochodnej zawierającej 4 cząsteczki $HgCl_2$. Na rys. 2 jest pokazany schemat wiązania pochodnej rtęci z mioglobiną.

pomiar intensywności promieni ugiętych na krysztale białka

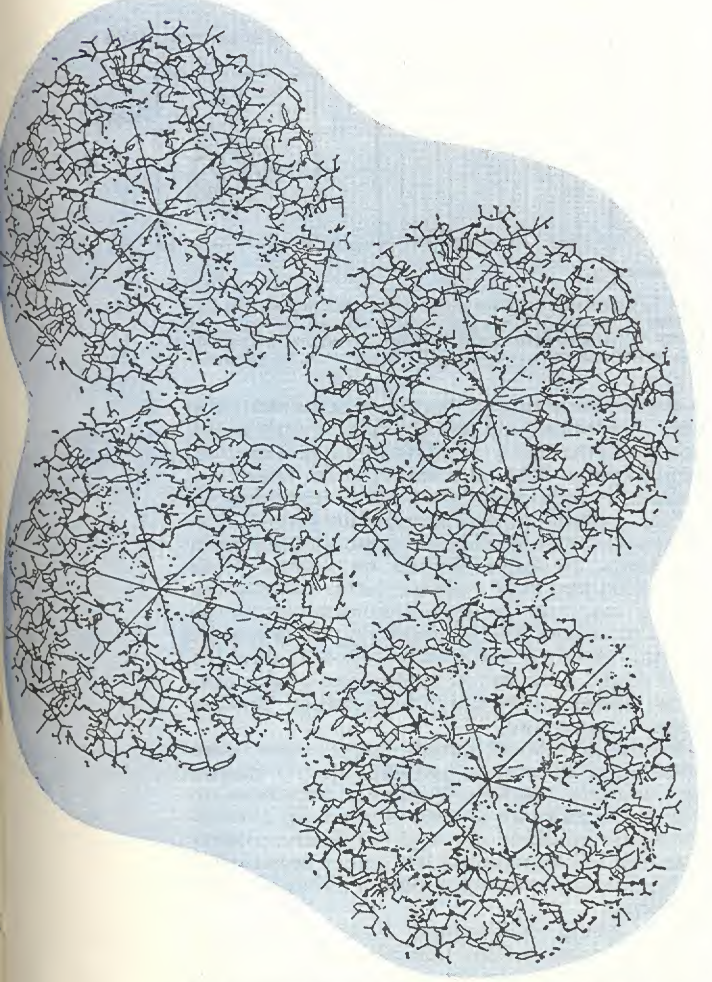
Metoda fotograficzna ma tę wielką zaletę, że rejestruje się przy jej użyciu wielką liczbę promieni ugiętych w jednym eksperymencie, ale potem trzeba zmierzyć intensywność każdej plamki na wywołanej błonie fotograficznej, co jest zadaniem żmudnym i ogromnie pracochłonnym. Ta wada metody fotograficznej była jednym z powodów, dla których zbudowano aparaty do bezpośredniego pomiaru intensywności promieni ugiętych metodą rejestracji licznikowej. Są to automatyczne czterokołowe dyfraktometry monokrystaliczne sterowane komputerami. Komputer służy do obliczania położeń kątowych kryształu i licznika scyntylacyjnego i przestawiania ich w położenia umożliwiające zmierzenie następnego refleksu. Metoda ta wymaga oddzielnego zbadania każdego maksimum dyfrakcyjnego i dlatego zebranie danych o intensywności około 10 000 promieni ugiętych zajmuje zwykle kilka dni. W ciągu tego czasu kryształ białka podlega naświetlaniu, co powoduje jego nieodwracalne uszkodzenie radiacyjne. Stanowi to jeden z głównych problemów dyfraktometrii rentgenowskiej i zmusza krystalografa do opracowania takiej strategii zbierania danych o intensywności promieni ugiętych, która pozwoli zbierać je z żądaną dokładnością, a jednocześnie zminimalizuje czas ekspozycji kryształu w wiązce promieni rentgenowskich. W ostatnich latach starano się zaradzić trudnościom badawczym przez opracowanie nowych aparatów i urządzeń pomiarowych. Skonstruowano automatyczne mierniki stopnia zaczernienia (gęstości optycznej) plamek na rentgenogramach, tzw. densytometry i dyfraktometry przystosowane do jednoczesnej rejestracji intensywności trzech do pięciu refleksów, co oczywiście skraca całkowity czas pomiaru trzy do pięciu razy. Opracowano także nowe rodzaje kamer fotograficznch.

interpretacja map gęstości elektronowej

Omówimy jeszcze krótko etap interpretacji map gęstości elektronowej (map Fouriera), pomijając pozostałe etapy jako zbyt specjalistyczne. Należy tu wyjaśnić pojęcie rozdzielczości. Najmniejsza odległość r, która może być rozróżniona na obrazie uzyskanym za pomocą promieni rentgenowskich, jest określona przez wyrażenie:

$$r \approx 0{,}71dm = 0{,}71\lambda/2\sin\theta(\max),$$

gdzie $\sin\theta(\max)$ oznacza maksymalną wartość $\sin\theta$ (θ — kąt odbłysku Bragga) dla refleksów uży-



Rys. 1. Ułożenie heksamerów insuliny w krysztale romboedrycznej insuliny 2-cynkowej. Kryształ składa się w istocie z dwu faz: stałej składającej się z cząsteczek białka stykających się w kilku miejscach i tworzących otwartą sieć i wypełnionej przez fazę ciekłą. Cząsteczki rozpuszczalnika przylegające do cząsteczek białka wykazują często wysoki stopień uporządkowania i są powiązane silnymi wiązaniami wodorowymi z powierzchniowymi grupami polarnymi białka. Natomiast rozpuszczalnik wewnątrz dużych kanałów, które mogą mieć szerokość do 2 mm, jest nieuporządkowany, tak jak to bywa w zwykłych cieczach (wg Blundella)

tych do obliczenia sum Fouriera, a *dm* jest odpowiednią minimalną odległością międzypłaszczyznową. Możliwa do osiągnięcia rozdzielczość dla kryształów białek jest ograniczona przez wyżej wspomniane nieuporządkowanie statyczne i termiczne. W praktyce krystalografii białek przyjęło się podawanie nominalnej rozdzielczości mapy gęstości elektronowej jako *dm*, tzn. minimalnej odległości międzypłaszczyznowej, dla której wartości czynników struktury *F* są włączone w szeregi Fouriera. Terminy „rozdzielczość 6 Å" (tzn. niska) i „rozdzielczość 2,0 Å" (tzn. wysoka) użyte przy podawaniu danych oznaczają, że dane dyfrakcyjne zostały zebrane do granicy tej właśnie odległości międzypłaszczyznowej.

rozdzielczość mapy

Po rozwiązaniu problemu fazowego można obliczyć syntezę Fouriera i wykreślić mapę gęstości elektronowej, na której krystalograf stara się zidentyfikować położenie atomów. Gdy cząsteczki są małe, mapa gęstości elektronowej może być obliczona przy wysokiej rozdzielczości 1 Å lub nawet wyższej i interpretacja takiej mapy jest zupełnie prosta. Wspomniane już nieuporządkowanie w krysztale białek może być tak duże, iż najlepsza możliwa rozdzielczość wyniesie jedynie 2 lub 3 Å. Zwykle można zidentyfikować główny łańcuch polipeptydowy, gdy rozdzielczość wynosi 3 Å, ale nie pozwala ona na rozróżnienie pojedynczych atomów.

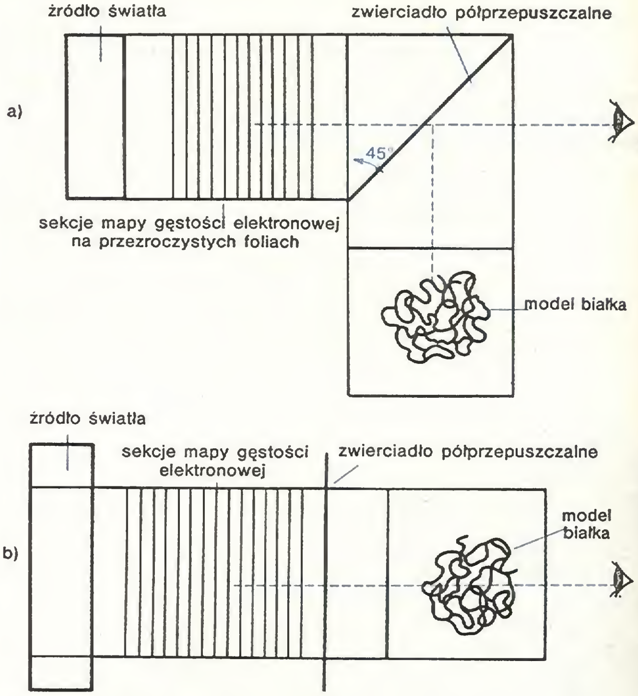
Przy rozdzielczości 2 Å jest możliwe zidentyfikowanie raczej grup niż indywidualnych atomów. Jeżeli jest znana szczegółowa struktura chemiczna białka, tj. kolejność aminokwasów wzdłuż łańcucha polipeptydowego, wtedy odpowiednie łańcuchy boczne można szybko rozpoznać. Jeśli taka wstępna budowa chemiczna jest znana tylko częściowo albo nie znana wcale, wtedy jest konieczne zidentyfikowanie poszczególnych aminokwasów wprost z mapy gęstości elektronowej. Cztery aminokwasy: tryptofan, fenyloalanina, tyrozyna i histydyna posiadają aromatyczne łańcuchy boczne, które można zwykle łatwo rozpoznać. Tryptofan wyróżnia się swoimi rozmiarami, a fenyloalanina jest zwykle ulokowana w obszarach hydrofobowych białka. Polarny charakter tyrozyny i histydyny sugeruje, że są one najprawdopodobniej ulokowane w obszarach hydrofilowych. Aminokwasy zawierające siarkę są zwykle łatwo rozróżnialne z powodu rozmiarów atomu siarki.

Wiele innych aminokwasów można grupować wg ich kształtu. Walina i treonina odgałęziają się od atomu C_β, podczas gdy leucyna, kwas asparaginowy i asparagina odgałęziają się od atomu C_γ itd. Uwagi powyższe ilustrują sposób, w jaki wiedza chemiczna może być spożytkowana w tego rodzaju rozważaniach. Interesującym przykładem takiej analizy może być interpretacja mapy gęstości elektronowej wykonana przez W. N. Lipscomba i współpracowników Harvard University w trakcie ich pracy nad strukturą enzymu trawiennego karboksypeptydazy *A*. Znano kolejność aminokwasów w kilku fragmentach enzymu, ale brakowało pewnych części i nie była znana kolejność fragmentów. W.N. Lipscomb zaproponował na podstawie badań rentgenograficznych kompletną kolejność dla 307 reszt aminokwasowych. Nieco później biochemik H. Neurath opublikował kolejność aminokwasów w tym enzymie, określoną metodami chemicznymi. Porównanie wyników wykazało, że fragmenty zostały poprawnie, co do kolejności, ulokowane na mapie gęstości elektronowej, ale liczba znanych fragmentów wynosiła 214 na ogólną liczbę 307 aminokwasów. W.N. Lipscomb oznaczał pozostałe 93 reszty i okazało się, że zidentyfikował on 60% z nich (tj. 56 reszt) poprawnie. 37 niepoprawnie określonych reszt wynikało z pomylenia waliny z treoniną, leucyny z kwasem asparaginowym i pewnych innych błędów. Zatem identyfikacja łańcuchów bocznych aminokwasów nie jest w żadnym razie jednoznaczna i trzeba w rozważaniach brać po uwagę czynniki stereochemiczne, otoczenie miejscowe i wiele innych aspektów.

struktura karboksypeptydazy *A*

Interesujące jest badanie struktury papainy — proteazy roślinnej, dla której chemicy błędnie określili kolejność aminokwasów, a co można było poprawić dopiero po przestudiowaniu mapy gęstości elektronowej.

Interpretację map gęstości elektronowej przeprowadza się za pomocą porównywania w specjalnych urządzeniach (rys. 3) modeli zbudowanych w skali



Rys. 3. Urządzenie optyczne do porównywania modelu molekularnego z mapami Fouriera gęstości elektronowej: a) lustro pod kątem 45° do map gęstości elektronowej, b) lustro równoległe do map gęstości elektronowej (wg T. L. Bundella, L. N. Johnsona)

2 cm/Å z otrzymanym obrazem gęstości elektronowej. Urządzenie takie, znane wśród krystalografów pod popularną nazwą skrzynki albo pudła Richardsa, zaprojektował F.M. Richards z Yale University. Arkusze z przezroczystego materiału z narysowanymi na nich cięciami trójwymiarowej mapy gęstości elektronowej montuje się pionowo w zaciemnionej części urządzenia. Następnie oświetla się wybrane cięcie mapy albo jego część i pozorny obraz oświetlonego obrazu rzuca się za pomocą półprzepuszczalnego lustra na obszar, w którym odbywa się budowanie modelu. Można w ten sposób studiować określoną część mapy, a budowanie modelu jest bardzo ułatwione dzięki porównaniu z pozornym obrazem gęstości elektronowej.

skrzynka Richardsa

Rozwinięto również metody komputerowe, które umożliwiają jednoczesną projekcję mapy i modelu na monitorze (odmiana kineskopu telewizyjnego). Pozwala to na konstruowanie precyzyjnego i stereochemicznie uzasadnionego modelu łańcucha polipeptydowego. Niedogodnością jest tutaj przedstawienie trójwymiarowej struktury w postaci dwuwymiarowego obrazu.

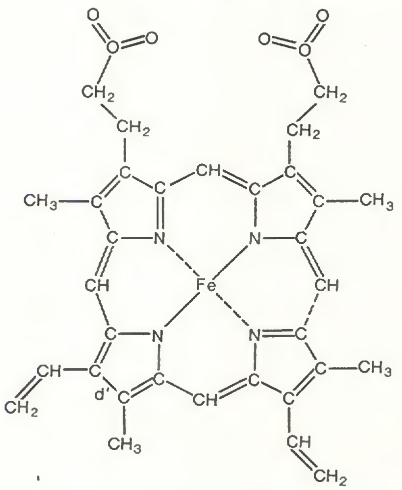
Lista analiz strukturalnych białek globularnych opublikowanych w literaturze naukowej do 1976 r. liczy około 90 pozycji (w tym m.in. 48 enzymów, 7 globin, 14 układów redoksowych, 5 hormonów, 8 immunoglobin). Liczba zbadanych białek jest niższa, gdyż wiele zespołów pracowało nad takimi samymi obiektami. Analizy te są wynikiem wytężonej pracy dużych zespołów uczonych przy zaangażowaniu znacznych środków materialnych. Uzyskano je w stosunkowo krótkim czasie — w ciągu ośmiu lat, poczynając od roku 1968.

Przykładowe analizy białek

hemoglobina

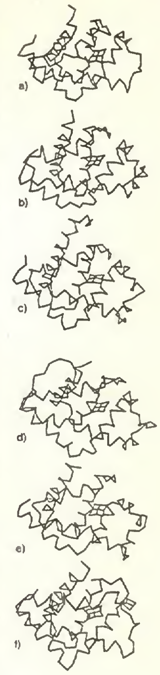
Przez lata hemoglobina była obiektem szczególnego zainteresowania chemików, biochemików i fizjologów. To łatwo dostępne białko stanowiące główny składnik czerwonych ciałek krwi (il. 115, tabl. 28), jest niezwykle ważne pod względem fizjologicznym, ponieważ zapewnia transport tlenu z płuc arteriami do tkanek i transport CO_2 żyłami z powrotem do płuc. Hemoglobina ssaków jest złożona z czterech podjednostek (→ Białka), które są parami identyczne. Każda jednostka zawiera 1 atom Fe usytuowany w środku grupy hemowej, zawierającej pierścień porfirynowy (rys. 4).

W. Conant stwierdził już w 1926 r., iż żelazo pozostaje w postaci dwuwartościowych jonów żelazawych podczas odwracalnego pobierania tlenu, natomiast w roztworze bez białka, żelazo grupy hemowej utlenia się raptownie do stanu trójwartościowego. Krzywa wiązania tlenu do hemoglobiny ma charakterystyczny kształt esowaty, który pokazuje, że powinowactwo do tlenu zależy od liczby cząsteczek tlenu już związanych.



Rys. 4. Grupa hemowa jest zasadniczym składnikiem hemoglobin, cytochromów i enzymów, takich jak kataliza i peroksydaza (wg T. L. Bundella, L. N. Johnsona)

Powinowactwo do tlenu zależy również od stężeń innych metabolitów, takich jak jony wodorowe i fosforany organiczne, które nie są związane z substratem.



Rys. 5. Porównanie różnych struktur mioglobin i hemoglobin pokazujące pofałdowanie charakterystyczne dla mioglobin. To pofałdowanie jest względnie stałe dla ssaków, owadów i globin wyizolowanych z pierścienic. Pokazane są tylko atomy α węgla: a) łańcuch α hemoglobiny końskiej, b) łańcuch β hemoglobiny końskiej, c) mioglobina kaszalota, d) hemoglobina minogi morskiej, e) hemoglobina gatunku *Chironomus*, f) hemoglobina gatunku *Glycera* (wg Love'a)

Hemoglobina stawia zatem szczególne problemy; po pierwsze, w jaki sposób otoczenie białkowe kontroluje własności atomu żelaza, i po drugie, jak związanie tlenu przez jedną z grup hemowych wpływa na powinowactwo do tlenu sąsiedniej grupy hemowej. Badania strukturalne dostarczyły odpowiedzi na te pytania. Analiza hemoglobiny i spokrewnionych z nią białek monomerycznych wykazała, że hem jest umieszczony w hydrofobowej kieszeni na powierzchni białka, zapewniającego środowisko o niskiej stałej dielektrycznej, w którym nasycenie tlenem jest faworyzowane, a utlenianie zachodzi trudno. Odwracalne łączenie się tlenu z grupą hemową powoduje zmianę stanu elektronowego żelaza — przejście z wysokospinowego stanu paramagnetycznego do niskospinowego stanu diamagnetycznego.

M. von Perutz w swoich badaniach rentgenograficznych hemoglobiny wykazał, że te współdziałające efekty są wyzwalane przez niewielkie ruchy atomu żelaza względem pierścienia porfirynowego, które mają miejsce w trakcie pobierania tlenu i towarzyszą zmianie stanu spinowego żelaza. Te przesunięcia atomu żelaza są przenoszone do innych części cząsteczki, tak że więzy trzymające białko w stanie dezoksy (odtlenionym) zostają rozluźnione i dzięki temu następna podjednostka może łatwiej pobrać tlen.

Strukturalne badania białek rozszerzyły naszą wiedzę o procesach ewolucji. Stwierdzono m.in., że lizozym ludzki i kurzy, cytochromy c: koński, makreli i tuńczyka oraz hemoglobiny: ludzka, końska, owadzia, minogi morskiej i ochotki tworzą szereg homologiczny, tzn. że rozwinęły się one ze wspólnego przodka — enzymu. Badania strukturalne wykazują, że zmiany aminokwasów, szczególnie w obszarze miejsca aktywnego, mają mały wpływ na strukturę. Szereg struktur hemoglobin i mioglobin pokazany jest na rys. 5. Ogólna struktura i umieszczenie hemu zostały w wysokim stopniu zachowane w trakcie ewolucji.

C. C. F. Blake Adv. in Protein Chem. 23, 59 (1968); D. M. Blow, T. A. Steitz Am. Rev. of Biochem. 39, 63 (1970); T. L. Blundell, L. N. Johnson *Protein Crystalography*, New York 1976; J. P. Glusker, K. N. Trueblood *Zarys rentgenografii kryształów*, Warszawa 1977; D. Sherwood *Crystals, X-rays and Proteins*, London 1976.

Kryształy ciekłe

Antoni Adamczyk

W artykule wykorzystano materiały udostępnione przez prof. Mariana Mięsowicza. Za zgodę na ich wykorzystanie autor składa prof. Mięsowiczowi serdeczne podziękowanie.

Ciekłe kryształy, odrębny stan materii

Ciekłe kryształy są oryginalnym rodzajem cieczy, której cechą wyróżniającą jest anizotropia własności fizycznych. Przez anizotropię rozumiemy, jak zwykle, zależność przebiegu zjawiska, tzn. zależność mierzonej wielkości fizycznej od kierunku, w którym wykonujemy pomiar lub od kierunku działania czynnika wywołującego zjawisko. Anizotropia — typowa cecha wielu kryształów, nigdy nie występuje w cieczach. Wyjątkiem są ciekłe kryształy, nazywane z tego powodu także cieczami anizotropowymi lub mezomorficznymi (o budowie pośredniej między cieczami a kryształami). Ciekłe kryształy powstają po stopieniu pewnych substancji krystalicznych lub wtedy, gdy w rozpuszczalniku o zdecydowanie określonym typie rozpuszczalności (polarnym lub niepolarnym) rozpuścimy substancję wykazującą obydwa typy rozpuszczalności, czyli tzw. substancję amfifilową. W pierwszym wypadku ciekłe kryształy nazywamy termotropowymi, w drugim zaś — liotropowymi. Tak termotropowe, jak i liotropowe ciekłe kryształy przechodzą do stanu zwykłej cieczy izotropowej po podgrzaniu ich do odpowiedniej temperatury. W chwili obecnej największe znaczenie praktyczne mają ciekłe kryształy termotropowe i do nich tylko ograniczymy nasz przegląd.

W trakcie ogrzewania kryształu substancji mogącej posiadać fazę mezomorficzną powstaje najpierw ciekły kryształ, a następnie w wyższej temperaturze przechodzi on w stan cieczy izotropowej. Przedział temperatu-



Fazy upadku kropli ciekłego kryształu na powierzchnię tej samej substancji

ry, w którym istnieje faza mezomorficzna, czyli przedział między temperaturą powstania ciekłego kryształu i temperaturą jego przejścia w ciecz izotropową, zależy od rodzaju substancji i może wynosić bądź ułamek stopnia, bądź też kilkadziesiąt, a nawet 100 stopni. W czasie ochładzania takiej substancji z cieczy izotropowej powstaje najpierw ciekły kryształ, który przy dalszym ochładzaniu ulega zwykłej krystalizacji. Temperatury wymienionych przejść fazowych, zarejestrowane w czasie ogrzewania i w czasie chłodzenia, na ogół nie pokrywają się ze sobą. Może się również zdarzyć, że po stopieniu kryształu powstaje od razu ciecz izotropowa i dopiero w czasie jej ochładzania pojawia się faza mezomorficzna (zawsze poniżej temperatury topnienia). Takie przejście do stanu mezomorficznego, występujące tylko w czasie chłodzenia, nosi nazwę przejścia monotropowego. Przez wiele lat po odkryciu ciekłych kryształów znano tylko substancje, których temperatura przejścia fazowego w stan mezomorficzny była znacznie wyższa od temperatury pokojowej, zwykle powyżej 100°C. Dopiero w ostatnich latach udało się otrzymać substancje, które występują w stanie mezomorficznym już w temperaturze pokojowej, a nawet w niższej. Odpowiednio dobrane mieszaniny mogą być ciekłymi kryształami w temperaturze —50°C, inne natomiast pozostają ciekłymi kryształami jeszcze w temperaturze +300°C.

przejście monotropowe

Ciekłe kryształy znalazły liczne zastosowania w różnych dziedzinach nauki i techniki. Same jednak własności tych niezwykłych cieczy stanowią pole bardzo atrakcyjnych prac badawczych.

Trochę historii

Ciekłe kryształy zostały odkryte już w 1888 r. przez austriackiego botanika F. Reinitzera, który zaobserwował pod mikroskopem polaryzacyjnym istnienie cieczy wykazujących dwójłomność optyczną. W rok później O. Lehmann stwierdził, że oryginalna ciecz jest nowym stanem materii i nazwał tę ciecz ciekłym kryształem.

Historia rozwoju badań ciekłych kryształów notuje trzy okresy rozkwitu. Pierwszy okres to lata wkrótce po ich odkryciu. Substancjami tymi zajmowali się wtedy głównie chemicy. Przeprowadzono syntezy wielu nowych związków chemicznych, mających fazy mezomorficzne i obserwowano prawidłowości przemian fazowych.

Następny okres — to dwudziestolecie międzywojenne, kiedy to pojawiły się m.in. prace G. Friedela na temat własności optycznych ciekłych kryształów. W wyniku tych prac G. Friedel zaproponował, do dziś obowiązujący, podział na ciekłe kryształy nematyczne, smektyczne i cholesterolowe. Wtedy także badano własności elektryczne i magnetyczne ciekłych kryształów i znaleziono m.in. tak ważną dla obecnych zastosowań łatwość orientowania się tych substancji w zewnętrznych polach elektrycznych i magnetycznych. Warto wzmiankować, że w owym okresie w Polsce (Kraków) intensywnie, jak na owe czasy, rozwijały się badania nad ciekłymi kryształami. Mamy tu na myśli pionierskie badania M. Jeżewskiego nad przenikalnością elektryczną ciekłych kryształów oraz badanie ich przewodnictwa elektrycznego i wpływu na nie pola elektrycznego i magnetycznego (M. Jeżewski, M. Mięsowicz). O polskich pracach nad anizotropią lepkości ciekłych kryształów będzie mowa dalej.

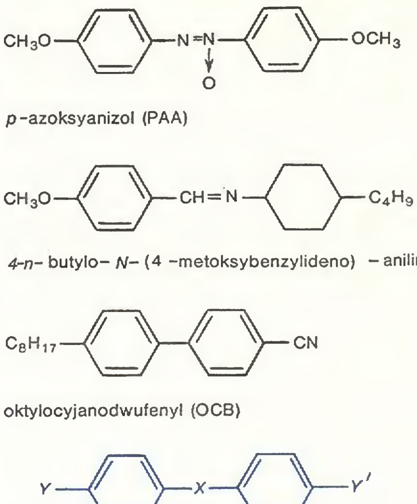
badanie własności optycznych, elektrycznych i magnetycznych

Obecny okres rozkwitu, trwający od ok. 10 lat, wiąże się z szerokim zastosowaniem ciekłych kryształów w technice i ze zrozumieniem roli tych substancji w organizmach żywych. Szeroko bada się i stosuje własności optyczne i elektrooptyczne ciekłych kryształów, bada się także ich udział w procesach biologicznych. Dotychczas zsyntezowano ok. 7 tys. związków ciekłokrystalicznych i ciągle pojawiają się doniesienia o odkryciu nowych takich substancji, a nawet całych ich grup.

Chemiczne własności substancji ciekłokrystalicznych i ich struktura molekularna

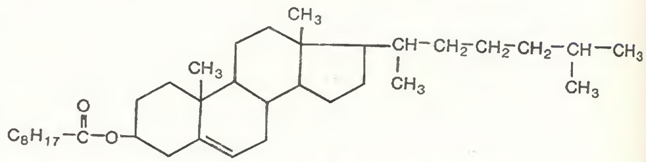
Istotną cechą substancji, które mogą występować jako ciekłe kryształy, jest wydłużony kształt ich molekuł, tak jak to widać na przykładach przedstawionych na rys. 1. U dołu rys. 1 zamieszczono schemat budowy

wydłużony kształt molekuł



Rys. 1. Wzory strukturalne kilku prostych molekuł substancji ciekłokrystalicznych i ogólny schemat ich budowy

częściej spotykanych związków ciekłokrystalicznych. Zwykle, chociaż nie jest to sztywna reguła, molekuła taka składa się z dwu pierścieni benzenowych, połączonych ze sobą bezpośrednio lub za pośrednictwem środkowej grupy atomów *X* i przeciwlegle (w położeniu para-) przyłączonych grup końcowych *Y* i *Y'*. Spotyka się także bardziej skomplikowaną budowę molekuł. Jako przykład mogą służyć popularne estry cholesterolu, np. pelargonian cholesterylu, którego wzór strukturalny jest przedstawiony na rys. 2.



Rys. 2. Wzór strukturalny molekuły cholesterolowego ciekłego kryształu — pelargonianu cholesterylu

Wydłużenie molekuł substancji mezomorficznych, czyli stosunek długości molekuł do ich szerokości, wynosi 6 i więcej. Na przykład, długość molekuły *p*-azoksyanizolu z rys. 1 wynosi 2 nm, zaś jej szerokość — 0,3–0,4 nm.

Pod względem struktury ciekłe kryształy dzielimy na trzy zasadnicze typy:

ciekłe kryształy nematyczne, w których molekuły pozostają względem siebie równoległe i nie mają żadnych dodatkowych ograniczeń we wzajemnym przesuwaniu się (rys. 3a);

struktura nematyczna

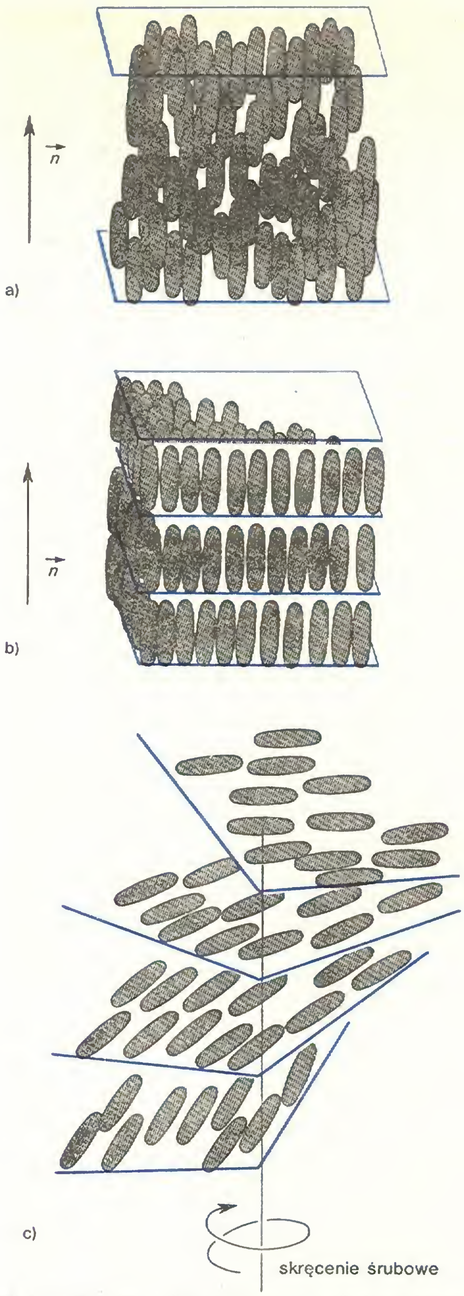
ciekłe kryształy smektyczne, w których równoległe względem siebie molekuły są rozmieszczone warstwami, zaś długie osie tych molekuł pozostają prostopadłe lub nieco nachylone względem tych warstw (rys. 3b);

struktura smektyczna

ciekłe kryształy cholesterolowe, w których molekuły są także rozmieszczone warstwowo, ale w ten sposób, że ich długie osie, pozostając wzajemnie równoległe, są jednocześnie równoległe do warstw. Wszystkie molekuły z warstwy następnej są skręcone o pewien stały kąt względem molekuł leżących w warstwie poprzedniej (rys. 3c). Wynikiem takiego skręcenia jest

struktura cholesterolowa

struktura śrubowa



Rys. 3. Struktura ciekłych kryształów: a) typ nematyczny, b) typ smektyczny, c) typ cholesterolowy

swoista struktura śrubowa o niesłychanie ciekawych własnościach optycznych.

Każdy z tych trzech typów struktur rozpada się na kilka rodzajów, różniących się szczegółami w uporządkowaniu molekuł. Tak więc, wyróżnia się trzy rodzaje nematyków, dwa rodzaje cholesterolowych ciekłych kryształów i kilkanaście rodzajów smektyków.

Geneza słowa „nematyk" bierze początek od greckiego *nema* — 'nić'. Nazwa struktury nematycznej została wyprowadzona od charakterystycznych, ruchliwych nici widocznych pod mikroskopem i będących zaburzeniami struktury. Smektyki wzięły nazwę od greckiego słowa *smektos* — 'mydlany', ponieważ ich mętność i lepkość przypomina stężone roztwory mydła. Nazwa cholesterolowych ciekłych kryształów pochodzi od estrów cholesterolu, u których ten typ struktury po raz pierwszy zaobserwowano.

polimezomorfizm

Niektóre substancje mają więcej niż jedną fazę ciekłokrystaliczną. Jest to tzw. polimezomorfizm. Istnieją substancje, w których występuje kilka faz smektycznych, u innych występuje najpierw faza smektyczna, zaś w wyższych temperaturach — nematyczna lub cholesterolowa itd. W toku ogrzewania substancji faza smektyczna, jeżeli występuje w ogóle, to występuje zawsze jako pierwsza, zaś fazy nematyczna i cholesterolowa pojawiają się dopiero w wyższych temperaturach i poprzedzają przejście substancji do stanu cieczy izotropowej. W substancjach jednoskładnikowych fazy nematyczna i cholesterolowa nawzajem się wykluczają, tzn., że jeżeli w jakiejś substancji występuje faza cholesterolowa, to tej substancji nie można przeprowadzić do fazy nematycznej w drodze zmian temperatury.

przejścia fazowe

Temperatury przejść fazowych kilku substancji ciekłokrystalicznych (w °C)

Substancja	T_{K-N}	T_{K-S}	T_{S-CH}	T_{S-N}	T_{N-I}	T_{CH-I}
p-azoksyanizol (PAA)	118	—	—	—	135	—
4-*n*-butylo-*N*-(4-metoksybenzylideno)-anilina (MBBA)	22	—	—	—	47	—
pelargonian cholesterylu (CN)	—	74	81	—	—	93
oktylocyjanodwufenyl (OCB)	—	21	—	32	40	40

Poniżej przedstawiono tabelę zawierającą temperatury przejść fazowych kilku popularnych substancji ciekłokrystalicznych. Obok nazwy każdej z nich umieszczono, zwyczajowo używane, skróty ich nazw angielskich. Temperaturę przejścia z fazy stałej do fazy nematycznej oznaczono przez T_{K-N}, z fazy nematycznej do izotropowej przez T_{N-I}, z fazy stałej do smektycznej przez T_{K-S}, z fazy smektycznej do cholesterolowej przez T_{S-CH}, z fazy smektycznej do nematycznej przez T_{S-N} oraz z fazy cholesterolowej do izotropowej przez T_{CH-I}. Kreska w kolumnie oznacza, że odpowiednie przejście fazowe nie występuje w danej substancji.

Model domenowy a model ośrodka ciągłego (model continuum)

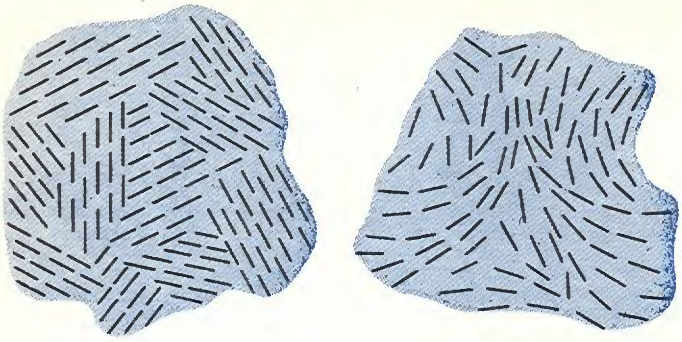
Anizotropowe własności ciekłych kryształów wynikają ze wzajemnie równoległego ustawienia się wydłużonych molekuł. Jest intuicyjnie oczywiste, że grupa równoległych względem siebie molekuł będzie miała inne własności wzdłuż długich osi molekuł, inne zaś w kierunku osi krótkich. W dużej objętości ciekłego kryształu kierunek uporządkowania molekuł ma charakter jedynie lokalny. Można więc przyjąć hipotezę, że molekuły tworzą ziarniste skupiska (domeny) i kierunek uporządkowania molekuł w sąsiednich domenach jest różny. Na podstawie danych doświadczalnych obliczono, że liczba jednakowo zorientowanych molekuł, tworzących hipotetyczną domenę, jest rzędu 10^6. Rozmiary domen są więc porównywalne z długością fali świetlnej i dlatego światło przechodzące przez warstwę ciekłego kryształu jest silnie rozpraszane. Ten mechanizm może tłumaczyć istnienie mętności charakterystycznej dla wszystkich ciekłych kryształów.

model domenowy

Molekuły w cieczach mają dużą swobodę przemieszczania się w przestrzeni, mogą więc bardzo łatwo przechodzić z jednej domeny do drugiej. Ta wymiana molekuł powoduje, że granice domen są silnie rozmyte. Wydaje się zatem, że bardziej odpowiednim modelem nematycznego ciekłego kryształu jest model ośrodka ciągłego (model continuum). W tym modelu kierunek uporządkowania molekuł zmienia się w sposób ciągły w przestrzeni, mimo że nadal zostaje zachowane dalekie uporządkowanie między tymi molekułami. „Dalekie" znaczy tutaj rozciągające się na od-

model continuum

ległości duże w porównaniu z rozmiarami samych molekuł. Różnicę między modelem domenowym i modelem continuum widać ze szkiców, przedstawionych na rys. 4. Wiele zjawisk optycznych i hydrodynamicznych świadczy o poprawności raczej modelu continuum, chociaż na gruncie modelu domenowego łatwo można wytłumaczyć kolektywny charakter oddziaływania ciekłego kryształu z polami elektrycznymi i magnetycznymi. Chodzi o to, że energia oddziaływania pojedynczej molekuły z polem jest o kilka rzędów wielkości za mała, aby móc przeciwdziałać niszczącemu działaniu ruchów termicznych. Jeżeli jednak przyjąć, że pole działa nie na jedną molekułę, lecz na cały ich zbiór zawarty w domenie, to już można wyjaśnić, dlaczego orientujące działanie wywierają pola o stosunkowo niewielkich natężeniach (zob. niżej).



Rys. 4. Model domenowy i model ośrodka ciągłego (continuum)

Równoległe uporządkowanie molekuł, charakterystyczne dla ciekłych kryształów, nie jest idealne. Drgania termiczne molekuł powodują odchylenia od równoległości i jedynie średni kierunek ich długości osi pozostaje niezmieniony. Ten średni kierunek oznacza się za pomocą wektora $\vec{n}$ o długości jednostkowej, zwanego dyrektorem (od ang. *direction* — 'kierunek'). Im wyższa jest temperatura ciekłego kryształu, tym większa amplituda drgań molekuł i tym gorsze uporządkowanie. Stopień tego uporządkowania można wyrazić liczbowo za pomocą tzw. parametru uporządkowania S. Dla nematyków:

stopień uporządkowania

$$S = \tfrac{1}{2}\langle 3\cos^2\theta - 1\rangle .$$

Znak $\langle\ \rangle$ oznacza uśrednienie względem wszystkich molekuł, należących do zbioru o jednakowej orientacji. Kąt θ jest utworzony między dyrektorem $\vec{n}$ i długą osią jednej wybranej molekuły. Gdy uporządkowanie znika (jak w cieczach izotropowych), czyli gdy każda wartość kąta θ jest jednakowo prawdopodobna, to średnia wartość $\cos^2\theta = {}^1/_3$ i parametr $S = 0$. Dla idealnego uporządkowania θ jest zawsze równe zeru i parametr $S = 1$. W nematykach, w zależności od temperatury, S ma wartość od ok. 0,4 do ok. 0,7.

Tekstury molekularne

Teksturą nazywamy sposób ułożenia krystalitów w materiale polikrystalicznym, ułożenie włókien w materiale kompozytowym itd. W wypadku ciekłych kryształów przez teksturę rozumiemy sposób przestrzennego ustawienia długich molekuł w cienkiej warstwie ciekłokrystalicznej. Jest to zatem tekstura molekularna. Tekstur molekularnych w ciekłych kryształach jest wiele rodzajów; tutaj wspomnimy jedynie o kilku najważniejszych z punktu widzenia zastosowań ciekłych kryształów.

Jeżeli w cienkiej warstwie nematyka, zawartej między dwiema powierzchniami szkła lub innego materiału stałego, długie osie molekuł leżą równolegle do obu powierzchni ograniczających i ponadto, gdy wszystkie molekuły mają jednakowy kierunek, to takie uporządkowanie nosi nazwę tekstury planarnej. Z tekstury planarnej można łatwo otrzymać teksturę skręconego nematyka (ang. *twisted nematics*). W tym celu obie płytki, między którymi znajduje się warstwa planarnie zorientowanego nematyka, należy skręcić względem siebie o 90°. Po takim obrocie kierunek molekuł przylegających do jednej powierzchni tworzy kąt prosty z kierunkiem molekuł leżących na drugiej powierzchni. Nematyk w głębi warstwy zostanie zdeformowany w sposób ciągły z wytworzeniem układu śrubowego, zupełnie podobnego do struktury cholesterolowej (ćwierć skoku śruby). Opisana tekstura jest obecnie najczęściej stosowana w układach elektrooptycznych.

tekstura planarna

Prostopadłe ustawienie molekuł względem powierzchni ograniczających warstwę ciekłego kryształu nosi nazwę tekstury homeotropowej. Teksturę tę można wytworzyć w nematykach i smektykach.

tekstura homeotropowa

Tekstura, która występuje tylko w cholesterolowych ciekłych kryształach i która jest niezbędna, jeśli chcemy wykorzystać oryginalne własności optyczne i termooptyczne tych substancji, nosi nazwę tekstury planarnej Grandjeana. Molekuły w warstwie cholesterolowej o tej teksturze pozostają stale równoległe do powierzchni ścianek ograniczających, zaś ich kierunek w przestrzeni ulega zmianie zgodnie ze spontanicznym skręceniem śrubowym struktury cholesterolowej. Innymi słowy, tekstura Grandjeana to takie uporządkowanie, w którym oś linii śrubowej jest zawsze prostopadła do płaszczyzny podłoża.

tekstura Grandjeana

Wreszcie ostatnią teksturą, o której tutaj wspomnimy, jest tekstura konfokalna, tworzona przez smektyki i cholesterolowe ciekłe kryształy. Substancje e wskutek warstwowej budowy mają tendencję do grupowania się w makroskopowe bryłki o dość skomplikowanych kształtach. Kształt tych bryłek wynika ze współogniskowego (konfokalnego) połączenia elips i hiperbol i stąd nazwa tej tekstury. Cienka warstwa ciekłego kryształu, w której została wytworzona tekstura konfokalna, ma mleczne, silne zmętnienie, znikające, gdy ciekły kryształ zostanie przeprowadzony w dowolną inną teksturę.

tekstura konfokalna

Niektóre zagadnienia optyki ciekłych kryształów

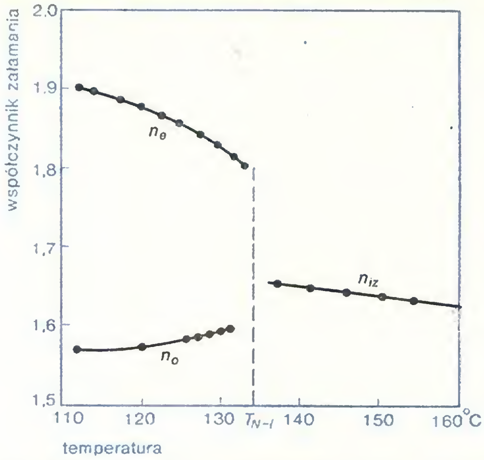
Ciekłe kryształy są cieczami dwójłomnymi optycznie. Znaczy to, że ich współczynnik załamania światła zależy od kierunku drgań wektora elektrycznego fali świetlnej względem osi optycznej cienkiej warstwy ciekłego kryształu. Stanowi to curiosum, ponieważ żadne inne ciecze nie wykazują dwójomności, jeżeli nie są poddane specjalnym działaniom deformującym.

Kiedy dwie płytki szklane, między którymi znajduje się cienka warstwa planarnie zorientowanego nematyka, umieścimy między dwoma skrzyżowanymi polaroidami, to początkowo ciemne pole widzenia zostanie silnie rozjaśnione. Rozjaśnienie to jest największe wtedy, gdy kąt między osią polaroidu a kierunkiem uporządkowania molekuł w nematyku jest równy 45°, najmniejsze — gdy kąt ten jest równy zeru. Gdy warstwę nematyka, umieszczoną między wspomnianymi płytkami szklanymi, zorientujemy homotropowo, to pole widzenia pozostanie ciemne bez względu na obroty płytki. Widać więc, że warstwa zorientowanego nematyka ma własności optycznie takie, jak płytka wycięta z kryształu jednoosiowego. Oś optyczna jest równoległa do warstwy zorientowanej planarnie i prostopadła do warstwy o teksturze homotropowej, czyli jest zawsze równoległa do kierunku długich osi molekuł w nematyku.

dwójłomność ciekłych kryształów

Współczynnik załamania wiązki światła spolaryzowanego, w której drgania wektora świetlnego są równoległe do osi optycznej, nazywamy współczynnikiem

nadzwyczajnym i oznaczamy symbolem n_e. Gdy drgania wektora świetlnego są prostopadłe do osi optycznej, to załamanie wiązki określa współczynnik zwyczajny, oznaczany n_o. Kryształ jest dodatni optycznie, gdy $n_e - n_o > 0$. Nematyki i smektyki są właśnie kryształami optycznie dodatnimi. Ujemne optycznie są tylko cholesterolowe ciekłe kryształy. Na rysunku 5



Rys. 5. Zależność współczynników załamania n_e i n_o w PAA od temperatury

przedstawiono zależność współczynników załamania nematyka od temperatury. Widać, że ze wzrostem temperatury, gdy maleje parametr uporządkowania S, maleje również anizotropia optyczna.

aktywność optyczna

Niezwykłe własności optyczne mają cholesterolowe ciekłe kryształy. Rozmieszczenie molekuł w tych substancjach na linii śrubowej powoduje, że mają one bezkonkurencyjną w całej przyrodzie skręcalność płaszczyzny drgań światła spolaryzowanego (aktywność optyczną). Jeżeli w innych substancjach aktywnych optycznie płaszczyzna drgań światła obraca się o kilka, a najwyżej o kilkanaście stopni na 1 mm drogi, to na tej samej drodze w cholesterolowych ciekłych kryształach płaszczyzna drgań ulega skręceniu nawet o kilkadziesiąt tysięcy stopni. Ponadto skręcalność ta może być łatwo sterowana przez zewnętrzne oddziaływania termiczne, chemiczne, elektryczne, magnetyczne i mechaniczne.

odbicie światła

Piękne zabarwienie, tak charakterystyczne dla cholesterolowych ciekłych kryształów, jest wywołane przez zjawisko selektywnego odbicia światła od ich powierzchni. Długość skoku śruby cholesterolowej jest porównywalna z długością fali światła i padające na warstwę cholesterolową światło białe odbija się w postaci wiązki, która padając na ekran daje obraz tęczy. Barwa światła selektywnie odbitego zależy od kąta padania światła białego i od kąta obserwacji. Selektywne odbicie ma więc charakter odbicia Bragga i jest podobne do odbicia promieni rentgenowskich od zwykłych kryształów. Barwy światła odbitego selektywnie są jaskrawe i czyste tylko wtedy, gdy warstwa odbijająca ma teksturę grandjeanowską.

dichroizm

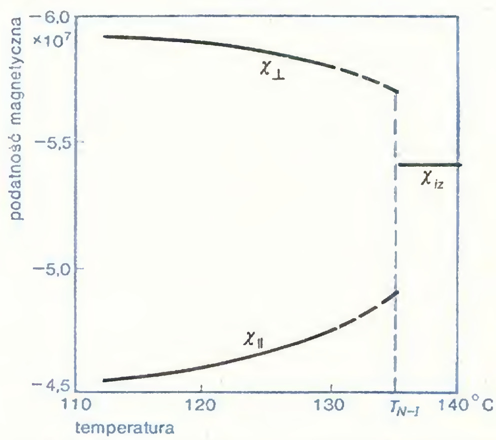
Ze śrubową strukturą cholesterolowych ciekłych kryształów wiąże się również zjawisko tzw. dichroizmu kołowego. Polega ona na tym, że w warstwie ciekłego kryształu, mającego teksturę planarną Grandjeana, światło może rozchodzić się tylko w postaci wiązki spolaryzowanej kołowo. Skrętność polaryzacji kołowej jest przeciwna do skrętności śruby cholesterolowej, tzn. że jeżeli śruba cholesterolowa jest lewoskrętna, to przechodzące przez warstwę światło jest spolaryzowane kołowo prawoskrętnie i odwrotnie. Jak wiadomo, światło spolaryzowane liniowo można rozłożyć na dwie składowe o przeciwnej polaryzacji kołowej. Dzięki temu wiązka światła liniowo spolaryzowanego, padająca na warstwę cholesterolowego ciekłego kryształu, zostanie podzielona na dwie wiązki kołowo spolaryzowane o przeciwnych skrętnościach, przy czym jedna z nich przechodzi przez ciekły kryształ, druga zaś zostaje całkowicie odbita.

Anizotropia diamagnetyczna i dielektryczna. Anizotropia przewodnictwa elektrycznego

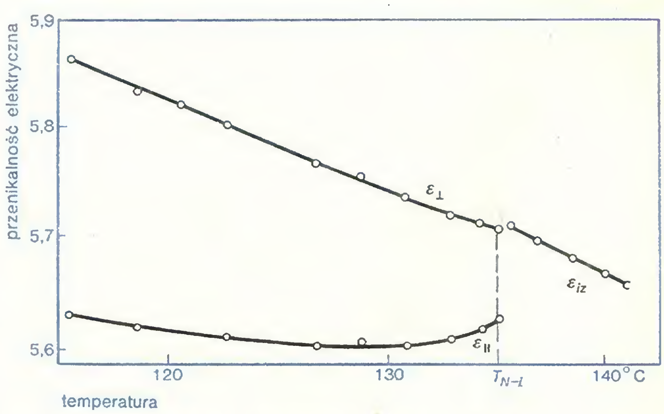
Ciekłe kryształy są diamagnetykami wykazującymi anizotropię diamagnetyczną wynikającą, w głównej mierze, z istnienia w ich molekułach pierścieni benzenowych. Diamagnetyzm pierścienia benzenowego ma szczególne własności: elektrony walencyjne, tworzące wiązanie typu π między atomami węgla mają swobodę ruchu w obszarze całego pierścienia, który dzięki temu może być przyrównany do pętli wykonanej z nadprzewodnika. W takiej pętli zewnętrzne pole magnetyczne indukuje silny prąd, którego własne pole jest skierowane przeciwnie do pola indukującego. Pierścień benzenowy ma tendencję do „uciekania" z pola, czyli do ustawiania się tak, aby strumień pola zewnętrznego przez jego płaszczyznę był minimalny. Pierścień benzenowy przyjmuje więc położenie równoległe do linii sił pola. Taką właśnie orientację względem pola przyjmują wszystkie ciekłe kryształy, złożone z molekuł zawierających pierścienie benzenowe.

anizotropia diamagnetyczna

Anizotropię diamagnetyczną ciekłych kryształów opisujemy ilościowo przez podanie różnicy podatności magnetycznych $\Delta\chi = \chi_{\parallel} - \chi_{\perp}$, gdzie $\chi_{\parallel}$ i $\chi_{\perp}$ oznaczają podatności magnetyczne ciekłego kryształu, w którym molekuły są zorientowane — odpowiednio — równoległe i prostopadłe do linii sił pola magnetycznego. W ciekłych kryształach wartości $\Delta\chi$ są rzędu 10^{-7}. Ze wzrostem temperatury wszystkie cechy anizotropowe ciekłych kryształów zanikają, maleje także anizotropia diamagnetyczna (rys. 6)



Rys. 6. Zależność podatności magnetycznej PAA od temperatury



Rys. 7. Zależność przenikalności elektrycznej PAA od temperatury

anizotropia dielektryczna

Ciekłe kryształy wykazują również anizotropię dielektryczną, tzn. ich przenikalność elektryczna, gdy molekuły są ustawione równolegle do kierunku pomiarowego pola elektrycznego ($\varepsilon_{\parallel}$), jest zawsze inna, niż przenikalność elektryczna ($\varepsilon_{\perp}$) próbki, w której molekuły są ustawione prostopadle do kierunku pola. Anizotropia dielektryczna $\Delta\varepsilon = \varepsilon_{\parallel}-\varepsilon_{\perp}$ maleje ze wzrostem temperatury (rys. 7). Wartość anizotropii $\Delta\varepsilon$, a nawet i jej znak, zależą także od częstości pola elektrycznego.

Molekuły w ciekłych kryształach mają z reguły niezerowy trwały elektryczny moment dipolowy. Gdy wektor momentu dipolowego jest skierowany prostopadle względem długiej osi molekuły, to ciekły kryształ w polu o małej częstości wykazuje ujemną anizotropię dielektryczną, czyli $\Delta\varepsilon = \varepsilon_{\parallel}-\varepsilon_{\perp} < 0$. Takimi substancjami są np. PAA i MBBA (rys. 1). I odwrotnie, kiedy wektor momentu dipolowego jest równoległy do długiej osi molekuły, to ciekły kryształ, jak np. OCB, ma dodatnią anizotropię dielektryczną, czyli $\varepsilon_{\parallel}-\varepsilon_{\perp} > 0$.

Ujemna anizotropia dielektryczna jest z reguły niewielka, zazwyczaj mniejsza od jedności, natomiast wartość $\Delta\varepsilon$ w ciekłych kryształach o dodatniej anizotropii dielektrycznej może dochodzić nawet do 20.

Każdy układ fizyczny osiąga stan równowagi wtedy, gdy jego energia swobodna jest możliwie najmniejsza. Jeżeli np. krawędzie okładek kondensatora powietrznego zetkniemy z powierzchnią dielektryka ciekłego, to po przyłożeniu do okładek napięcia elektrycznego zauważymy, że ciecz dielektryczna jest wciągana między okładki. Kondensator „pije" tę ciecz tym obficiej, im większa jest przenikalność elektryczna cieczy i im większe jest natężenie pola w kondensatorze. Przy stałym ładunku na okładkach kondensatora jego energia będzie odwrotnie proporcjonalna do wartości przenikalności elektrycznej cieczy wprowadzonej między okładki.

Jeżeli teraz przestrzeń między okładkami wypełnimy cieczą anizotropową dielektrycznie i do kondensatora przyłożymy napięcie, to — dzięki tendencji układu do obniżania swej energii swobodnej — ciecz przyjmie taką konfigurację, aby pole „widziało" jej największą przenikalność elektryczną. Zatem ciekły kryształ o ujemnej anizotropii dielektrycznej ustawia się w polu elektrycznym tak, że długie osie molekuł są prostopadłe do linii sił pola, powstaje tekstura planarna. Z kolei długie osie molekuł w ciekłych kryształach o dodatniej anizotropii dielektrycznej przyjmują położenie równoległe do linii sił pola elektrycznego (tekstura homeotropowa).

Taki sposób orientowania się ciekłych kryształów w polu elektrycznym jest bardzo często zaburzany przez przepływ prądu jonowego, któremu towarzyszą elektrohydrodynamiczne, często bardzo burzliwe przepływy ciekłego kryształu. Przepływy elektrohydrodynamiczne, towarzyszące prądowi jonowemu

orientacja molekuł w polu elektrycznym

mogą powodować orientację długich osi molekuł zawsze w kierunku wektora natężenia pola elektrycznego, niezależnie od znaku anizotropii dielektrycznej. Fakt, że stałe pole elektryczne orientuje molekuły nematyka o ujemnej anizotropii dielektrycznej równolegle do kierunku linii sił (a nie prostopadle, jak nakazywałyby reguły elektrostatyki) zaobserwowany już został w wyżej wymienionych pracach M. Jeżewskiego i M. Mięsowicza.

Uporządkowanie warstwy ciekłego kryształu przez zewnętrzne pole elektryczne lub magnetyczne jest osiągalne dopiero wtedy, gdy natężenie pola orientującego przekroczy pewną wartość progową. Ta wartość progowa zależy od wielkości anizotropii diamagnetycznej i dielektrycznej, od wielkości odpowiednich współczynników sprężystości ciekłego kryształu i od grubości jego warstwy. Jeżeli tę grubość oznaczymy przez d, to wartości progowe natężeń pola magnetycznego H_t i pola elektrycznego E_t spełniają przybliżone zależności $H_t d = C_1$ oraz $E_t d = C_2$, gdzie C_1 i C_2 są stałymi zależnymi od rodzaju ciekłego kryształu.

przewodnictwo elektryczne

Przewodnictwo elektryczne ciekłych kryształów ma, prawie wyłącznie, charakter jonowy; ich przewodność σ zawarta jest przeważnie w przedziale od 10^{-8} do $10^{-12}\,\Omega^{-1}\mathrm{cm}^{-1}$. Jak łatwo się domyśleć, również i przewodnictwo elektryczne ma w ciekłych kryształach charakter anizotropowy. Jeżeli przez $\sigma_{\parallel}$ oznaczymy przewodność próbki, w której jony poruszają się równolegle do kierunku długich osi molekuł, zaś przez $\sigma_{\perp}$ przewodność uwarunkowaną ruchem jonów w kierunku prostopadłym do tych osi, to stwierdzamy, że w nematykach zawsze $\sigma_{\parallel} > \sigma_{\perp}$. Stąd wniosek, że jony napotykają mniejszy opór, jeżeli ich ruch jest zgodny z kierunkiem długich osi molekuł. Dla określonego nematyka stosunek obu przewodności jest zazwyczaj bliski 1,5 i niewiele zależy od ich wartości.

W smektykach relacja między przewodnościami jest odwrotna. W tych ciekłych kryształach ruch jonów jest łatwiejszy wzdłuż warstw smektycznych, a nie wzdłuż długich osi molekuł (jak pamiętamy, molekuły w smektykach są prostopadłe lub prawie prostopadłe do warstw). Wobec tego w smektykach przewodność elektryczna $\sigma_{\parallel} < \sigma_{\perp}$.

Orientacja molekuł na granicy ciekły kryształ–ciało stałe

Ciekłe kryształy są jedynymi substancjami, których własności optyczne mogą być radykalnie zmieniane już przez słabe oddziaływania powierzchniowe. Ich ciekłość, dalekie uporządkowanie molekularne i anizotropia optyczna powodują, że dokładnie odwzorowują one stany powierzchni ciał stałych i pozwalają te stany wizualizować. Ta zdolność ciekłych kryształów została wykorzystana do wytwarzania określonych tekstur molekularnych, głównie w cienkich warstwach wchodzących w skład elementów elektrooptycznych.

Odpowiednia obróbka powierzchni szkła, półprzewodnika czy metalu może spowodować, że molekuły ciekłych kryształów będą przyjmowały położenie równoległe do tych powierzchni i będą leżały w wyznaczonym kierunku lub też będą do tych powierzchni prostopadłe. W pierwszym wypadku wytwarzamy teksturę planarną, w drugim zaś — homeotropową.

wytwarzanie tekstury planarnej

Teksturę planarną można uzyskać przez jednokierunkowe polerowanie powierzchni ciał stałych papierem, tkaniną itp. lub przez ukośne napylanie warstw dielektryków i metali. Mikrorowki, powstające na powierzchniach w czasie polerowania lub też wydłużone wysepki substancji napylanych ukośnie ustawiają molekuły ciekłego kryształu, naniesionego na tak przygotowaną powierzchnię, w kierunku równoległym do kierunku polerowania lub napylania. Molekuły kontaktujące się z powierzchnią stałą pociągają za sobą molekuły z wnętrza warstwy ciekłokrystalicznej tak, że jednorodnie zorientowana warstwa może mieć grubość nawet 1 mm.

wytwarzanie tekstury homeotropowej

Aby wytworzyć teksturę homeotropową należy spowodować, aby pierwsza warstwa molekuł, kontaktująca się z powierzchnią ciała stałego, przyjmowała względem tej powierzchni położenie prostopadłe. W tym celu na powierzchnię stałą nanosi się bardzo cienką warstewkę substancji o wydłużonych molekułach, mających na jednym końcu grupę polarną, pn. warstewkę lecytyny lub odpowiedniego detergentu. Grupa ta wiąże się z powierzchnią podłoża, reszta zaś molekuły wystaje prostopadle nad powierzchnią. Molekuły ciekłego kryształu naniesionego na tak przygotowaną powierzchnię ustawiają się równolegle do molekuł warstwy inicjującej, tworząc teksturę homeotropową.

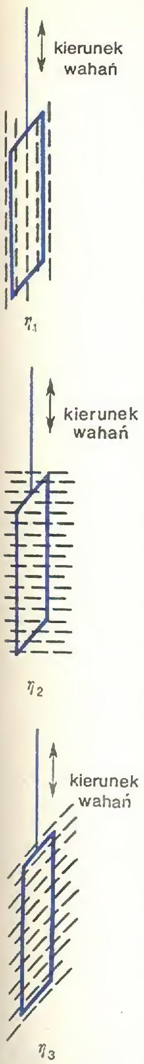
Ciekłokrystaliczne urządzenia elektrooptyczne działają głównie na zasadzie deformowania za pomocą pola elektrycznego tekstury ciekłego kryształu, narzuconej przez opisane oddziaływania powierzchniowe. Po ustaniu działania pola te same oddziaływania powierzchniowe przywracają pierwotną teksturę warstwie ciekłego kryształu.

Anizotropia lepkości. Trzy główne współczynniki lepkości nematyków

Pomiary wpływu orientacji molekuł w ciekłych kryształach na współczynniki lepkości tych substancji prowadzone były już dawno. Stosowano wtedy tradycyjne metody wiskozymetryczne. Badano m.in. wpływ silnego pola magnetycznego, o indukcji rzędu 1 T, na prędkość przepływu cieczy nematycznych przez rurki kapilarne o średnicy ok. 0,1 mm. Wyniki tych eksperymentów, wskutek wadliwego doboru warunków, były jednak negatywne: włączanie pola, prostopadłego do osi przepływu, nie zmieniało prędkości cieczy w kapilarze. Jeszcze w r. 1934 twierdzono, że orientacja molekuł nie ma wpływu na współczynnik lepkości.

W 1933 r. w Krakowie M. Mięsowicz podjął systematyczne badania własności reologicznych ciekłych kryształów. Wysunął rewolucyjną wówczas ideę anizotropii lepkości tych substancji, a następnie w opracowanym przez siebie bardzo pomysłowym i oryginalnym eksperymencie wykrył tę anizotropię i zmierzył odpowiednie współczynniki lepkości. Zastosował specjalną technikę pomiaru, której nowość polegała na przyjęciu prostokątnej geometrii przepływu i na zastosowaniu bardzo małych gradientów prędkości. Zbudowano urządzenie, w którym cienka płytka szklana (0,1 mm) o powierzchni kilku cm² była zawieszana w ciekłokrystalicznym PAA lub *p*-azoksyfenetolu. Płytka mogła wykonywać bardzo powolne wahania w płaszczyźnie pionowej (rys. 8). Z wielkości tłumienia wahań można było wyznaczyć odpowiednie współczynniki lepkości. Badana ciecz była orientowana przez zewnętrzne pole magnetyczne. W odróżnieniu od poprzednich rezultatów stwierdzono bardzo silny wpływ pola magnetycznego na lepkość obu ciekłych kryształów. Określono trzy różne współczynniki lepkości: η_1 — gdy molekuły leżały wzdłuż poruszającej się płytki, η_2 — gdy były ustawione do niej prostopadle, i η_3 — gdy leżały w poprzek tej płytki. Z pomiarów dla PAA (w temperaturze 122°C) uzyskano wartości: $\eta_1 = 2{,}4$ m·Pa·s, $\eta_2 = 9{,}2$ m·Pa·s, $\eta_3 = 3{,}4$ m·Pa·s. Jak widać, anizotropia współczynnika lepkości jest bardzo duża ($\eta_2/\eta_1 \approx 4$). Okazało się, że jest największą ze wszystkich anizotropii. Poprzednio obserwowany brak efektu przy przepływie przez cienkie kapilary w porównaniu z tymi rezultatami był dowodem silnej orientacji powodowanej przez przepływ z gradientem prędkości.

W ostatnich latach została stworzona ogólna hydrodynamika teoretyczna ciekłych kryształów, oparta na modelu continuum. W równaniach tej hydrodynamiki występuje sześć współczynników Lesliego. Za ich pomocą można wyrazić trzy współczynniki lepkości Mięsowicza, bo tak są nazywane w literaturze współczynniki η_1, η_2 i η_3, a także inne współczynniki lepkości występujące przy skomplikowanych deformacjach ciekłych kryształów.



Rys. 8. Położenie molekuł nematyka względem poruszającej się płytki przy wyznaczaniu trzech współczynników lepkości

Kilka zagadnień dynamiki

W ostatnich latach (po 1965 r.) okazało się, że wiele ciekawych zjawisk z fizyki ciekłych kryształów można opisać ilościowo przy użyciu przedstawionej wyżej anizotropii lepkości. Chodzi tu o szeroki zakres zjawisk, w których rolę odgrywają ruchy molekularne, w tym głównie: rozpraszanie światła; anizotropia absorpcji i anizotropia prędkości fal ultradźwiękowych; orientacja przez ruch cieczy; orientacja przez przepływ prądu elektrycznego; przewodnictwo cieplne; dyfuzja i samodyfuzja.

Ze względu na ograniczone ramy tego artykułu, dla przykładu zajmiemy się tylko zjawiskiem rozpraszania światła oraz zjawiskiem samodyfuzji.

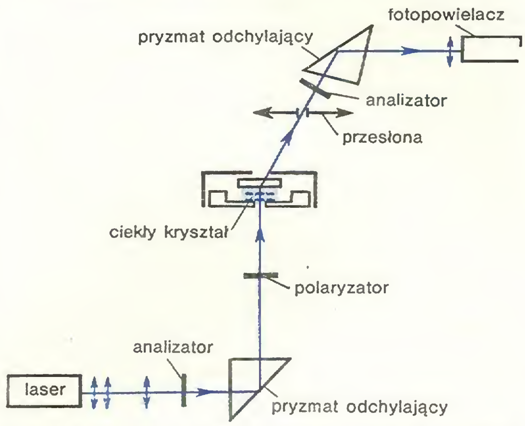
rozpraszanie światła

Niedawno stwierdzono, że rozpraszanie światła w ciekłych kryształach nematycznych daje wiele ciekawych informacji na temat ich struktury. Okazało się także, co jest pewną niespodzianką, że badanie rozpraszania światła potwierdza zasadnicze założenia hydrodynamiki ciekłych kryształów. Ruch termiczny molekuł powoduje wystąpienie fluktuacji wektora uporządkowania (dyrektora), te zaś — dzięki istnieniu anizotropii optycznej ciekłych kryształów — dają silne ropraszanie i depolaryzację światła. Fluktuacje wektora orientacji, będącego funkcją położenia $\vec{r}$ i czasu t, możemy zapisać w prosty sposób:

$$\vec{n}(\vec{r}, t) = \vec{n}_0 + \delta\vec{n}(\vec{r}, t).$$

Odchylenie $\delta\vec{n}(\vec{r}, t)$, jest właśnie miarą fluktuacji i ma silny wpływ na natężenie światła rozpraszanego pod określonym kątem. Schemat aparatury do pomiaru natężenia światła rozproszonego użytej w Laboratorium w Orsay jest przedstawiony na rys. 9.

pomiary rozpraszania światła



Rys. 9. Aparatura do pomiaru rozpraszania światła w ciekłych kryształach (Laboratorium Orsay). Źródłem światła jest laser: fotopowielacz rejestruje natężenie światła rozproszonego pod określonym kątem

Pomiary rozpraszania światła pozwoliły znaleźć w niezależny sposób wartości współczynników η_1, η_2 i η_3. Z pomiarów tych wynikają interesujące wnioski na temat struktury cieczy nematycznych, które przede wszystkim pozwalają poddać w wątpliwość słuszność modelu domenowego nematyków.

Nie będziemy zajmować się tutaj szczegółowo zagadnieniem absorpcji fal ultradźwiękowych w ciekłych kryształach. Zwrócimy tylko uwagę, że badania te, przeprowadzone przy częstościach rzędu kilku MHz, dały w wyniku takie same wartości η_1, η_2 i η_3, jakie zostały otrzymane wcześniej metodami w zasadzie statycznymi. Mechanizm przekazywania pędu między molekułami jest zatem identyczny w bardzo szerokim zakresie częstości.

samodyfuzja w nematykach

Samodyfuzja jest szczególnym rodzajem dyfuzji, w której jedne molekuły ulegają przemieszczeniu wśród innych molekuł tej samej substancji. W Krakowie, w zespole kierowanym przez J. Janika i od dawna specjalizującym się w badaniach strukturalnych i w fizyce molekularnej, podjęto intensywne prace nad zagadnieniem ciekłych kryształów. W szczególności zespół zajął się problemem samodyfuzji w nematykach.

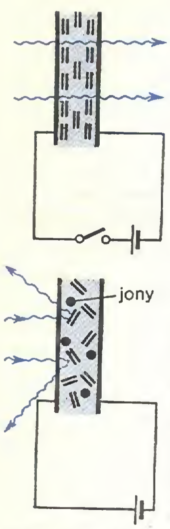
Do badań samodyfuzji zastosowano metodę kwazielastycznego rozpraszania neutronów. Wiązka powolnych neutronów o określonej energii, wysyłanych np. przez reaktor jądrowy, po rozproszeniu na badanej próbce daje tzw. linię, której rozmycie zależy od nieuporządkowanych ruchów molekuł w próbce. Zgodnie z opisem teoretycznym, dyfuzja prostopadła, w której molekuły przemieszczają się prostopadle do kierunku orientacji, opisywana współczynnikiem $D_\perp$, jest powolniejsza od dyfuzji równoległej ze współczynnikiem dyfuzji $D_\parallel$, w której molekuły przemieszczają się równolegle względem siebie. Zespół J. Janika stwierdził doświadczalnie, że istotnie zachodzi zależność $D_\parallel > D_\perp$.

Zastosowanie ciekłych kryształów

Niezwykłe zainteresowanie, jakie budzą obecnie ciekłe kryształy, jest spowodowane ich przydatnością w wielu gałęziach nauki i techniki. Prace oryginalne i artykuły przeglądowe na temat własności i zastosowań ciekłych kryształów ukazują się w książkach i czasopismach z dziedziny fizyki, elektroniki, chemii, biologii, farmacji, medycyny, techniki materiałowej itd.

Jak dotychczas, szerokie zastosowania znalazły prawie wyłącznie nematyki i cholesterolowe ciekłe kryształy, chociaż ostatnio prowadzi się badania również nad zastosowaniem smektyków.

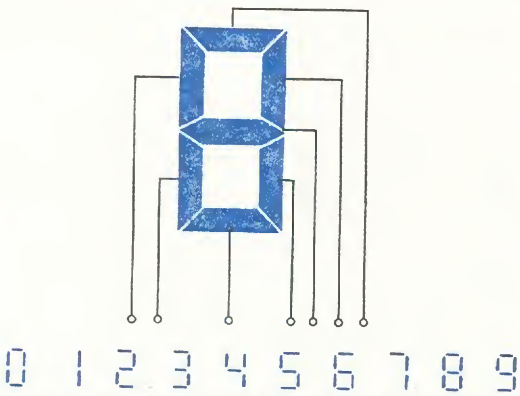
Przede wszystkim wykorzystuje się egzotyczne połączenie w ciekłych kryształach anizotropii optycznej i cech ciekłości, dzięki którym dają się one łatwo orientować lub deformować pod wpływem najróżnorodniejszych oddziaływań zewnętrznych.



Rys. 10. Dynamiczne rozpraszanie światła w nematykach. Analogiczny efekt występuje także przy orientacji molekuł prostopadle do powierzchni elektrod

odczytniki

W odczytnikach stosowanych w zegarkach elektronicznych, minikalkulatorach, ekranach radarowych z pamięcią, elektrooptycznych urządzeniach pamięciowych itd. wykorzystuje się różnorodne zjawiska elektrooptyczne. Początkowo do tych celów wykorzystywano zjawisko dynamicznego rozpraszania światła (rys. 10). Między dwie płytki szklane, pokryte przewodzącymi warstwami tlenku cyny SnO_2 lub tlenku indu In_2O_3, wprowadzano cienką warstwę nematyka o grubości ok. 30 μm. Nematyk był zorientowany homeotropowo lub planarnie i był zupełnie przezroczysty. Kiedy do elektrod przykładano napięcie, to płynący prąd jonowy burzył pierwotne uporządkowanie i warstwa nematyka stawała się mleczno biała. Gdy warstwa tlenkowa na jednej z elektrod była wytrawiona w kształcie siedmiosegmentowego pola numerycznego, jak na rys. 11, to po przyłączeniu napięcia do odpowiednich segmentów otrzymywano zobrazowanie żądanej cyfry w postaci mlecznej figury na przezroczystym tle. Jednakże w czasie działania odczytnika tego typu przebiegają szkodliwe procesy elektrochemiczne, powodujące stopniowy rozkład chemiczny ciekłego kryształu, toteż obecnie technikę dynamicznego rozpraszania prawie całkowicie zarzucono.



Rys. 11. Siedmiosegmentowe pole cyfrowe, stosowane w odczytnikach ciekłokrystalicznych

W późniejszych rozwiązaniach odczytników materiał i kształt elektrod pozostały w zasadzie bez zmian, zmienił się jednak sposób oddziaływania na ciekły kryształ. Obecnie prawie wyłącznie stosuje się elektrostatyczne sterowanie warstwą ciekłego kryształu. Przewodzącą warstwę tlenkową oddziela się od ciekłego kryształu bardzo cienką powłoką napylonego dielektryka, co eliminuje przepływ prądu jonowego, ale nie przeszkadza we wnikaniu pola elektrycznego do wnętrza warstwy ciekłego kryształu. W tak uformowanych komórkach wykorzystuje się w zasadzie cztery efekty elektrooptyczne lub ich modyfikacje:

elektrostatyczne sterowanie

1) Komórkę, w której nematyk o ujemnej anizotropii dielektrycznej jest zorientowany homeotropowo przez oddziaływania powierzchniowe, umieszcza się między skrzyżowanymi polaroidami. Przed przyłożeniem napięcia pole widzenia jest ciemne. Włączenie napięcia do odpowiednich segmentów powoduje ich rozjaśnienie (nematyk o ujemnej anizotropii dielektrycznej orientuje się prostopadle do linii sił pola elektrycznego).

2) Między skrzyżowanymi polaroidami umieszcza się komórkę zawierającą warstwę planarnie zorientowanego nematyka o dodatniej, tym razem, anizotropii dielektrycznej. Ponieważ molekuły takiego nematyka orientują się równolegle do kierunku pola elektrycznego, to jednorodne i jasne początkowo pole widzenia ulega zaciemnieniu po włączeniu napięcia.

3) Jeżeli w komórce umieszczonej jak w punkcie 2) wytworzymy teksturę skręconego nematyka, która ma własność skręcania o 90° płaszczyzny drgań światła spolaryzowanego, to pole widzenia między skrzyżowanymi polaroidami będzie początkowo jasne. Po włączeniu napięcia molekuły nematyka orientują się równolegle do kierunku pola, tekstura śrubowa skręconego nematyka zostaje zniszczona i zamieniona w teksturę homeotropową. Znika zdolność skręcania płaszczyzny drgań i pole widzenia ulega zaciemnieniu. Opisany efekt jest obecnie najczęściej wykorzystywany w konstrukcjach odczytników ciekłokrystalicznych.

4) Pod działaniem pola elektrycznego zachodzi indukowana przemiana fazowa cholesterolowego ciekłego kryształu w nematyk. Jeżeli warstwa cholesterolowa o dodatniej anizotropii dielektrycznej i teksturze konfokalnej zostanie poddana działaniu pola elektrycznego, to molekuły, dążąc do ustawienia się równolegle względem kierunku pola, wytworzą teksturę homeotropową, a tym samym cholesterolowy ciekły kryształ ulegnie przemianie w nematyk. Mleczna białość warstwy, wywołana istnieniem tekstury konfokalnej, zniknie i warstwa będzie zupełnie przezroczysta tak długo, jak długo będzie działało pole sterujące.

Odczytniki ciekłokrystaliczne, których działanie oparte jest na opisanych efektach elektrooptycznych, nadają się szczególnie dobrze jako układy pośredniczące między mikrominiaturowymi urządzeniami elektronicznymi a człowiekiem. Ich sterowanie odbywa się na drodze elektrostatycznej przy napięciach rzędu kilku woltów i praktycznie bezprądowo, mogą więc współpracować z układami logicznymi bez użycia wzmacniaczy.

zdolność orientowania innych molekuł

Cenną, z punktu widzenia zastosowań, własnością ciekłych kryształów jest ich zdolność orientowania molekuł innych substancji. Jeżeli substancja rozpuszczona w ciekłym krysztale ma wydłużone molekuły, to zostaną one zorientowane zgodnie z kierunkiem molekuł ciekłego kryształu. Ten efekt jest wykorzystywany do orientacji długich molekuł w czasie ich badania metodami spektroskopii EPR i NMR (→ Spektroskopia rezonansów magnetycznych), do tworzenia polimerów o zorientowanym usieciowaniu oraz w chromatografii gazowej do rozdzielania stereoizomerów wielu substancji. Uporządkowane roztwory substancji o długich molekułach można przeprowadzić w stan stały bez naruszenia uporządkowania ciekłokrystalicznego. Dokonuje się tego przez gwałtowne oziębienie roztworu do temperatury ciekłego azotu, w której następuje zamrożenie stanu orientacji.

ciekłe ferromagnetyki

Bardzo ciekawe zastosowanie znalazł efekt wzajemnego oddziaływania orientującego między ciekłym kryształem i bardzo drobnymi ziarnami ferromagnetyka. Koloidalne mieszaniny ciekłego kryształu i rozpylonego magnetytu mają własności ciekłego ferromagnetyka. Ciecze takie dają się namagnesować, są przyciągane przez magnes itd.

Łatwość zmiany orientacji molekularnej w ciekłych kryształach została wykorzystana do badań stanu powierzchni ciał stałych. Niektóre ciekłe kryształy mają zdolność spontanicznego tworzenia tekstur homeo-

badania powierzchni

tropowych, czyli ich molekuły samorzutnie przyjmują położenie prostopadłe do powierzchni ciał stałych. Jeżeli taka powierzchnia ma obszary o innym ładunku elektrycznym czy o innym składzie chemicznym, to tekstura homeotropowa zostanie w tych obszarach zaburzona. Zaburzenie to łatwo wykrywać przez obserwację warstwy ciekłego kryształu w świetle spolaryzowanym.

efekt Kerra

Spośród efektów optycznych w nematykach należy jeszcze wymienić efekt Kerra, czyli pojawienie się dwójłomności optycznej w cieczach pod działaniem pola elektrycznego o dużym natężeniu. Substancje nematyczne w temperaturze, w której znajdują się w fazie izotropowej, mają bardzo dużą stałą Kerra. Sterowanie elementami elektrooptycznymi, działającymi na zasadzie efektu Kerra jest jednak utrudnione przez konieczność stosowania wysokich napięć.

Długość skoku linii śrubowej w strukturze cholesterolowej, decydująca o barwie światła selektywnie odbitego, zależy w istotny sposób od temperatury, składu chemicznego ciekłego kryształu, od działania pól elektrycznych i magnetycznych, a także od oddziaływań mechanicznych. Stwarza to pole do wielu zastosowań, przede wszystkim w termografii.

termografia

Można komponować mieszaniny cholesterolowych ciekłych kryształów tak, aby barwa selektywnego odbicia pojawiała się w ściśle określonej temperaturze, a ponadto można regulować zakres temperatur, w jakim barwa powinna zmienić się od czerwonej do fioletowej. Można np. uzyskać taką mieszaninę, aby barwa czerwona przypadała w temperaturze 34°C, zaś barwa fioletowa w 36°C. Jeżeli taką mieszaniną pokryjemy fragment skóry człowieka, np. rękę, to otrzymamy barwną mapę rozkładu temperatury na skórze, jak to pokazano na il. 24 (tabl. 8). Można wytworzyć mieszaniny, w których przedział między pojawieniem się barwy czerwonej i fioletowej wynosi zaledwie 0,5°C. Warstwy takich czułych termicznie mieszanin stosuje się do budowy noktowizorów, do otrzymywania hologramów w zakresie podczerwieni i mikrofal, do wykrywania defektów materiałów, do badań modeli lotniczych itd.

Barwa selektywnego odbicia zależy także od obecności w powietrzu par wielu substancji organicznych. Fakt ten wykorzystuje się do wykrywania zanieczyszczeń toksycznych powietrza, do analizy zapachów (sztuczny nos) itp.

Wymienione zastosowania stanowią tylko przykłady najpopularniejsze. Z nowszych zastosowań możemy tutaj jedynie zasygnalizować wykorzystywanie ciekłych kryształów do budowy laserów cieczowych, do konstrukcji światłowodów i filtrów optycznych, do wykonywania siatek dyfrakcyjnych i do badań topografii powierzchni ciał stałych metodą interferencji światła w cienkich warstwach cholesterolowych ciekłych kryształów.

Ciekłe kryształy w strukturach biologicznych

Zjawisko życia jest niezaprzeczalnie najbardziej skomplikowanym zespołem procesów fizycznych i chemicznych przebiegających w ośrodkach ciekłych. Wielu z tych elementarnych procesów fizykochemicznych jeszcze nie znamy, podobnie jak często nie znany, czy nawet obecnie nie przeczuwany, jest charakter łańcucha związków przyczynowo-skutkowych między tymi procesami.

Znaczna część substancji, z których zbudowane są organizmy żywe, ma strukturę ciekłokrystaliczną i dlatego bez ciekłych kryształów życie w ogóle nie byłoby możliwe. W żywych komórkach ciekłe kryształy występują przede wszystkim jako elementy strukturalne błon biologicznych, ale także jako roztwory wydłużonych makromolekuł i wirusów.

Ciekłe kryształy są cieczami, w których molekuły są w znacznym stopniu względem siebie uporządkowane, i dlatego obecność struktur ciekłokrystalicznych w błonach biologicznych narzuca błonom celową organizację, dostosowaną do ich funkcji w żywej komórce. Jednocześnie ciekłość takiego układu uporządkowanego zapewnia błonom dużą elastyczność i podatność na oddziaływania zewnętrzne, a także — łatwy transport odpowiednich substancji oraz wymianę zużytych lub uszkodzonych fragmentów.

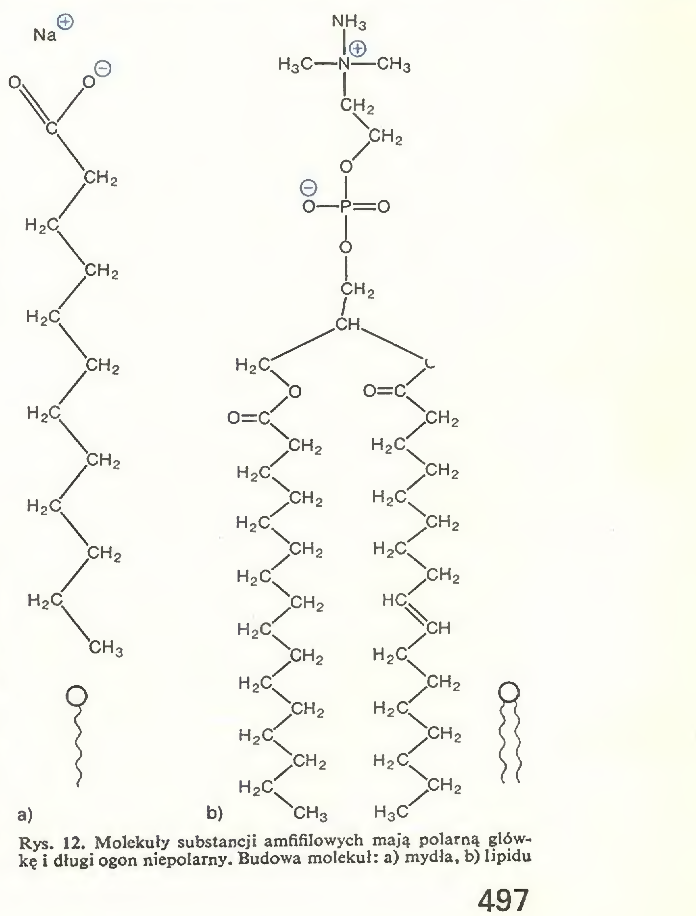
To, że błony biologiczne istotnie zawierają struktury ciekłokrystaliczne, zostało stwierdzone metodami dyfrakcji promieniowania rentgenowskiego, elektronowego rezonansu paramagnetycznego oraz kalorymetrii różnicowej, a także na podstawie badania przemian fazowych wywołanych termicznie w lipidach i w błonach.

struktura błon biologicznych

Bazę strukturalną każdej błony biologicznej, stanowi podwójna warstwa molekuł lipidów, uporządkowanych względem siebie tak jak molekuły w ciekłym krysztale smektycznym. Podwójne warstwy lipidowe powstają spontanicznie dzięki szczególnej budowie molekuł tych substancji, określającej ich zachowanie się w kontakcie z rozpuszczalnikiem (→ Błony komórkowe).

Wzajemne rozpuszczanie się substancji określa zasada „swój do swego", tzn. że substancje polarne są dobrze rozpuszczalne tylko w rozpuszczalnikach polarnych i odwrotnie. Tak np. alkohol metylowy CH_3OH lub kwas octowy CH_3COOH, dzięki polarnym grupom —OH i —COOH, rozpuszczają się dobrze w wodzie — rozpuszczalniku silnie polarnym. Parafina, złożona głównie z niepolarnych węglowodorów, nie rozpuszcza się w wodzie, jest natomiast dobrze rozpuszczalna w niepolarnym benzenie. Takie substancje jak alkohol metylowy i kwas octowy nazywamy substancjami hydrofilowymi (lubiącymi wodę), zaś węglowodory — substancjami hydrofobowymi (obawiającymi się wody). Gdy jednak w molekułach

substancje hydrofilowe i hydrofobowe



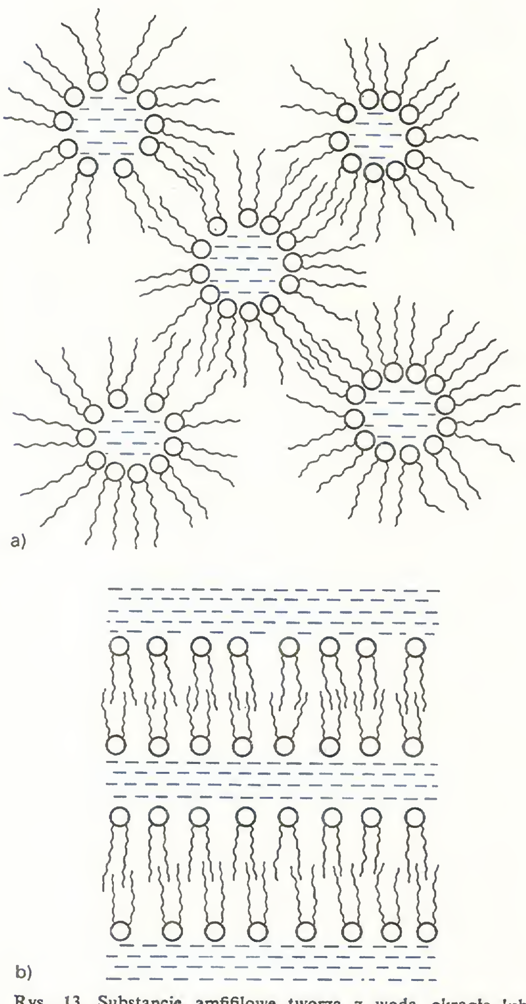
Rys. 12. Molekuły substancji amfifilowych mają polarną główkę i długi ogon niepolarny. Budowa molekuł: a) mydła, b) lipidu

substancje amfifilowe

kwasów i alkoholi wzrasta długość łańcucha węglowodorowego, to hydrofobowe działanie tego łańcucha równoważy hydrofilowe działanie grupy polarnej i takie substancje — mające obojnacze cechy rozpuszczalności — nazywamy substancjami amfifilowymi. Typowymi substancjami amfifilowymi, poza wymienionymi już długołańcuchowymi kwasami i alkoholami, są również mydła (sole kwasów tłuszczowych, rys. 12a) oraz właśnie lipidy (rys. 12b). Wspólną cechą molekuł tych związków jest posiadanie przez nie polarnej główki i jednego lub dwu długich ogonów niepolarnych.

ciekły kryształ liotropowy

Po dodaniu do substancji amfifilowej niewielkiej ilości wody powstaje uporządkowana struktura ciekłokrystaliczna. Woda zostaje podzielona na bardzo cienkie warstewki lub nici, otoczone ze wszystkich stron przez grupy polarne molekuł amfifilowych. Otrzymany ciekły kryształ, zbudowany z kulistych lub wydłużonych miceli, zawierających rdzeń wodny (rys. 13a), lub z cienkich płytek, zwanych lamelami (rys. 13b),



Rys. 13. Substancje amfifilowe tworzą z wodą okrągłe lub wydłużone micele (a) oraz płytkowe lamele (b)

nazywamy ciekłym kryształem liotropowym, czyli powstającym pod wpływem działania rozpuszczalnika. Pojedyncze liotropowe lamele, takie jak na rys. 13b, stanowią bazę strukturalną błon biologicznych.

połączenia z białkami

Ciekłokrystaliczny, dwuwarstwowy szkielet błon biologicznych, w którym molekuły lipidów są zwrócone ku sobie łańcuchami węglowodorowymi, tworzy skomplikowane połączenia z białkami. Substancje białkowe o różnej konformacji pokrywają błonę z obu stron i w wielu miejscach ją przenikają. Białka umieszczone w błonie spełniają wiele różnorodnych funkcji biologicznych, a poza tym dzięki ich obecności błona staje się sztywniejsza, struktura zaś ciekłokrystaliczna lipidów nie ulega zniszczeniu pod wpływem nadmiaru wody. Przypuszcza się (model Brockerhoffa), że działanie białek polegające na utrwalaniu struktury ciekłokrystalicznej błon jest rezultatem wiązań wodorowych, jakie powstają między grupami karbonylowymi fosfolipidów i sfingolipidów jako akceptorami a białkami, cholesterolem i wodą — jako donorami wodoru. Mimo usztywniającego działania białek, błona ma ciągle naturę cieczową. Świadczy o tym duża ruchliwość molekuł białek w błonie, a także ruchliwość antygenów i molekuł lipidów w płaszczyźnie błony.

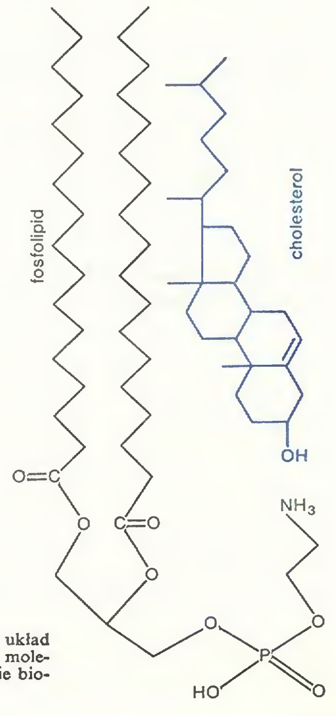
termotropowe fazy mezomorficzne

Badania spektroskopowe, rentgenograficzne i termograficzne dowiodły, że lipidy mogą tworzyć nie tylko liotropowe, lecz także termotropowe fazy mezomorficzne, tzn., że przechodzą w stan cieczy anizotropowej także bez udziału wody i jedynie pod wpływem ogrzewania. Termotropowe przejścia tych substancji do fazy mezomorficznej występują jednak w wysokich temperaturach i wszystkie ciekłe kryształy, spotykane w organizmach żywych, mają bez wyjątku charakter liotropowy, wynikający z oddziaływań lipidów z wodą. Usunięcie wody z błon powoduje, że lipidy przechodzą w stan krystaliczny, co oznacza przerwanie procesów życiowych w komórkach. Lipidy, kwasy nukleinowe i proteiny są aktywne biologicznie tylko w obecności wody i odwodnienie powoduje śmierć komórki, mimo że jej skład chemiczny pozostanie nie zmieniony.

Ciekłe kryształy liotropowe, zawarte w strukturach biologicznych, ulegają również przemianie w zwykłą ciecz po ich odpowiednim podgrzaniu. Interesujące jest, że temperatura przejścia w stan cieczy izotropowej wielu liotropowych ciekłych kryształów lipidów i wody wynosi 41°C, a powyżej tej temperatury, już niebezpiecznej dla człowieka, ustają mechanizmy komórkowe u wielu pierwotniaków i organizmów wielokomórkowych. Widać więc, że procesy życia ulegają zahamowaniu nie tylko wtedy, gdy lipidy przechodzą w stan krystaliczny, lecz także wtedy, gdy znajdą się one w fazie cieczy izotropowej.

cholesterol

Inną, szczególnie ważną dla struktur błonowych substancją jest cholesterol. Prawdopodobnie molekuły cholesterolu przyjmują położenie równoległe względem molekuł lipidów, jak to przedstawiono na



Rys. 14. Hipotetyczny układ molekuły cholesterolu i molekuły fosfolipidu w błonie biologicznej

rysunku 14, i przez osłabienie oddziaływań między łańcuchami węglowodorowymi utrzymują ciekłokrystaliczny charakter membran w dużym przedziale temperatur.

Relacje między ciekłokrystalicznym szkieletem błon biologicznych i lekami stanowią ważny problem farmakologiczny. Na przykład dobra rozpuszczalność w lipidach jest, jak stwierdzono, warunkiem aktywności leków działających na układ nerwowy. Środki usypiające i znieczulające, leki psychotropowe i narkotyki działają na układ nerwowy poprzez ingerencję w strukturę i funkcje jego cienkich błon ciekłokrystalicznych. Jedne z tych substancji mogą unieruchamiać kationy K^+, Na^+ i Ca^{++}, przez co znoszą lub znacznie upośledzają przenoszenie sygnałów elektrycznych w nerwach, inne zaś, jak np. znany narkotyk LSD, wywołują lokalne zmiany w uporządkowaniu lipidów w błonach, prowadzące do niekontrolowanych, halucynogennych reakcji systemu nerwowego.

leki działające na układ nerwowy

Obecnie przypuszcza się, że również powstanie komórek tkanek nowotworowych związane jest ze zmianami w strukturze błon, ponieważ w komórkach tych zanika zdolność reagowania na sygnały przychodzące od komórek otaczających. Komórki rakowate „tępieją", w dużym stopniu zatracają specjalizację i rozmnażają się w sposób nie kontrolowany. Zmiany te mogą być spowodowane przez czynniki chemiczne, mechaniczne, przez działanie wirusów itd.

biofizyczne modele ciekłokrystaliczne

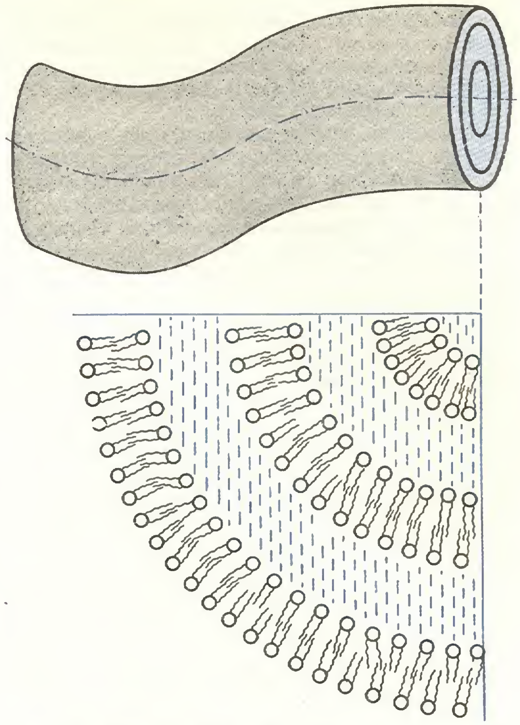
Badania błon lipidowych i uporządkowanych układów makromolekuł dostarczają więcej informacji o istocie życia niż jakikolwiek inny dział nauki. Dlatego duże znaczenie przypisuje się badaniom modeli błon biologicznych wytworzonych sztucznie w postaci podwójnych warstw lipidowych. Warstwy takie umieszcza się jako przegrody między dwoma roztworami elektrolitów, zaś mechanizmy przewodzenia, selektywność jonową oraz działanie określonych domieszek bada się przez pomiar natężenia prądu oraz asymetrii przewodnictwa w funkcji napięcia i w zależności od składu obu roztworów.

Nowe możliwości stwarza modelowanie błon przy użyciu warstw ciekłych kryształów o strukturze cholesterolowej. Z powodu ułożenia molekuł wzdłuż linii śrubowej, grubość tych warstw kilkaset razy przewyższa grubość warstw smektycznych. Dzięki temu warstwy cholesterolowe mogą być łatwo obserwowane przy użyciu mikroskopu optycznego. Istnieje więc możliwość modelowania kształtu błon w zadanych warunkach przestrzennych oraz ich rozmieszczenia wewnątrz komórek o określonej specjalizacji, a także istnieje możliwość badania zarówno biogenezy, jak i morfogenezy błon.

Upakowanie pofałdowanych i przenikających się warstw cholesterolowych jest podobne do rozmieszczenia błon wielu organów specjalnych, np. błon receptorów w oku, rozwiniętych błon mitochondriów czy błon w chloroplastach u roślin (il. 109, 110, tabl. 27). Uwarstwienie kropelek cholesterolowych ciekłych kryształów (il. 111, tabl. 27) jest często analogiczne do obserwowanego za pomocą mikroskopu elektronowego uwarstwienia roztworów biologicznych.

Niezwykle ważną ze względu na przewodzenie impulsów nerwowych strukturą membranową jest otoczka mielinowa aksonów, zbudowana z cylindrycznych, współosiowych warstw lipidowych, tzw. liposomów (rys. 15). Zdolność tworzenia figur mielinowych mają lipidy w połączeniu z wodą i w obecności substancji

figury mielinowe



Rys. 15. Schemat budowy tworu mielinowego wytworzonego w mieszaninie lipidów i wody

białkowych. Kształt, rozgałęzienia i strukturę wewnętrzną tworów mielinowych można łatwo modelować także w ciekłych kryształach nematycznych, po dodaniu do nich estrów cholesterolu lub niektórych substancji aktywnych optycznie (il. 112, tabl. 27).

Badania cienkich warstw ciekłych kryształów, a zwłaszcza związków między ich morfologią a procesami transportu masy, energii i ładunku elektrycznego w powiązaniu z towarzyszącymi im procesami chemicznymi, pozwolą znacznie przybliżyć rozwiązanie tajemnicy materii ożywionej. Badania biofizycznych modeli ciekłokrystalicznych mogą się okazać czynnikiem integrującym prace w zakresie biochemii, biofizyki, fizjologii itd. i mogą dostarczyć odpowiedzi na trudne pytania z pogranicza tych dyscyplin. Prawdziwą trudność stanowi jednak formułowanie poprawnych pytań.

A. Adamczyk, Z. Strugalski *Ciekłe kryształy*, Warszawa 1976; L. M. Blinov *Zjawiska elektrooptyczne w ciekłych kryształach*, Post. Fiz., **28**, 237 (1977); I. G. Czistiakow *Żydkije kristałły*, Moskwa 1970; P. G. de Gennes *The Physics of Liquid Crystals*, Oxford 1974; S. Friberg, R. F. Gould (ed.) *Lyotropic Liquid Crystals and the Structure of Biomembranes*, Washington 1956; M. Mięsowicz 50 *lat polskich badań nad ciekłymi kryształami*, Post. Fiz. **26**, 129 (1975).

Współczesne teorie symetrii w krystalografii

Zygmunt Trzaska Durski

klasyczna teoria symetrii kryształów

Rozwój klasycznej teorii symetrii kryształów został ukoronowany opracowaniem przez J. S. Fiodorowa, A. M. Schoenfliesa i W. Barlowa, w latach osiemdziesiątych ubiegłego stulecia, teorii 230 grup przestrzennych. Wydawało się wówczas, że prace nad teorią symetrii kryształów zostały tym samym całkowicie zakończone, że nic więcej w tej dziedzinie nie będzie można zrobić. Jednak już wkrótce w pracach samego

Fiodorowa pojawiły się elementy uogólniające teorię symetrii. Fiodorow wprowadził bowiem do krystalografii pojęcie „symetrii pozornej" kryształów. Okazało się więc, że nie wszystko w teorii symetrii jest już wiadome.

W naszym stuleciu rozwój teorii symetrii przebiegał w dwu kierunkach. Jeden — dotyczył doskonalenia i opracowywania szczegółów klasycznej teorii symetrii, drugi — uogólnień tej teorii. Przyjrzyjmy się bliżej niektórym fragmentom tych prac.

doskonalenie teorii symetrii

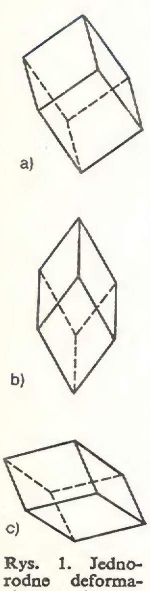
Doskonalenie teorii symetrii obejmowało prace nad symetrią różnorodnych figur skończonych i nieskończonych. I tak np.: 1) stwierdzono istnienie siedmiu grup symetrii bordiur, tzn. szlaków ornamentacyjnych (P. Niggli, G. Polya); 2) wyprowadzono 31 grup symetrii wstęg (A. Speiser); 3) opracowano teorię symetrii rdzeni i stwierdzono istnienie dla nich 75 grup symetrii (C. Hermann); 4) ustalono istnienie 80 grup symetrii warstw (K. Herman, E. Alexander, C. Hermann, L. Weber); 5) stwierdzono możliwość występowania najgęstszych ułożeń kul o jednakowym promieniu tylko w 8 grupach przestrzennych (N. W. Biełow).

Do tego rodzaju prac należy też zaliczyć prace nad krystalograficznymi i strukturalnymi odmianami postaci prostych, nad postaciami krawędziowymi i wierzchołkowymi (→ Budowa kryształów). Stosowane wyżej nazwy: bordiura, wstęga, rdzeń, warstwa — nie mają znaczenia potocznego, lecz są ściśle określonymi terminami naukowymi. Rdzeń definiuje się np. jako figurę z jednym kierunkiem szczególnym, nie mającą szczególnych punktów i płaszczyzn. Szczególnymi nazywają się tu takie punkty, proste i płaszczyzny, które nie mają w danej figurze punktów, prostych i płaszczyzn sobie równoważnych. Punktem szczególnym jest np. środek kwadratu, a szczególną prostą — czterokrotna oś symetrii kwadratu (w kwadracie nie ma bowiem innego środka lub innej osi czterokrotnej).

przekształcenia izometryczne

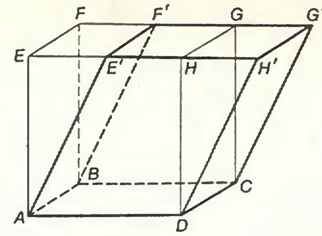
Uogólnień klasycznej teorii symetrii można było poszukiwać przez zmianę lub rozszerzenie podstawowych założeń teorii, a takimi są np. pojęcia: równości figur, figury symetrycznej, przekształcenia symetrycznego. Utworzone dotychczas nowe teorie symetrii są głównie oparte na różnym rozumieniu pojęcia równości figur. Pojęcie to zostało sformułowane w XIX w. przez A. F. Moebiusa. Według Moebiusa dwie figury są równe wtedy, gdy każdemu punktowi jednej figury odpowiada punkt w drugiej figurze, a odległość między dwoma dowolnymi punktami jednej figury jest równa odległości między dwoma takimi samymi punktami drugiej figury. W klasycznej teorii symetrii wszystkie przekształcenia symetryczne figur (→ Budowa kryształów) są przekształceniami izometrycznymi, tzn. nie zmieniającymi odległości między punktami. W wyniku swoistych rewizji pojęcia równości figur i rezygnowania z izometryczności przekształceń wyrosły nowe, ogólniejsze w porównaniu z klasyczną, teorie symetrii.

Pierwsze teorie uogólniające klasyczną teorię symetrii zostały przedstawione w pracach J. S. Fiodorowa o symetrii pozornej (1901) i C. M. Violi o harmonii kryształów (1904). Później, w 1925 r. D. W. Naliwkin opublikował pracę o symetrii krzywoliniowej, w latach 1947–1951 W. I. Michiejew opracował teorię homologii kryształów, a w 1960 r. A. W. Szubnikow sformułował teorię symetrii podobieństwa. Jednak najbardziej doniosłą w skutkach, mającą ogromny wpływ na dalszy rozwój i na dalsze uogólnienie teorii symetrii, okazała się teoria antysymetrii A. W. Szubnikowa (1951).



Rys. 1. Jednorodne deformacje: sześcian (a) przechodzi w romboedr w wyniku rozciągnięcia (b), lub ściśnięcia (c) wzdłuż trójkrotnej osi symetrii (ustawionej pionowo na rys. a)

Symetria pozorna J. S. Fiodorowa wynikła z jego prac nad deformacjami jednorodnymi figur geometrycznych i kryształów. Jednorodnymi, tzn. nie zmieniającymi charakteru figury: figura wyjściowa i pochodna mają np. taką samą liczbę wierzchołków, krawędzi, ścian. Deformacje jednorodne powstają w rezultacie rozciągnięcia (lub ściśnięcia) albo „przesunięcia" figur. W wyniku jednorodnej deformacji figura zmienia swoją symetrię. Tak więc operacje rozciągania (lub ściskania) i przesunięcia nie są operacjami symetrii, lecz jak wykazał później W. I. Michiejew jednorodne deformacje są operacjami homologii. Sześcian poddany rozciągnięciu lub ściśnięciu przekształca się



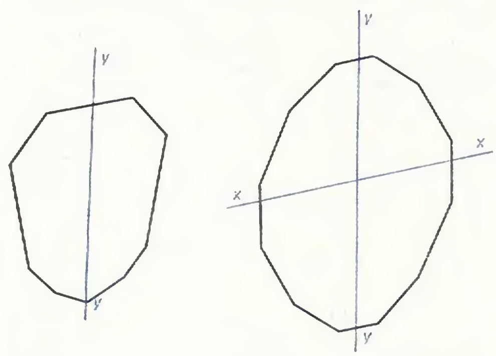
Rys. 2. Jednorodna deformacja: przesunięcie. Sześcian *ABCDEFGH* przechodzi w jednoskośny równoległościan *ABCDE'F'G'H'*

w romboedr (rys. 1), a poddany przesunięciu przemienia się w równoległościan jednoskośny (rys. 2) lub trójskośny Możliwe są także przekształcenia odwrotne (np. romboedr → sześcian). Dowolny wielościan poddany jednorodnej deformacji pozostaje wielościanem. Przy rozciąganiu (lub ściskaniu) objętość wielościanu ulega zmianie, przy przesunięciu objętość pozostaje bez zmiany. W obydwu przypadkach zmienia się postać równoległościanu. Badanie jednorodnych deformacji doprowadziło Fiodorowa do określenia operacji symetrii pozornej. W celu scharakteryzowania symetrii pozornej Fiodorow wprowadził pojęcie elementów symetrii pozornej: osi i płaszczyzny symetrii pozornej (np. odbicie w płaszczyźnie symetrii pozornej może przekształcić trójkąt ukośnokątny w trójkąt prostokątny). Według Fiodorowa wielościanami pozornie symetrycznymi względem siebie są: wielościan rzeczywiście symetryczny (tj. mający klasyczne elementy symetrii) i wielościan dowolny mający taką samą liczbę tak samo ułożonych ścian (jak wielościan pierwszy).

symetria pozorna

harmonia

C.M. Viola opisał i scharakteryzował harmoniczne własności kryształów oraz samą harmonię, z której bezpośrednio wynika symetria. Figura przestrzenna może być harmoniczna względem środka harmonii, płaszczyzny lub osi harmonii (są to elementy harmonii). Na przykład płaszczyzną harmonii danej figury



Rys. 3. Figury harmoniczne względem jednej (*yy*) i dwóch (*xx*, *yy*) płaszczyzn harmonii

jest taka płaszczyzna, która dzieli na połowy odległości między odpowiednimi punktami znajdującymi się na równoległych prostych (rys. 3). Równoległe do siebie promienie rzutujące płaszczyzny harmonii tworzą z tą płaszczyzną pewien kąt. Płaszczyzna symetrii jest więc tym szczególnym rodzajem płaszczyzny harmonii, przy którym promienie rzutujące są prostopadłe do płaszczyzny. Harmonia Violi jest identyczna z symetrią pozorną Fiodorowa.

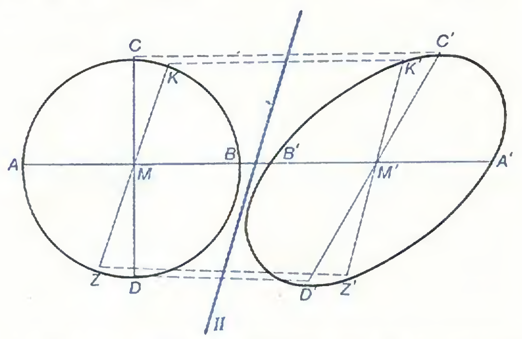
homologia

W. I. Michiejew opierając się na pracach Fiodorowa i Violi opracował teorię homologii kryształów i wykazał, że zarówno symetria pozorna, jak i harmonia są homologią. Homologia jest własnością figur polegającą na jednoznacznej odpowiedniości między wszystkimi ich elementami, przy tym odpowiadające sobie elementy figur są jednorodne, lecz niekoniecznie jednakowe. Dwie homologiczne figury nie są sobie równe,

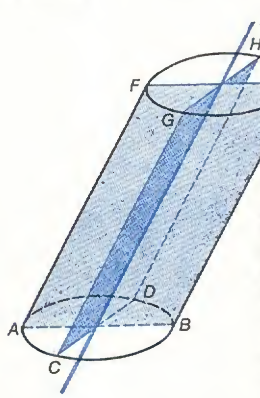
tylko są do siebie podobne w tym sensie, że dowolnej prostej w jednej z nich odpowiada prosta w drugiej, a dowolnej płaszczyźnie w pierwszej figurze odpowiada płaszczyzna w drugiej itd. Homologiczne względem siebie są np. dowolne dwa trójkąty — niezależnie od długości ich boków i kątów między bokami. Homologicznymi względem siebie są także każde dwa dowolne prostopadłościany, jak np. sześcian i słup tetragonalny zamknięty dwuścianem. Przekształceniem homologicznym nazywa się każde przekształcenie przeprowadzające daną figurę w figurę względem niej homologiczną. Pierwszymi przekształceniami homologicznymi zastosowanymi w krystalografii były jednorodne deformacje Fiodorowa.

elementy homologii

W homologii kryształów odpowiednikiem elementów symetrii z klasycznej teorii symetrii są elementy homologii. Do nich należą, prócz zwykłych elementów symetrii, płaszczyzna homologii, osie homologii (o takiej samej krotności jak w symetrii klasycznej) oraz osie inwersyjne homologii. Płaszczyzna homologii Π jest płaszczyzną ukośnego odbicia, tzn. że promienie rzutujące (które dla płaszczyzny symetrii są do niej prostopadłe) tworzą z płaszczyzną Π pewien kąt. Na rys. 4 przedstawiono przekształcenie homologiczne kuli w wyniku działania płaszczyzny Π. Oś ho-



Rys. 4. Kula po odbiciu w płaszczyźnie homologii Π przechodzi w elipsoidę



Rys. 5. Elementy homologii ukośnego walca o eliptycznej podstawie są następujące: oś homologii o nieskończonej krotności A_∞, płaszczyzna homologii Π równoległa do płaszczyzny podstawy, płaszczyzna symetrii $ABEF$, 2-krotna oś symetrii równoległa do płaszczyzny podstawy, środek symetrii oraz nieskończony zbiór płaszczyzn homologii przechodzących (czyli równoległych) przez A_∞

mologii A_n jest prostą, za pomocą której przeprowadza się ukośny kołowy lub eliptyczny obrót. Przy takim obrocie każdy punkt układu porusza się na ogół po elipsie (lub po kole), przy czym elipsa czy koło leżą w płaszczyźnie, która może tworzyć z osią pewien kąt (rys. 5). Elementy symetrii są rodzajem elementów homologii.

klasy homologii i harmonie

Istnieje 215 klas homologii, czyli różnych możliwych w kryształach zespołów elementów homologii. Podobnie jak klasy symetrii łączy się w układy krystalograficzne, tak klasy homologii łączy się w harmonie (jest ich 7, a mianowicie: harmonia trójskośna, jednoskośna, rombowa, trygonalna, tetragonalna, heksagonalna i regularna). W każdej klasie homologii można wyprowadzić postacie proste homologii, w których ściany tworzące jedną postać homologii mogą należeć do różnych krystalograficznych postaci prostych. Mi-

przestrzeń grupy homologii

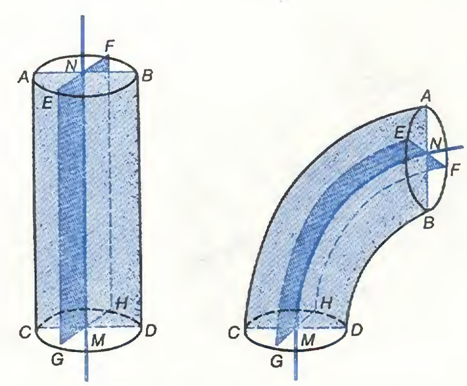
chiejew sformułował również pojęcie przestrzennej grupy homologii i prawidłowego zespołu figur homologicznych (odpowiedników grup przestrzennych i zespołu pozycji symetrycznie równoznacznych). Poddając zespół punktów symetrycznie równoznacznych deformacjom jednorodnym, można otrzymać prawidłowy zespół figur homologicznych, przy czym elementy symetrii grupy przestrzennej przemieniają się w elementy homologii i tworzą grupę przestrzenną homologii.

symetria krzywoliniowa

W celu opisywania symetrii organizmów żywych geolog i paleontolog — D. W. Naliwkin stworzył pojęcie symetrii krzywoliniowej. Według Naliwkina w przyrodzie istnieje wiele obiektów, które łatwo wyginają się i wyprostowują i tym samym stają się raz symetryczne krzywoliniowo, a drugi raz — symetryczne prostoliniowo. Przejście z jednego stanu w drugi jest nadzwyczaj łatwe i właściwie niezauważalne. Zwróćmy uwagę np. na żmiję. Jeśli leży ona nieruchomo wyprostowana, to zarys jej ciała wykazuje symetrię względem płaszczyzny symetrii (pionowej). Wystarczy jednak, aby żmija poruszyła się, wygięła, aby jej symetria prostoliniowa znikła. Czy w wyniku poruszenia się, pełznięcia żmii, symetria jej ciała w ogóle zginęła? Oczywiście — nie, nastąpiła tylko przemiana symetrii prostoliniowej w symetrię krzywoliniową. Płaszczyzna symetrii zmieniła się w zakrzywioną powierzchnię symetrii. Podobnie jest z człowiekiem. Ciało człowieka, np. stojącego, ma pionową płaszczyznę symetrii. Co się stanie, gdy człowiek nieco się wygnie w bok? Czy ciało straci przez to symetrię płaszczyznową? Oczywiście że nie, zajdzie wtedy jedynie przemiana płaszczyzny symetrii w zakrzywioną powierzchnię symetrii (tj. płaszczyznę symetrii krzywoliniowej). Podobnie i zgięty kryształ może stać się symetryczny krzywoliniowo. Postacie symetryczne prostoliniowo są szczególnym rodzajem postaci o symetrii krzywoliniowej. W świecie nieorganicznym przeważają postacie proste, słabo wygięte, w organicznym — silnie wygięte, a nawet skręcone.

elementy symetrii krzywoliniowej

Symetria krzywoliniowa może powstawać w wyniku zgięcia lub skręcenia zwykłych symetrycznych figur. Elementami symetrii krzywoliniowej mogą być: oś symetrii krzywoliniowa, oś symetrii spiralna, powierzchnia symetrii zakrzywiona (tj. płaszczyzna symetrii krzywoliniowej), powierzchnia symetrii spiralna, a także — środek symetrii, osie symetrii, płaszczyzna symetrii. Elementy symetrii są szczególnymi rodzajami elementów symetrii krzywoliniowej. Jeżeli walec (rys. 6) zegnie się tak, aby jego tworząca przemieniła się w łuk koła, to płaszczyzna symetrii $ABCD$ walca pozostanie płaszczyzną symetrii także i dla zgiętego walca, natomiast płaszczyzna $EFGH$ przemieni się w płaszczyznę symetrii krzywoliniowej; oś symetrii MN o nieskończonej krotności (L_∞) zmieni się w oś symetrii krzywoliniowej. Tak więc figury niesymetryczne (w pojęciu symetrii klasycznej) można rozpatrywać jako krzywoliniowo symetryczne.



Rys. 6. Walec po zgięciu staje się figurą o krzywoliniowych elementach symetrii

Symetria krzywoliniowa jest rodzajem homologii, a symetria, symetria pozorna i harmonia są szczególnymi rodzajami symetrii krzywoliniowej. Dla symetrii krzywoliniowej podjęto również próby określenia możliwych kombinacji elementów symetrii krzywoliniowej i wyprowadzenia niektórych klas symetrii krzywoliniowej i płaskich postaci „krzywokrawędziowych" (P.L. Dubow).

symetria podobieństwa

Chyba każdy zna z lat dziecinnych rosyjską drewnianą zabawkę — składaną laleczkę, w której po otwarciu znajduje się następna laleczka — mniejsza, a w tej jeszcze jedna, którą też można otworzyć i w jej wnętrzu znaleźć następną jeszcze mniejszą itd. Takich laleczek, jedna w drugiej jest kilka lub kilkanaście (rys. 7). Wszystkie laleczki są właściwie takie same,



Rys. 7. Symetria podobieństwa: drewniane laleczki „matrioszki"

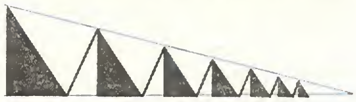
ale różni je ich wielkość. Laleczki są w pewnym sensie jednakowe, wyczuwa się tutaj pewien rytm, pewne powtarzanie się, a jeśli występuje powtarzanie się, to jest i symetria. Tym razem jest to jednak dziwna symetria. Figury są jednakowe, ale różnią się wielkością. Każdym dwu punktom jednej figury odpowiadają dwa punkty drugiej figury, ale odległość między nimi jest inna. Każdej prostej czy płaszczyźnie w jednej figurze odpowiada prosta czy płaszczyzna w drugiej, ale o innej wielkości. H. Weyl, który pierwszy poddał myśl matematycznego opisania takich prawidłowości, nazwał taką „symetrię" podobieństwem, a A.W. Szubnikow, który taki opis stworzył nazwał ją symetrią podobieństwa.

symetria podobieństwa w przyrodzie i sztuce

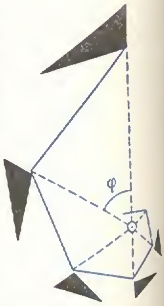
Historia powstania i rozwoju pojęcia symetrii podobieństwa sięga w daleką przeszłość, a tworzyli ją uczeni, malarze i architekci. Już Leonardo da Vinci stosował symetrię podobieństwa w postaci perspektywy. Symetrię podobieństwa obserwuje się często w przyrodzie, np. w ułożeniu liści na młodych pędach roślin, w prawidłowym ułożeniu ziarenek w kwiatach słonecznika czy rumianku, w spiralnych postaciach muszli amonitów, w stożkowych kształtach wielu drzew (np. jodła), w szkieletowych postaciach kryształów i w tzw. piramidach wzrostu kryształów (il. 63, 64, 66, tabl. 17).

W nauce o symetrii podobieństwa za jednakowe — jak już wspomniano — uważa się nie tylko rzeczywiście jednakowe figury, ale i wszystkie do nich podobne, tj. figury o tym samym kształcie. Analogicznie do operacji symetrii można wprowadzić operacje symetrii podobieństwa, które są swoistymi analogami translacji, odbić w płaszczyznach zwykłych i poślizgu, oraz

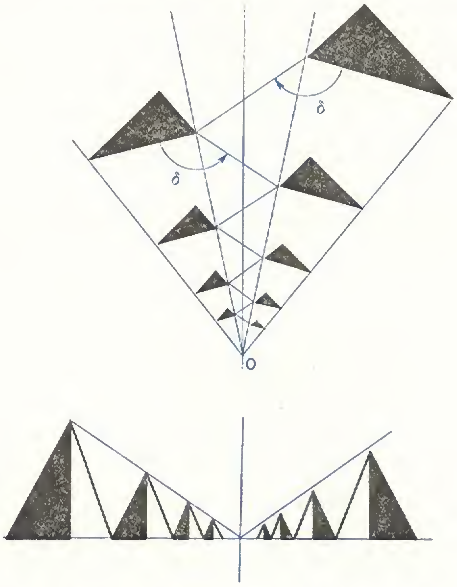
Rys. 8. Symetria podobieństwa: przykład operacji K



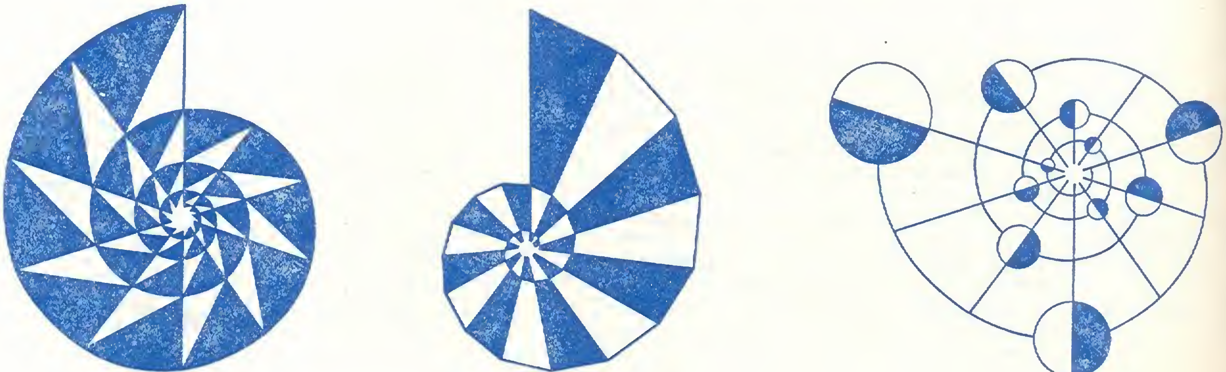
obrotów wokół osi zwykłych i śrubowych, z tą tylko różnicą, że wszystkie te operacje związane są z równoczesnym zwiększeniem (lub zmniejszeniem) skali figur i odległości między figurami. Operacja K jest najprostszą operacją symetrii podobieństwa; polega ona na przenoszeniu wszystkich podobnych części figury w położenia równoległe, przy jednoczesnym zwiększaniu lub zmniejszaniu skali tych części i odległości między nimi n razy (rys. 8). Operacja L składa się z kolejno przeprowadzanych obrotów podobnych części figury wokół nieruchomej osi o pewien stały kąt φ, którym towarzyszy odpowiednie przesuwanie figury w kierunku do osi (lub od osi, jeśli operacja jest prowadzona z wnętrza figury; rys. 9). Odpowiadające sobie punkty podobnych części figury powinny leżeć na spirali logarytmicznej. Operację L można nazwać spiralnym ruchem wokół osi podobieństwa. Łatwo można zauważyć, że operacja K jest równoważna operacji



Rys. 9. Konstrukcja figury mającej symetrię podobieństwa L



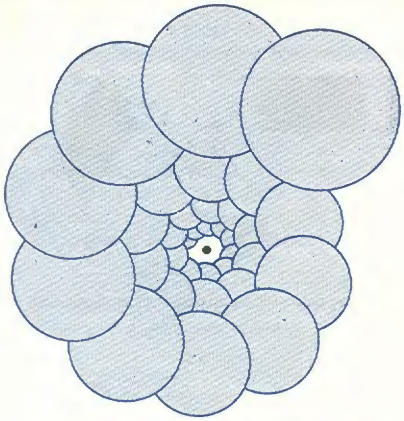
Rys. 10. Przykłady figur mających symetrię podobieństwa M



$\varphi = -\pi/5$ $\varphi = 2\pi/15$ $\varphi = -3\pi/5$

Rys. 11. Przykłady figur mających symetrię podobieństwa L

L przeprowadzanej przy $\varphi = 2\pi$ (tj. $K = L(\varphi = 2\pi)$). Na rys. 11 pokazano przykłady figur skonstruowanych przy zastosowaniu operacji L. Operacja M jest analogiem operacji odbicia zwykłego lub poślizgowego w symetrii klasycznej; nazywa się ją odbiciem w płaszczyźnie podobieństwa (rys. 10). Operacje K, L i M są operacjami symetrii podobieństwa figur dwu-



Rys. 12. Trójwymiarowa figura powstała w wyniku operacji N symetrii podobieństwa

wymiarowych (płaskich). Operacja N jest operacją symetrii podobieństwa figur przestrzennych (rys. 12) i można ją interpretować jako iloczyn operacji $L(\varphi \neq 2\pi)$ i operacji $K = L(\varphi = 2\pi)$, pod warunkiem, że obydwie osie są wzajemnie równoległe. W rezultacie operacja N powoduje ruch podobnych części figury po rozszerzającej się linii śrubowej.

Operacjami symetrii podobieństwa są oczywiście także operacje klasycznej symetrii, stanowiące szczególny rodzaj ogólniejszych operacji symetrii podobieństwa.

Operacje symetrii podobieństwa można składać ze sobą, tworzyć ich kombinacje. Powstają przy tym grupy symetrii podobieństwa. Kilka przykładów figur należących do różnych grup pokazano na rys. 13.

antysymetria

Jak już wspomniano poprzednio, A. W. Szubnikow, opierając się na wcześniejszych pracach A. Speisera (1927) i L. Webera (1929), sformułował jeszcze jedną nową teorię symetrii, mianowicie teorię antysymetrii. W przyrodzie często spotyka się przedmioty podobne do siebie a właściwie równe sobie w pewien szczególny sposób, mianowicie: jak pozytyw i negatyw fotograficzny jednego i tego samego przedmiotu, jak medal i odcisk medalu, jak pozyton i elektron, jak +1 i −1, jak śruba i gwint, jak postacie wzrostu i postacie rozpuszczania kryształów, jak prawa biała rękawiczka z czarnym mankietem i czarna prawa rękawiczka z białym mankietem. Szubnikow ten rodzaj równości figur nazwał ich antyrównością. Ogólnie biorąc, za antyrówne można uważać figury o jednakowym kształcie i jednakowych rozmiarach, lecz przeciwstawne sobie pod względem jakiejś własności: np. barwy, znaku, ładunku elektrycznego, kierunku momentu magnetycznego.

antyrówność figur

Stosunki symetryczne, jakie mogą zachodzić między antyrównymi przedmiotami, figurami, Szubnikow nazwał antysymetrią.

przekształcenie antysymetryczne

Przekształcenie antysymetryczne można zdefiniować jako przekształcenie symetryczne przemieniające figurę białą w czarną lub zmieniające znak figury (+ na −), czyli mówiąc ogólnie — zmieniające własność figury na przeciwną. Szubnikow sformułował więc pojęcie antysymetrii jako zasadnicze rozszerzenie symetrii klasycznej w wyniku dodania zmiany własności fizycznej.

W teorii antysymetrii każdemu punktowi przypisuje się znak + lub − w dowolnym znaczeniu, interpretowanym zwykle jako własności fizyczne: znak ładunku elektrycznego, czarny lub biały kolor itp. Przekształceniem antysymetrycznym nazywa się więc izometryczne przekształcenie figury w siebie samą przy równoczesnej zmianie znaku każdego punktu. W teorii antysymetrii rozpatruje się figury charakteryzujące się trzema własnościami: +1, −1, 0 (np. figura biała, czarna i szara). Figury szare nazywa się także figurami neutralnymi.

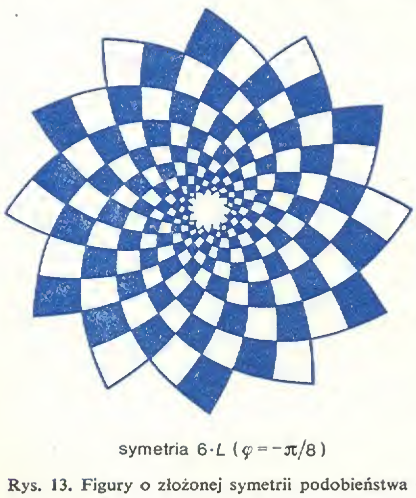
W klasycznej teorii symetrii operacje symetrii przeprowadzane w przestrzeni trójwymiarowej na trójwymiarowych figurach skończonych można formalnie rozpatrywać jako operacje w przestrzeni czterowymiarowej. Operacje takie pozostawiają nie zmienioną czwartą współrzędną (x_4) równą zeru dla wszystkich punktów figury. W tym ujęciu operacje antysymetrii można rozpatrywać również jako operacje w przestrzeni czterowymiarowej, które nie zmieniając bezwzględnych wartości współrzędnych punktów figury względem osi X_4, zmieniają ich znak.

elementy antysymetrii

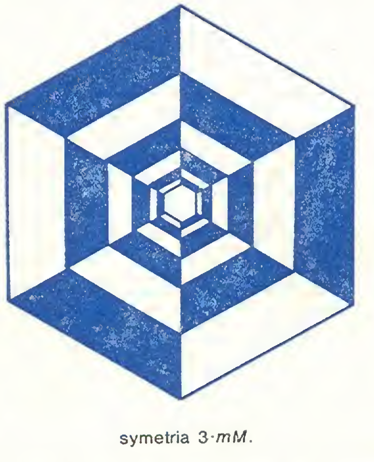
Analogicznie do elementów symetrii istnieją elementy antysymetrii — środek, osie i płaszczyzna antysymetrii, antytranslacja, śrubowe osie antysymetrii i antypłaszczyzny poślizgu. Przekształcenie antysymetryczne działa zawsze w połączeniu ze zwykłymi elementami symetrii. Działania różnych elementów antysymetrii i porównanie ich z przekształceniami symetrycznymi przedstawiono na rys. 14 i 15. Rysunek 16 pokazuje działanie elementów antysymetrii (niektórych) na asymetryczny czworościan.

punktowe grupy antysymetrii

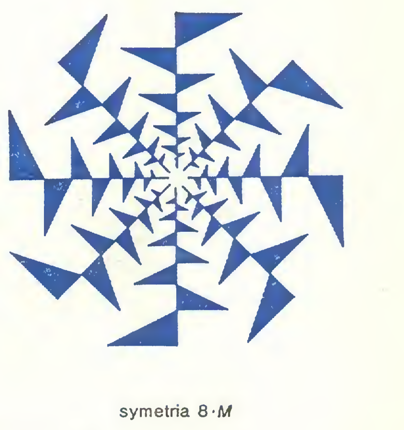
Szubnikow wykazał, że istnieje tylko 122 krystalograficznych punktowych grup antysymetrii (tzn. klas antysymetrii; klasy antysymetrii nazywa się także klasami symetrii magnetycznej), a w tym: 32 grupy polarne (jednobarwne, np. tylko białe), odpowiadające figurom, których punkty symetrycznie równoznaczne



symetria $6 \cdot L$ ($\varphi = -\pi/8$)



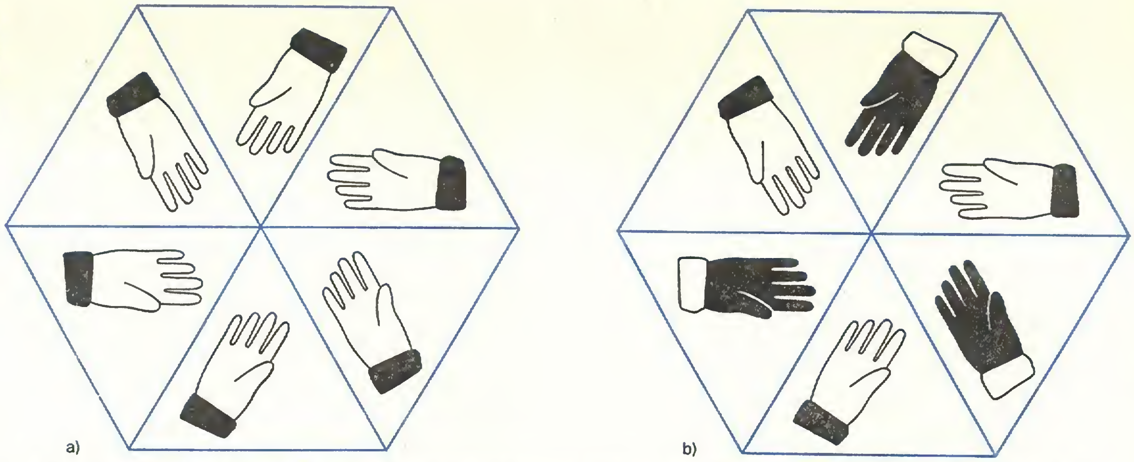
symetria $3 \cdot mM$.



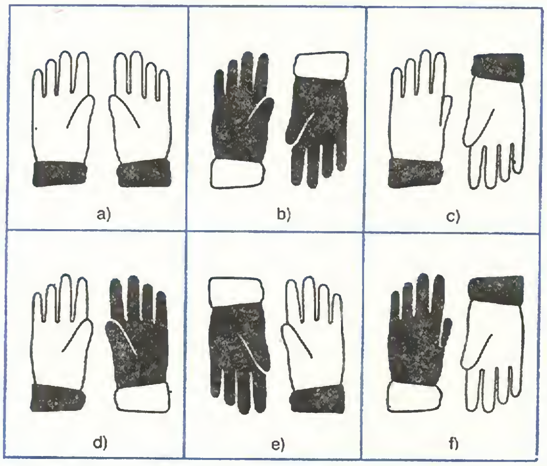
symetria $8 \cdot M$

Rys. 13. Figury o złożonej symetrii podobieństwa (liczba przed symbolem operacji oznacza krotność osi symetrii)

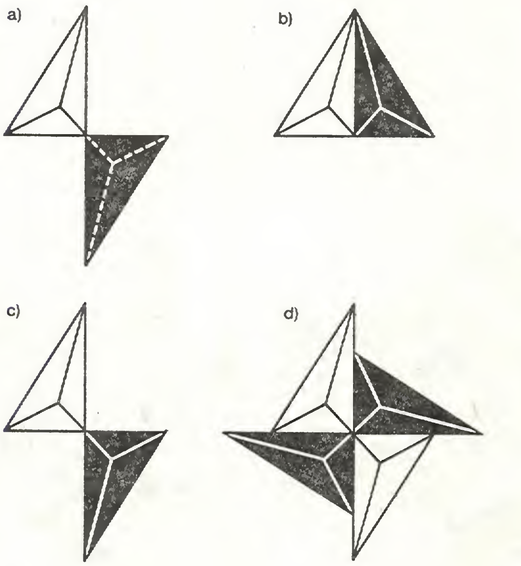
przykłady figur symetrycznych i antysymetrycznych



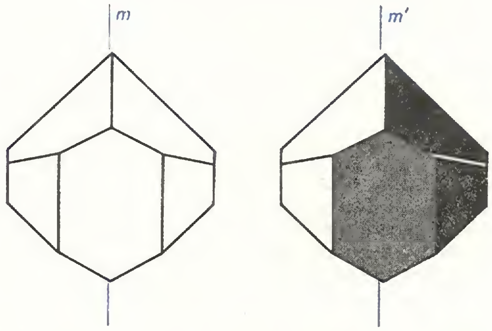
Rys. 14. Działanie 6-krotnej osi symetrii prostopadłej do płaszczyzny rysunku (a) obraz działanie 6-krotnej osi antysymetrii 6′ (b); obróceniu figury o 60° w płaszczyźnie rysunku towarzyszy zmiana barwy figury



Rys. 15. Figury symetryczne (a, b, c) i antysymetryczne (d, e, f). Symetria figur: a) płaszczyzna symetrii *m*; b) dwukrotna oś symetrii 2; c) środek symetrii $\bar{1}$; d) płaszczyzna antysymetrii *m′*; e) dwukrotna oś symetrii 2′, f) środek antysymetrii $\bar{1}'$



Rys. 16. Działanie elementów antysymetrii na asymetryczny czworościan: a) środek antysymetrii $\bar{1}'$, b) płaszczyzna antysymetrii *m′*, c) dwukrotna oś antysymetrii 2′, d) czterokrotna o antysymetrii 4′



Rys. 17. Przykład postaci prostych w klasie symetrii *m* oraz w klasie antysymetrii *m′*

mają jeden znak (barwę), i pokrywające się z klasycznymi grupami punktowymi; 32 grupy neutralne (szare), odpowiadające figurom, których punkty nie mają żadnego znaku; 58 grup o mieszanej polarności (dwubarwnych, czarno-białych), odpowiadających figurom, których punkty symetrycznie równoznaczne mają różne znaki. Szubnikow wprowadził również pojęcie postaci prostych dla grup o mieszanej polarności (rys. 17).

grupy szubnikowskie

Uwzględniając przekształcenie antysymetryczne otrzymuje się 22 dwubarwne typy sieci Bravais'go (prócz 14 zwykłych, jednobarwnych). Grupami szubnikowskimi nazwano, na cześć twórcy teorii, grupy przestrzenne symetrii magnetycznej, tj. grupy przestrzenne uwzględniające antysymetrię. Grup tych jest 1651, a w ich skład wchodzi: 230 przestrzennych grup fiodorowskich (polarnych, białych); 230 grup neutralnych (szarych) oraz 1191 grup o mieszanej polarności (czarno-białych). Grupy szubnikowskie zostały wyprowadzone po raz pierwszy przez A. M. Zamorzajewa (1953) i niezależnie od niego przez N. W. Biełowa, N. N. Neronową i T. S. Smirnową (1955).

antysymetria wielokrotna

Rozwój teorii symetrii nie zakończył się z chwilą pojawienia się prac Szubnikowa o antysymetrii i podobieństwie. W krótkim czasie pojawiły się dalsze uogólnienia tych teorii. Sam Szubnikow, w 1945 r., a więc jeszcze na kilka lat przed ukazaniem się jego książki o antysymetrii stwierdził, że jednemu punktowi można przypisać nie tylko jeden, lecz równocześnie kilka znaków. Tę myśl, aby antysymetrię uogólnić, Zamorzajew i jego szkoła rozwinęli w postaci antysymetrii wielokrotnej, nazwanej też przez nich antysymetrią różnego rodzaju.

W teorii antysymetrii wielokrotnej każdemu punktowi figury przypisuje się skończoną liczbę *l* znaków + lub −, mających różne fizyczne interpretacje.

Pierwszy znak może oznaczać np. barwę białą lub czarną, drugi — znak ładunku elektrycznego itp. Jeżeli np. $l = 2$, tj. gdy punktowi przypisane są dwa znaki, to w wyniku przekształcenia antysymetrycznego pierwszego, drugiego czy trzeciego (mieszanego) rodzaju ulegnie zmianie tylko pierwszy, tylko drugi, albo obydwa znaki równocześnie, np. $(++) \rightleftarrows (-+)$, $(++) \rightleftarrows (+-)$, $(++) \rightleftarrows (--)$. Ogółem istnieje $2^l - 1$ różnych rodzajów antysymetrii.

Pojęciem bliskim antysymetrii wielokrotnej są „grupy złożone" wprowadzone do teorii symetrii przez A. L. Mackaya (1957).

symetria barwna

Innym uogólnieniem antysymetrii jest symetria barwna N. W. Biełowa i T. N. Tarchowej (1956). Śrubowe osie symetrii, płaszczyzny poślizgu i centrowane typy sieci Bravais'go powodują, że zespół punktów symetrycznie równoznacznych (tj. zespół prawidłowy punktów) w położeniu ogólnym rozmieszcza się w komórce elementarnej na kilku poziomach. Zabarwiając na różne kolory rzuty tych punktów na płaszczyznę prostopadłą do krystalograficznej osi Z, otrzymuje się wielobarwny zespół prawidłowy punktów. Dla osi 2_1, 4_2, 6_3, płaszczyzn c, n i sieci A, I, F będzie to dwubarwny zespół punktów. Dwie barwy na płaszczyźnie rzutu odpowiadają więc ułożeniu punktów symetrycznie równoznacznych na dwóch poziomach związanych ze sobą osią śrubową lub płaszczyzną poślizgu. Stosunki symetryczne istniejące w figurach dwubarwnych są po prostu antysymetrią. Inaczej jest, gdy elementy symetrii lub sama sieć powodują rozmieszczenie zespołu punktów symetrycznie równoznacznych na 3, 4 lub 6 poziomach (symetria krystalograficzna). Tego rodzaju rozmieszczenie dają osie 3_1, 3_2, 6_2, 6_4 (trzy poziomy); osie 4_1, 4_3 i płaszczyzna d (cztery poziomy) oraz osie 6_1, 6_5 (sześć poziomów). Rozmieszczenie punktów na trzech poziomach powoduje też sieć romboedryczna (R). Jeśli punkty symetrycznie równoznaczne znajdujące się na różnych poziomach zostaną różnie zabarwione, to na płaszczyźnie rzutu otrzyma się 3-, 4- lub 6-barwny prawidłowy zespół punktów. Mamy wówczas do czynienia z symetrią wielobarwną (barwną). Osie 3_1, 3_2, 6_2, 6_4, 4_1, 4_3, 6_1, 6_5, płaszczyzna d oraz sieć Bravais'go typu R są wielobarwnymi elementami symetrii. Wszystkie przedstawione wyżej uogólnienia symetrii klasycznej obejmuje kwazisymetria, nazywana też P-symetrią (A. M. Zamorzajew). P-symetria jest z kolei szczególnym rodzajem kryptosymetrii (A. Niggli, H. Wondratschek). Kryptosymetria nie obejmuje symetrii kompleksowej, którą z kolei można rozszerzyć do kwaternionowej.

kwazisymetria i kryptosymetria

symetria konformacyjna i afiniczna

Geometrycznym uogólnieniem symetrii podobieństwa jest symetria konformacyjna, w której występują zakrzywione elementy symetrii (kule, koła). Innym geometrycznym rozszerzeniem symetrii podobieństwa jest symetria afiniczna.

Omówione wyżej teorie z biegiem czasu podlegały dalszemu rozwojowi. Rozpoczęto m.in. wyprowadzanie przestrzennych grup homologii i badanie antysymetrii figur homologicznych. Z połączenia barwnej symetrii z antysymetrią powstała barwna antysymetria, którą następnie uogólniono w postaci barwnej antysymetrii różnego rodzaju. Również na symetrię podobieństwa przeniesiono idee antysymetrii prostej i antysymetrii wielokrotnej, a także rozszerzono ją za pomocą pojęć symetrii barwnej i antysymetrii barwnej. Idee antysymetrii i symetrii barwnej przeniesiono także na symetrię konformacyjną. Zostały już zapoczątkowane i rozwijają się nadal prace nad teorią grup przestrzennych Fiodorowa w przestrzeniach wielowymiarowych i geometriach nieeuklidesowych; np. dla $n = 4$ znaleziono punktowe grupy krystalograficzne i wyprowadzono wszystkie sieci Bravais'go.

krystalografia matematyczna

Wszystkie prezentowane wyżej teorie różnych rodzajów symetrii wraz z ich uogólnieniami stanowią obecnie treść nowej, znajdującej się w pełni rozwoju gałęzi krystalografii — krystalografii matematycznej.

Czy osiągnięto już kres rozwoju teorii symetrii? Nie, gdyż w teorii tej nadal istnieją luki, poza tym na opracowanie czekają dalsze problemy, a mianowicie różnorodne syntezy fizycznych i geometrycznych uogólnień teorii symetrii, wyprowadzenie postaci prostych w symetrii podobieństwa, klasyfikacja postaci prostych dla grup punktowych homologii, dokładniejsze zbadanie praw symetrii w obcych jeszcze dzisiaj krystalografom przestrzeniach n-wymiarowych oraz w przestrzeniach nieeuklidesowych geometrii, np. w pseudoeuklidesowej geometrii Minkowskiego, czy w przestrzeni Łobaczewskiego.

Niektóre z przedstawionych wyżej teorii przede wszystkim antysymetria i symetria podobieństwa już obecnie mają znaczenie praktyczne. Antysymetria np. znalazła zastosowanie w opisie magnetycznej struktury kryształów, a także w klasyfikacji bliźniaczych i równoległych zrostów kryształów. Stała się także źródłem twórczej inspiracji w sztuce, czego przykładem są słynne grafiki M. C. Eschera (il. 65, tabl. 17). Symetria podobieństwa natomiast umożliwiła opisanie kształtów muszli amonitu czy spirali wzrostu kryształu, które wymykały się dotąd wszelkiemu ścisłemu opisowi.

N. V. Belov, A. V. Shubnikov *Colored Symmetry*, Oxford 1964; *Idei J. S. Fiodorowa w sowriemiennoj kristałłografii i minierałłogii*, Leningrad 1970; *Issledowanija po diskrietnoj gieometrii*, Kiszyniew 1974; E. I. Galjarskij, A. F. Palistrant, A. M. Zamorzajew *Cwietnaja simmetrija, jeje obobszczenija i priłożenija*, Kiszyniew 1978; W. A. Kopcik, Kristałłografija 12, 755 (1967); W. A. Kopcik *Szubnikowskije gruppy*, Moskwa 1966; W. I. Michiejew *Gomologija kristałłow*, Leningrad 1961; A. Niggli *Antisymmetry, Colour Symmetry and Degenerate Symmetry*, w: *Advances in Structure Research by Diffraction Methods*, vol. 1, New York 1964; G. M. Popow, I. I. Szafranowskij *Kristałłografija*, Moskwa 1972; I. I. Szafranowskij *Simmietrija w prirodie*, Leningrad 1968; A. W. Szubnikow, W. A. Kopcik *Simmietrija w naukie i iskusstwie*, Moskwa 1972; A. W. Szubnikow *Simmietrija i antisimmietrija koniecznych figur*, Moskwa 1951; A. W. Szubnikow, Kristałłografija 5, 489 (1960); B. K. Wajnsztejn *Sowriemiennaja kristałłografija* t. 1, Moskwa 1979; H. Weyl *Symetria*, Warszawa 1960; A. M. Zamorzajew *Tieorija prostoj i krotnoj antisimietrii*, Kiszyniew 1976.

FIZYKA CIAŁA STAŁEGO

Metale

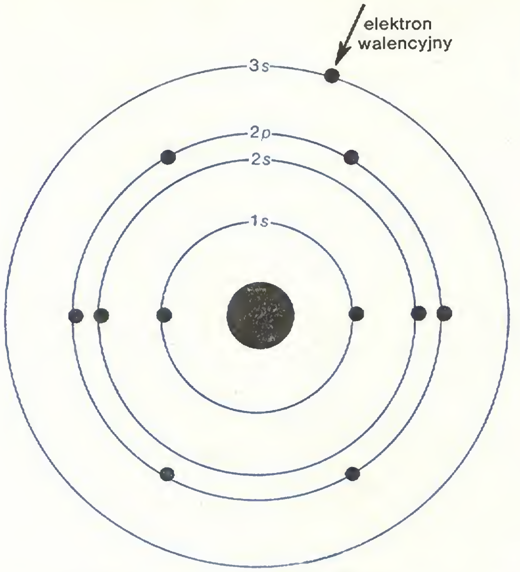
Jacek Furdyna

W życiu codziennym rozpoznajemy metale po ich charakterystycznych cechach zewnętrznych. W temperaturze pokojowej są to ciała stałe o dużej wytrzymałości, mające specyficzny połysk, nie przepuszczające światła, plastyczne (tzn. kowalne) oraz dobrze przewodzące ciepło i elektryczność. Te pozornie niezależne od siebie właściwości mają wspólne podłoże na szczeblu atomowym: elektrony w zewnętrznych powłokach atomów metali (tzw. elektrony walencyjne) są bardzo słabo związane z resztą atomu — tworzą gaz elektronowy. Obecność tego gazu jest, z kolei, podstawą wiązań międzyatomowych metalu i, jak pokażemy, prowadzi do właściwości odróżniających metale od innych ciał stałych.

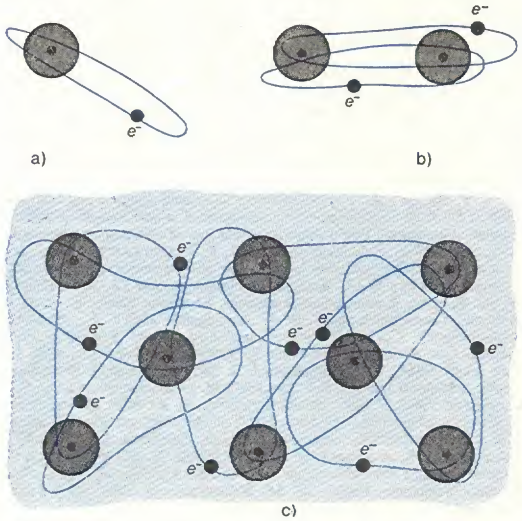
Zacznijmy od atomu metalu o jednym elektronie walencyjnym, np. atomu sodu (rys. 1). Elektron wa-

wiązanie metaliczne

lencyjny (powłoka 3*s* na rys. 1) otacza chmurą elektronową jon sodu Na^+, składający się z jądra i silnie z nim związanych pozostałych elektronów w powłokach 1*s*, 2*s* i 2*p*. Jeżeli przez zbliżenie dwóch atomów sodu utworzymy cząsteczkę Na_2, elektrony walencyjne



Rys. 1. Przykład pierwiastka jednowartościowego — model atomu sodu



Rys. 2. Elektrony walencyjne (e^-) w sodzie: a) w atomie; b) w cząsteczce; c) w krysztale

gaz elektronów walencyjnych

(dzięki swemu słabemu wiązaniu z atomem macierzystym) będą się swobodnie poruszać „na terenie" całej cząsteczki (rys. 2). Podobnie, jeżeli zbliżymy do siebie większą ilość atomów, tworząc w ten sposób kryształ sodu, elektrony walencyjne nie pozostaną zlokalizowane przy „swoich" atomach, lecz będą się poruszać w objętości całego kryształu. Wynika to z zasady nieokreśloności. Elektrony o dużym stopniu lokalizacji mogłyby mieć według tej zasady duże wartości pędu, a więc i wysoką energię. Układ w swoim naturalnym dążeniu do najniższego stanu energii dąży więc równocześnie do delokalizacji elektronów. Tak więc metal składa się z sieci dodatnich jonów, zanurzonych w gazie swobodnie poruszających się elektronów walencyjnych, które straciły bezpośredni związek z atomami macierzystymi i stanowią niejako wspólną własność wszystkich jonów równocześnie. Ten gaz elektronowy tworzy rodzaj kleju przenikającego przestrzeń międzyjonową, który przez oddziaływanie elektrostatyczne „wciąga" w siebie dodatnie jony, gęsto je przez to upakowując. Takie wzajemne przyciąganie jonów i gazu elektronowego stanowi istotę wiązania metalicznego.

Mechaniczne właściwości metali

Wiązanie metaliczne różni się tym od wiązań kowalentnych lub jonowych (→ Chemia kwantowa), że działa bezkierunkowo i nie zależy w zasadzie od wartościowości jonu. Ponieważ wynika ono z bezpośredniego oddziaływania różnych ładunków, wiązanie w metalach jest przy tym bardzo silne. Te cechy (bezkierunkowość i silne przyciąganie) prowadzą do gęstego upakowania (ułożenia) atomów w metalu.

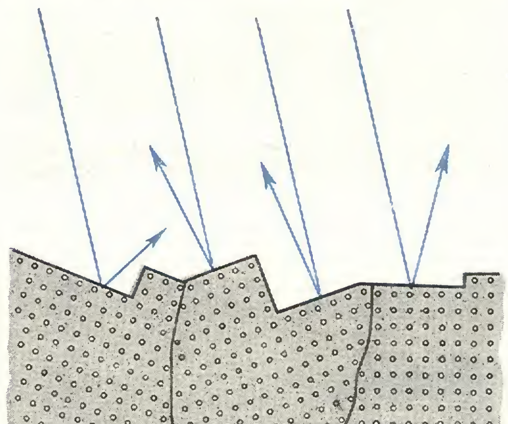
struktura

Najgęstsze upakowanie atomów jest możliwe w trzech strukturach krystalicznych: w układzie regularnym płasko centrowanym, w układzie heksagonalnym najgęstszego upakowania oraz (w nieco mniejszym stopniu) w układzie regularnym przestrzennie centrowanym (→ Budowa kryształów). Struktury te podobne są do różnych sposobów układania jednakowych kul (przykłady widujemy na straganach z owocami). Układy te są również proste, tzn. mają wysoki stopień symetrii. Fakt, że prawie wszystkie metale tworzą kryształy o jednej z trzech powyższych struktur, wynika więc bezpośrednio ze specyficznych właściwości „elektronowego kleju".

stopy

Wspomniana już bezkierunkowość wiązań w metalu i niezależność od wielkości ładunku jonu ma również dalsze konsekwencje. W krysztale miedzi, np. możemy dany atom miedzi (a więc atom jednowartościowy) bez trudności zastąpić dwuwartościowym atomem, np. cynku. Atom obcego metalu odda swoje elektrony walencyjne otaczającemu go gazowi elektronowemu, wiążąc się z siecią kryształu przez powstały w ten sposób ładunek, zupełnie tak samo jak sąsiadujące z nim atomy miedzi. Atomy różnych metali mogą się więc mieszać i układać w dowolnych proporcjach na wspólnej sieci krystalicznej. Ta obojętność wiązania na rodzaj atomu (byleby łatwo oddawał elektrony) jest właściwością, dzięki której możemy tworzyć stopy metali w szerokim zakresie składów, jak również spawać i lutować różne od siebie metale.

spawanie



Rys. 3. Odbicie światła przez ziarna metalu o różnych orientacjach krystalograficznych

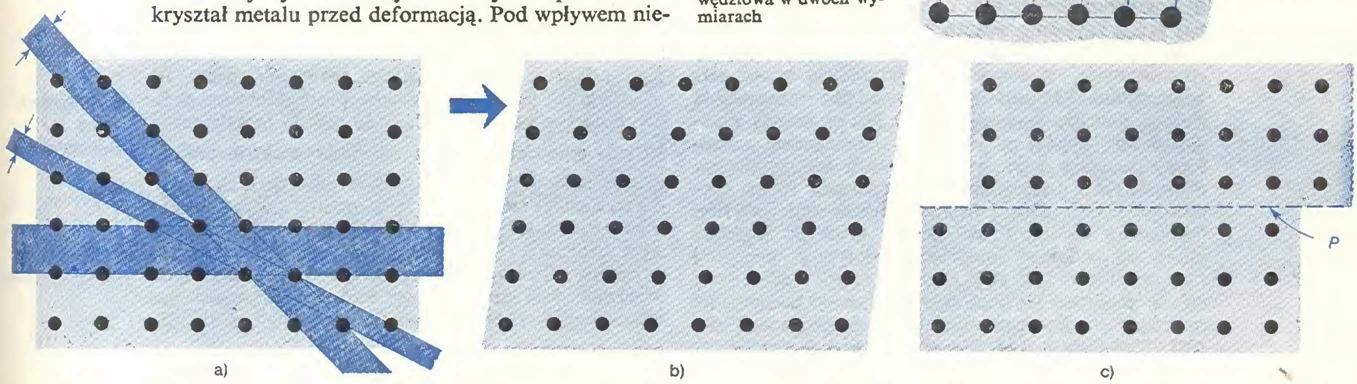
Obiekty metalowe, z którymi stykamy się na co dzień, są przeważnie polikrystaliczne, składające się z wielu ziaren, które często możemy rozróżnić gołym

mikrostruktura

okiem dzięki sposobowi, w jaki pojedyncze ziarna odbijają światło (rys. 3). Każde z ziaren jest monokryształem, oddzielonym od przyległych ziaren warstwą o grubości jednego lub dwu atomów. Krystaliczne uporządkowanie atomów zmienia się więc raptownie na przestrzeni tej warstwy. Mimo to, ze względu na bezkierunkowy charakter wiązania metalicznego, siła kohezji sąsiadujących ze sobą ziaren nie różni się w zasadzie od siły wiążącej atomy wewnątrz kryształu. Rozdzielenie ziaren jest sprawą nader trudną. Stąd też pochodzi wysoki stopień wytrzymałości wyrobów metalowych, co stanowi o ich ogromnej przydatności.

plastyczność (kowalność)

Szczególną cechą metalu jest to, że może on być kuty, gięty i walcowany w różne kształty, zachowując przy tym swą pierwotną wytrzymałość i inne właściwości fizyczne. Proces takiej deformacji plastycznej jest schematycznie pokazany na rys. 4. Punktami zaznaczono jony w sieci kryształu. Rys. 4a przedstawia kryształ metalu przed deformacją. Pod wpływem nie-



Rys. 4. Deformacja metalu w trzech stadiach: a) kryształ przed deformacją; b) odkształcenie sprężyste po przyłożeniu siły zewnętrznej (strzałka); c) odkształcenie plastyczne (nieodwracalne) wskutek poślizgu wzdłuż płaszczyzny *P*, po przekroczeniu granicy sprężystości

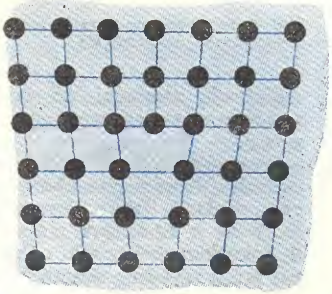
wielkiej siły zewnętrznej (strzałka na rys. 4b) następuje deformacja sprężysta; po usunięciu siły kryształ powraca do pierwotnego kształtu. Jeżeli jednak siła przyłożona (np. uderzenie młotem) przewyższa pewną granicę, metal odkształca się nieodwracalnie. Proces deformacji plastycznej pokazany jest na rys. 4c, gdzie cała płaszczyzna atomów prześlizguje się poziomo o jedną (całkowitą) stałą sieci, zabierając ze sobą górną część kryształu. Rzeczą zasadniczą jest to, że mimo zewnętrznego makroskopowego odkształcenia na szczeblu atomowym odtwarza się po poślizgu (o jedną lub kilka stałych sieci kryształu) pierwotna sytuacja: jony „wskakują" w identyczne położenie równowagi (identyczne otoczenie), w jakim znajdowały się przed deformacją. Mamy więc w rezultacie taki sam materiał i takie same właściwości fizyczne.

Jak widać z rys. 4, plastyczna deformacja metalu wynika z poślizgu wzdłuż płaszczyzn w krysztale. Na rys. 4a liniami przerywanymi oznaczyliśmy trzy różne płaszczyzny sieciowe. W zasadzie poślizg jest możliwy wzdłuż każdej takiej płaszczyzny. Najłatwiej jest jednak uzyskać poślizg wzdłuż płaszczyzn o największym zagęszczeniu atomów (a więc poziomo lub pionowo). Wynika to z dwóch względów: płaszczyzny najgęstszego upakowania są z konieczności najbardziej od siebie oddalone (największa odległość międzypłaszczyznowa A — rys. 4a), co zmniejsza opór wynikający z potencjału jonów w sąsiednich płaszczyznach; poza tym, w trakcie poślizgu wzdłuż takiej płaszczyzny, jony mijają położenia równowagi z największą częstością.

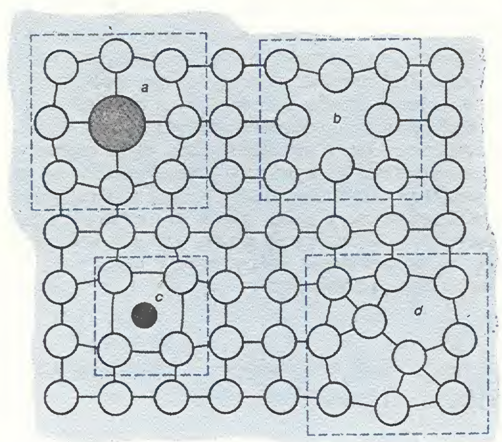
Ze względu na wysoki stopień symetrii, regularne układy krystaliczne mają wiele płaszczyzn o gęstym upakowaniu atomów, co umożliwia łatwy poślizg jednocześnie w kilku kierunkach. Dzięki temu metale o strukturze regularnej (np. miedź, złoto, ołów) mogą z łatwością zmieniać kształt (są łatwo kowalne), bez powstawania przy tym pęknięć i dziur. Metale o strukturze heksagonalnej, a więc o niższej symetrii, są kowalne w mniejszym stopniu, łatwiej się kruszą i mniej nadają do obróbki mechanicznej.

defekty w metalach

Na mechaniczne własności metali ogromny wpływ mają defekty składu lub struktury kryształów. Odróżniamy kilka podstawowych rodzajów defektów. Jeżeli płaszczyzna atomów nagle się urywa, tworząc krawędź (rys. 5), powstałe w ten sposób zaburzenie



Rys. 5. Dyslokacja krawędziowa w dwóch wymiarach

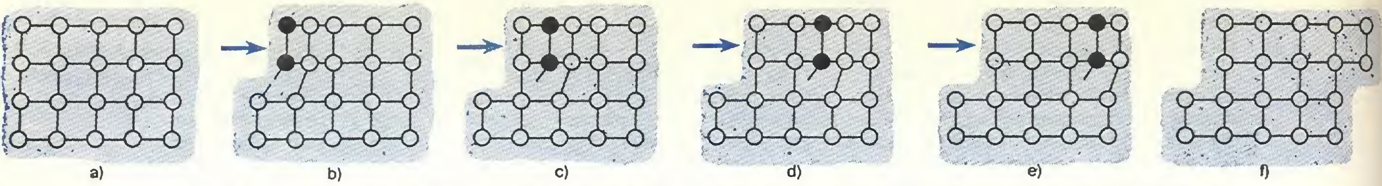


Rys. 6. Defekty punktowe: atom obcy w węźle (*a*); luka w sieci tzw. wakans (*b*); obcy atom w pozycji międzywęzłowej (*c*); dodatkowy atom własny w pozycji międzywęzłowej, tzw. defekt Frenkla (*d*)

sieci nazywamy dyslokacją. Jest to przykład tzw. defektu liniowego. Sieć kryształu może być również zaburzona defektami punktowymi, z których cztery pokazane są schematycznie na rys. 6.

wpływ dyslokacji

Rysunek 7 ilustruje (w dwóch wymiarach), jak ważne jest zjawisko dyslokacji dla plastyczności metali. Pod wpływem zewnętrznej siły (strzałka) wiązania linii atomów zostają zerwane, tworząc krawędź w płaszczyźnie kryształu (rys. 7b, czarne punkty). Pod dalszym naporem siły powstaje krawędź w sąsiedniej płaszczyźnie, zaś pierwotna krawędź „goi się" przez wskoczenie atomów w położenie równowagi. Proces ten powtarza się — mówimy wtedy o ruchu dyslokacji. Ruch dyslokacji możemy nawet oglądać pod mikroskopem. Jak widać z rys. 7, przemieszczanie się



Rys. 7. Schematyczny obraz przemieszczania się dyslokacji pod wpływem siły (strzałka). W rezultacie pierwotny kryształ (a) zostaje odkształcony plastycznie (f)

dyslokacji prowadzi do odkształcenia kryształu, z zachowaniem jego wewnętrznej struktury mikroskopowej.

metale „twarde" i „miękkie"

Szczególne znaczenie dyslokacji polega na tym, że od nich w dużej mierze zależy, czy metal jest „twardy", czy „miękki". Pokazaliśmy już, że kowalność metali łączy się z istnieniem płaszczyzn poślizgu. Dzięki dyslokacjom poślizg taki odbywać się może stopniowo przez zrywanie wiązań w jednym tylko szeregu atomów na raz, a nie na przestrzeni całej płaszczyzny równocześnie, wymaga więc względnie małej siły zewnętrznej. Możemy to porównać do przesuwania dużego dywanu po podłodze. Jeżeli zechcemy przesunąć cały dywan za jednym pociągnięciem, natrafiamy na duży opór. Możemy natomiast zrobić na dywanie fałdę i popychając ją, przesunąć w ten sposób dywan przy użyciu niewspółmiernie mniejszej siły. Obliczenia wykazują, że idealny kryształ, zupełnie wolny od dyslokacji, musi ulec deformacji sprężystej (rys. 4b), rzędu 3 do 10%, zanim nastąpi poślizg (rys. 4c). W praktyce kryształ czystego metalu zaczyna się odkształcać plastycznie już przy deformacjach ok. 0,01%, właśnie dzięki obecności w nim dyslokacji.

Ruch dyslokacji możliwy jest dzięki regularności sieci kryształu. Zaburzenia porządku sieci (np. atomy obce, atomy międzywęzłowe) stanowią więc przeszkodę w poruszaniu się dyslokacji. Toteż, wprowadzając takie defekty, czynimy metale twardszymi. Złoto czternastokaratowe, zawierające domieszki, jest o wiele bardziej wytrzymałe („twardsze") niż złoto zupełnie czyste. Zaburzeniem sieci są również połączenia ziaren w polikryształach, można więc uzyskać twardszy metal przez zmniejszenie rozmiaru ziaren.

hartowanie

Hartując metal przez kucie lub walcowanie, wprowadzamy zaburzenie sieci, którym są — paradoksalnie — nowe dyslokacje. Jeżeli gęstość dyslokacji jest dostatecznie duża, dyslokacje poruszające się wzdłuż przecinających się płaszczyzn po prostu wchodzą sobie w drogę, uniemożliwiając dalszy ruch i przez to czyniąc metal mniej kowalnym.

Ruch elektronów w metalu — obraz klasyczny

Wprawdzie od początku podkreślaliśmy rolę elektronów walencyjnych w mechanizmie kohezji, wiążącym metal w ciało stałe, jednak nasza uwaga skupiała się dotychczas głównie na dodatnich jonach, na ich układzie i ruchu. Zajmiemy się teraz bezpośrednio samymi elektronami walencyjnymi, którym przypisujemy niektóre najciekawsze i najbardziej charakterystyczne właściwości metali.

Pełny obraz dynamiki elektronów w metalach jest w zasadzie możliwy tylko w ramach mechaniki kwantowej. Zanim jednak do tego przystąpimy, opiszemy

model Drudego

pokrótce model klasyczny wprowadzony przez P. Drudego w 1900 r., prawie natychmiast po odkryciu elektronu przez J. J. Thomsona. Na podstawie modelu Drudego będziemy w stanie opisać, przynajmniej w sposób jakościowy, główne zjawiska elektryczne, termiczne i optyczne zachodzące w metalach. Model klasyczny daje przy tym intuicyjny, uchwytny dla wyobraźni wgląd w procesy elektronowe. Jego usterki natomiast uwypuklają subtelności rzeczywistej sytuacji i istotę poprawek, jakie do obrazu klasycznego wprowadza mechanika kwantowa.

Model Drudego traktuje elektrony walencyjne metalu jako gaz elektronów poruszających się swobodnie po całej objętości ciała. W trakcie swego ruchu elektrony napotykają jony metalu, z którymi (wg dosłownej interpretacji modelu Drudego) zderzają się i zostają w ten sposób rozpraszane w różnych kierunkach. Można by stąd wnioskować, że średnia droga między zderzeniami (tzw. droga swobodna) porównywalna jest z odległością między jonami, a więc jest rzędu 1 Å.

przewodnictwo elektryczne

Obraz ten tłumaczy zjawisko przewodnictwa elektrycznego w metalach. Jeżeli do próbki metalu przyłożymy pole elektryczne $\vec{E}$, swobodny elektron zostaje przyśpieszony (w kierunku przeciwnym do $\vec{E}$ ze względu na ujemny ładunek). Przyśpieszanie trwa aż do momentu zderzenia, które rozprasza elektron w dowolnym kierunku, po czym proces przyśpieszania zaczyna się od nowa. Na skutek tych przeciwstawnych wpływów pola elektrycznego i zderzeń powstaje stan równowagi — elektron porusza się wtedy ze średnią prędkością proporcjonalną do pola $\vec{E}$, tzw. prędkością unoszenia (dryfową). Jak więc widać, wypadkowy transport ładunku elektrycznego, tzn. prąd elektryczny, proporcjonalny jest do pola $\vec{E}$. Obraz Drudego daje nam zatem mikroskopowe uzasadnienie prawa Ohma.

Traktowanie elektronów w metalu jako gazu swobodnych, niezależnych od siebie cząstek klasycznych prowadzi również do wniosku, że z każdym elektronem wiąże się energia kinetyczna $\frac{3}{2}kT$, gdzie k jest stałą Boltzmanna, a T — temperaturą bezwzględną kryształu.

przewodnictwo cieplne

Dzięki swobodzie poruszania się elektrony mogą zatem z łatwością pośredniczyć w przewodzeniu ciepła, co tłumaczy bardzo duże na ogół przewodnictwo cieplne w metalach.

Niestety ten prosty, klasyczny obraz gazu elektronowego zawodzi przy zastosowaniu do kilku innych kluczowych zjawisk. Rozpatrzmy np. wyżej wspomniane założenie modelu, że w temperaturze T każdy elektron walencyjny wnosi energię kinetyczną $\frac{3}{2}kT$. Zastosowanie tego założenia do zagadnienia ciepła właściwego prowadzi do zupełnie błędnej wartości, przewyższającej wartość eksperymentalną o kilka rzędów wielkości. Co więcej, ponieważ każdy elektron ma moment magnetyczny związany ze spinem, należałoby się spodziewać, że przy przyłożeniu pola magnetycznego elektronowe momenty magnetyczne uporządkują się wzdłuż pola i metal wykaże wysoką wartość namagnesowania.

podatność magnetyczna

Tamczasem doświadczalna wartość podatności magnetycznej metalu jest znowu o kilka rzędów wielkości mniejsza od wartości przewidywanej na podstawie klasycznego obrazu gazu elektronowego. Nie rozpatrujemy tu metali ferromagnetycznych, których właściwości magnetyczne (→ Teoria magnetyzmu) wiążą się z głębszymi powłokami atomu. Eksperymentalne wartości ciepła właściwego i podatności magnetycznej sugerują więc, że jedynie drobna część elektronów walencyjnych, a nie cały gaz elektronowy, bierze udział w tych procesach.

Obraz Drudego również nie jest w stanie wyjaśnić, dlaczego przewodnictwo elektryczne metali bardzo szybko wzrasta z obniżaniem się temperatury, często osiągając w temperaturze ciekłego helu (w bardzo czystych kryształach) wartości 10^4–10^5 razy większe

niż w temperaturze pokojowej. W końcu założenie, że droga swobodna elektronu w metalu jest rzędu stałej sieci kryształu, również zupełnie nie zgadza się z doświadczeniem. Na przykład w bardzo czystych kryształach galu w temperaturze ciekłego helu długość drogi swobodnej jest kilkaset milionów razy większa niż odległość międzyatomowa kryształu.

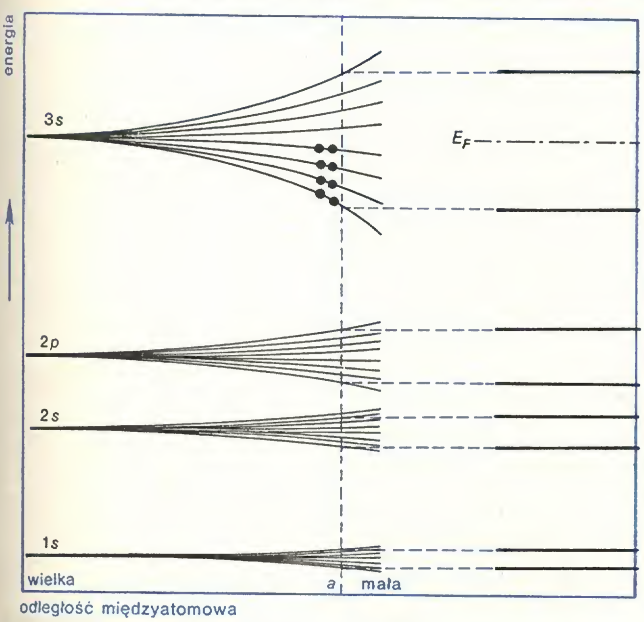
Ruch elektronów w metalu — obraz kwantowy

Model przedstawiający elektrony walencyjne metalu jako gaz swobodnie poruszających się ładunków jest w zasadzie zgodny z rzeczywistością. Błąd modelu Drudego wynika z dwóch dalszych, mylnych założeń: po pierwsze z założenia, że wszystkie elektrony mają jednakowy wpływ na wartość ciepła właściwego i podatności magnetycznej; po drugie z założenia, że elektrony ulegają rozpraszaniu na stacjonarnych jonach w sieci kryształu. Te usterki modelu wynikają bezpośrednio z traktowania elektronu jako cząstki klasycznej. Jak się zatem przedstawia rzeczywista (a więc kwantowa) sytuacja?

Wróćmy do pojedynczego atomu sodu. Według praw mechaniki kwantowej elektrony atomu znajdują się na specyficznych poziomach energetycznych (oznaczonych na rys. 1 jako powłoki 1*s*, 2*s*, 2*p*, i 3*s*). Liczba elektronów przypadających na każdą powłokę jest ściśle ograniczona zakazem Pauliego, skąd też wynika struktura powłokowa poszczególnych atomów (→ Chemia kwantowa).

poziomy energetyczne

Stany kwantowe elektronów w ciele stałym wywodzą się bezpośrednio ze struktury elektronowej atomu. Możemy uwidocznić to w sposób następujący. Wyobraźmy sobie atomy w węzłach sieci fikcyjnego kryształu, w którym odległość międzyatomowa (stała sieci) może być dowolnie zmieniana. Jeżeli odległość międzyatomowa jest wielokrotnie większa niż w krysztale rzeczywistym, atomy są odizolowane i ich kwantowe stany energetyczne nie różnią się od stanów pojedynczego atomu (lewa strona rys. 8). Gdy stała sieci fikcyjnego kryształu stopniowo się zmniejsza, oddziaływanie między atomami rozszczepia każdy poziom



Rys. 8. Powstawanie pasm energetycznych wskutek oddziaływania między atomami, na przykładzie sodu (*a* oznacza stałą sieci kryształu rzeczywistego). Jak widać, pasmo odpowiadające elektronom walencyjnym jest tylko częściowo obsadzone — do energii Fermiego E_F

elektronowy na szereg podpoziomów. Z każdego poziomu powstaje więc pasmo stanów energetycznych.

pasma energetyczne

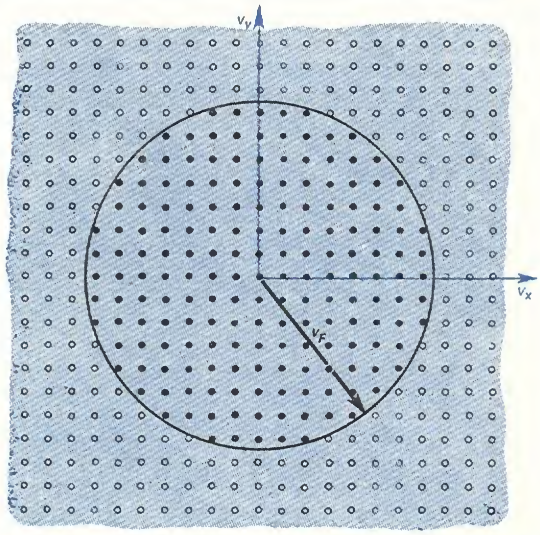
Jak widać z rys. 8 (i jak należałoby się spodziewać), zewnętrzne powłoki atomu są najbardziej rozszczepione wskutek oddziaływania z pozostałymi atomami kryształu. Widzimy również, że elektrony w krysztale nie mogą mieć energii w przedziałach pomiędzy dozwolonymi pasmami (tzw. w przerwach energetycznych).

Wielkość pasm i przerw energetycznych w danym krysztale zależy więc od rzeczywistej odległości międzyatomowej *a* (linia przerywana na rys. 8) oraz od rodzaju danego pierwiastka. Natomiast liczba stanów elektronowych wewnątrz pasma równa się liczbie atomów w krysztale (ściślej — liczbie komórek elementarnych w krysztale). Ponieważ w ciele stałym koncentracja atomów jest rzędu 10^{23} na cm^3, stany te leżą tak blisko siebie, że energię wewnątrz pasma możemy w praktyce uważać za ciągłą.

Kwantowy gaz elektronowy

Rozważmy najwyższe pasmo energetyczne, odpowiadające elektronom walencyjnym (3*s* na rys. 8). W krysztale składającym się z *N* atomów pasmo zawiera *N* stanów kwantowych. Każdemu z tych stanów przypisujemy prędkość $\vec{v}$, która odróżnia dany stan od pozostałych. Wygodnie jest przedstawić to za pomocą tzw. przestrzeni prędkości, pokazanej w dwóch wymiarach na rys. 9. Chodzi tu o przestrzeń formalną, matematyczną, gdzie wzdłuż osi *x*, *y*, *z* odkładamy składowe v_x, v_y, v_z wektora prędkości $\vec{v}$. Ponieważ stany elektronowe w pasmie odpowiadają dy-

przestrzeń prędkości



Rys. 9. Przestrzeń prędkości w dwóch wymiarach. Punkty czarne oznaczają stany obsadzone, kółka puste odpowiadają stanom nieobsadzonym; $\vec{v}_F$ oznacza prędkość Fermiego

skretnym wartościom $\vec{v}$, każdy taki stan można przedstawić oddzielnym punktem w przestrzeni prędkości. Wszystkie stany pasma będą w ten sposób przedstawione siecią *N* punktów (rys. 9). Każdemu z dozwolonych stanów (a więc każdemu punktowi) przypisujemy dyskretną wartość energii kinetycznej $E = {}^1/_2\, mv^2$. Tak więc stany mające tę samą wartość *v*, a różniące się jedynie kierunkiem ruchu (rys. 9, stany leżące na okręgu o promieniu *v*) odpowiadają tej samej wartości energii.

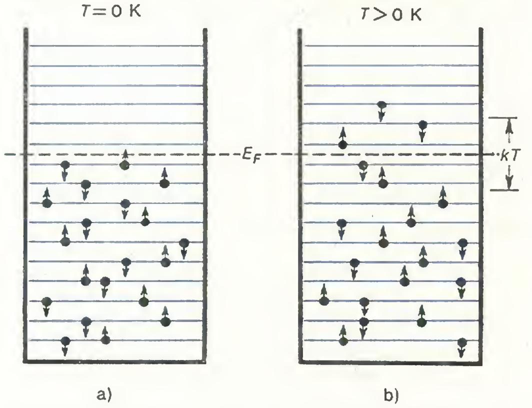
zakaz Pauliego

Zgodnie z zakazem Pauliego, w danym stanie kwantowym mogą znajdować się najwyżej dwa elektrony (o odwrotnych kierunkach spinów). Odnosi się to nie tylko do stanów elektronowych w pojedynczych atomach (o czym wspominaliśmy powyżej), lecz również

i do stanów w pasmach energetycznych kryształu. Zastosowanie zakazu Pauliego do pasma elektronów walencyjnych wprowadza ogromnie ważne, zupełnie nowe (a więc obce obrazowi klasycznemu) właściwości. Elektrony walencyjne metalu stopniowo obsadzają stany wewnątrz pasma, począwszy od najniżej położonych poziomów energetycznych (tzn. od najniższych wielkości v) wzwyż, po dwa elektrony na każdy stan.

pasmo przewodnictwa

Pasmo może więc pomieścić $2N$ elektronów. Jeżeli w danym materiale mamy mniej niż $2N$ elektronów walencyjnych (np. w sodzie mamy ich dokładnie N), pasmo jest tylko częściowo obsadzone. Pasmo takie nazywamy z przyczyn, które wyjaśnią się poniżej, pasmem przewodnictwa. W stanie podstawowym (co odpowiada temperaturze zera bezwzględnego) pasmo przewodnictwa jest obsadzone do pewnej wartości

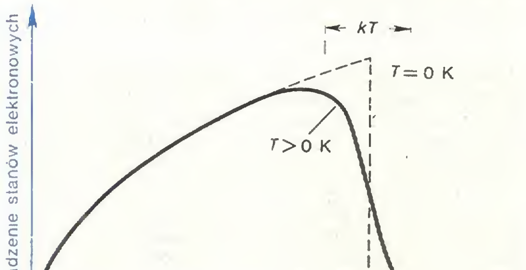


Rys. 10. Rozkład Fermiego–Diraca: a) dla $T = 0$ K; b) dla temperatury T wyższej od 0 K

energia Fermiego

energii E_F, określonej przez liczbę elektronów walencyjnych metalu (rys. 10a). Energię tę nazywamy energią Fermiego. Charakterystyczną cechą każdego metalu jest posiadanie, w stanie podstawowym, takiego częściowo obsadzonego pasma przewodnictwa.

Punktem wyjściowym dla omówienia wielu zjawisk elektronowych w metalach jest rozkład energetyczny elektronów w pasmie przewodnictwa, otrzymany przez E. Fermiego i P. Diraca przy uwzględnieniu zakazu Pauliego. Rozkład Fermiego-Diraca, określający liczbę elektronów przypadającą na dany obszar energii, pokazaliśmy na rys. 11. Wzrost obsadzenia elektronowego (tzn. liczba elektronów przypadająca w danym przedziale energii) wraz ze wzrostem energii można zrozumieć rysując koło o promieniu v (rys. 9): im większa wartość energii (a więc wielkość v), tym więcej odpowiada jej stanów. W temperaturze 0 K rozkład urywa się skokowo na energii Fermiego (krzywa przerywana na rys. 11), w temperaturze wyższej od 0 K — jest krzywą ciągłą (rys. 11).



Rys. 11. Obsadzenie stanów elektronowych w funkcji energii w temperaturze zera bezwzględnego (stan podstawowy metalu) oraz w temperaturze T (przy $T = 300$ K $kT/E_F \approx 0{,}005$)

rozkład Fermiego-Diraca energii elektronów

Rozkład Fermiego–Diraca uwidacznia dwie zasadnicze różnice pomiędzy kwantowym gazem elektronowym a klasycznym gazem Drudego. W gazie klasycznym wszystkie elektrony reagują na wpływ temperatury T, zyskując średnią energię kinetyczną $^3/_2kT$. W gazie kwantowym natomiast (jak widać z rys. 10b i 11), jedynie elektrony odpowiadające stanom w bezpośredniej bliskości energii Fermiego przechodzą ze wzrostem temperatury T na wyższe poziomy energetyczne. Po drugie, w gazie klasycznym przy $T = 0$ K wszelki ruch ustaje całkowicie, podczas gdy w gazie kwantowym zakaz Pauliego „wymaga", aby elektrony poruszały się ze skończoną prędkością nawet w temperaturze zera bezwzględnego (rys. 9). W metalach, ze względu na dużą koncentrację elektronów walencyjnych, prędkości te dochodzą do 10^8 cm/s. W gazie klasycznym prędkości takie odpowiadałyby wzbudzeniu termicznemu 50 000 K.

prędkość Fermiego

Górną granicę prędkości dla $T = 0$ K, odpowiadającą energii Fermiego, nazywamy prędkością Fermiego v_F. Prędkość tę przedstawiliśmy na rys. 9 jako promień v_F. Ma to następujące znaczenie fizyczne: w temperaturze $T = 0$ K wszystkie obsadzone stany znajdują się wewnątrz kuli o promieniu v_F. Powierzchnia kuli o promieniu v_F stanowi zatem granicę między stanami obsadzonymi i nieobsadzonymi. Powierzchnię tę nazywamy powierzchnią Fermiego.

powierzchnia Fermiego

Powierzchnia Fermiego jest oczywiście pojęciem czysto formalnym, przedstawiającym rozkład prędkości w kwantowym gazie elektronowym. Jej ogromna przydatność przy omawianiu metali wynika stąd, że w wielu kluczowych zjawiskach elektronowych (przewodnictwo elektryczne, zachowanie się ciepła właściwego, podatności magnetycznej) udział biorą jedynie stany elektronowe na powierzchni Fermiego lub położone w bezpośredniej jej bliskości.

Możemy to wyjaśnić następująco. Dozwolone stany elektronowe, choć dyskretne, położone są energetycznie bardzo blisko siebie. Energia ΔE pobrana przez elektron z zewnętrznego pola elektrycznego, z energii cieplnej, względnie z pola magnetycznego jest wprawdzie znikomo mała w porównaniu z E_F, ale jest porównywalna z odstępami dzielącymi poziomy energetyczne wewnątrz pasma. Może więc ona powodować zmiany w rozkładzie elektronów. Przejścia elektronowe są jednak możliwe jedynie wtedy, gdy w przedziale energii ΔE, liczonej od poziomu obsadzonego przez dany elektron, znajdują się stany wolne. Zatem będą zachodziły tylko przejścia elektronów ze stanów w bezpośrednim sąsiedztwie powierzchni Fermiego (jak to pokazaliśmy dla pobudzenia termicznego, tzn. dla energii $\Delta E = {^3/_2}kT$, na rys. 10b i 11), bowiem stany powyżej energii E_F są nieobsadzone. Rozważmy jednak elektron na poziomie głęboko poniżej energii Fermiego. Wszelkie stany w zasięgu energii ΔE od tego poziomu są całkowicie obsadzone. Takie głębokie poziomy są więc niejako „zablokowane" i nie biorą wskutek tego udziału w zjawiskach, w których zmiany energii powodowane polem zewnętrznym są małe w porównaniu z E_F.

właściwości magnetyczne metalu

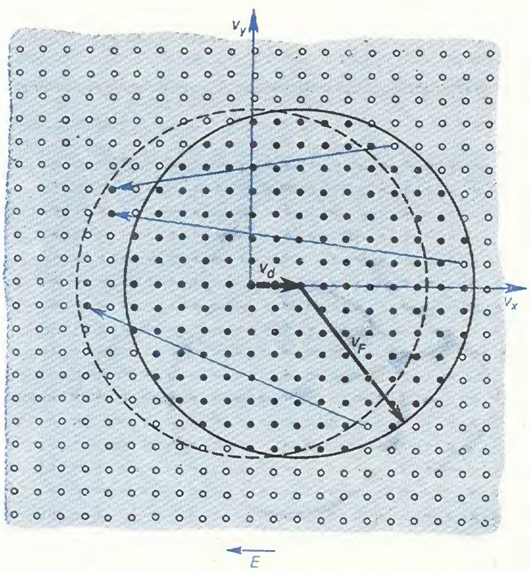
Omawiając niepowodzenia teorii Drudego pokazaliśmy, że ciepło właściwe i podatność magnetyczna metali mają takie wartości, jak gdyby jedynie znikoma liczba elektronów miała na nie wpływ. Możemy to teraz natychmiast wytłumaczyć na podstawie rozkładu Fermiego–Diraca. Jak już mówiliśmy, jedynie elektrony obsadzające stany tuż przy powierzchni Fermiego mogą reagować na zmiany temperatury, pobierać energię cieplną i wpływać na wartość ciepła właściwego. Z podatnością magnetyczną jest podobnie. W stanie podstawowym metalu na każdy dozwolony stan elektronowy wewnątrz powierzchni Fermiego przypadają dwa elektrony o odwrotnych kierunkach spinu. W tej sytuacji wypadkowy moment magnetyczny metalu równa się oczywiście zeru. Aby namagnesować metal przez przyłożenie pola magnetycznego, należy część spinów odwrócić, co (ze względu na zakaz Pauliego i na całkowite obsadzenie poziomów

położonych poniżej E_F) można uzyskać jedynie przez przejścia elektronów na poziomy powyżej E_F. Ponieważ energia oddziaływania elektronów z polem magnetycznym jest mała w stosunku do E_F, spośród wszystkich elektronów pasma przewodnictwa tylko te odpowiadające poziomom tuż przy powierzchni Fermiego przyczyniają się do właściwości magnetycznych metalu.

Przewodnictwo elektryczne

Sprawa przewodnictwa elektrycznego w metalach wymaga dokładniejszego omówienia w świetle kwantowego obrazu. Elektrony walencyjne w metalu (tzn. w pasmie przewodnictwa) są w ciągłym ruchu. Jak widać jednak z rys. 9, w nieobecności pola elektrycznego każdemu elektronowi poruszającemu się w danym kierunku odpowiada inny elektron poruszający się z dokładnie taką samą prędkością w kierunku przeciwnym. W rezultacie, mimo ruchu ładunków, wypadkowy prąd elektryczny jest równy zeru. Jeżeli teraz przyłożymy pole elektryczne $\vec{E}$, spowoduje ono przejścia, które odpowiadają przyspieszaniu (a więc zwiększaniu wartości v) elektronów poruszających się w kierunku przeciwnym do pola (ze względu na ujemny ładunek elektronów) oraz hamowaniu (tzn. zmniejszaniu v) elektronów poruszających się w kierunku pola. W rezultacie powierzchnia Fermiego (tzn. powierzchnia, wewnątrz której znajdują się stany obsadzone) zostaje przesunięta w przestrzeni prędkości, jak na rys. 12.

przesunięcie powierzchni Fermiego



Rys. 12. Przesunięcie kuli Fermiego w przestrzeni prędkości pod wpływem pola elektrycznego. Koło zaznaczone linią przerywaną odpowiada rozkładowi prędkości w nieobecności pola. Wskutek rozproszenia elektrony powracają do nieobsadzonych stanów pierwotnego rozkładu (strzałki niebieskie)

Przyspieszanie elektronów przez pole $\vec{E}$ (a więc przesuwanie powierzchni Fermiego) postępowałoby w nieskończoność, gdyby nie rozpraszanie elektronów. Wskutek rozpraszania otrzymujemy przejścia ze świeżo obsadzonych (pod wpływem pola $\vec{E}$) stanów do stanów nieobsadzonych, jak pokazano strzałkami na rys. 12. W rezultacie, dzięki współzawodnictwu pomiędzy przyspieszaniem i rozpraszaniem, przesunięcie powierzchni Fermiego osiąga pewną stałą wartość v_u, proporcjonalną do wielkości pola E. Jest to średnia prędkość elektronów w pasmie przewodnictwa; nazywamy ją prędkością unoszenia.

prędkość unoszenia

Należy zauważyć, że większość stanów prędkości wewnątrz przesuniętej kuli Fermiego była obsadzona również i w nieobecności pola. Prądy elektronowe wynikające z tych stanów znoszą się (jak znosiły się przed przyłożeniem pola) i nie przyczyniają się do wypadkowego prądu elektrycznego. Ponieważ w praktyce przesunięcie v_u jest nadzwyczaj małe w porównaniu z promieniem v_F, jedynie stany położone zupełnie blisko powierzchni Fermiego ulegają zmianie pod wpływem pola $\vec{E}$. Możemy więc powiedzieć, że tylko elektrony na powierzchni Fermiego biorą udział w przewodnictwie elektrycznym.

Widzimy również na podstawie tego obrazu, że gdyby wszystkie stany pasma (wszystkie punkty na rys. 9) były obsadzone, żadne zmiany w rozkładzie prędkości elektronów nie mogłyby nastąpić pod wpływem pola. Istnienie pasm tylko w części zapełnionych jest więc konieczne dla przewodzenia prądu elektrycznego i dlatego też pasma takie nazywamy pasmami przewodnictwa.

pasmo przewodnictwa

Rozpraszanie elektronów

Według dosłownej interpretacji modelu klasycznego elektrony są rozpraszane przez zderzenia z dodatnimi jonami metalu, co ogranicza drogę swobodną elektronu do kilku angstremów. Jak już wspominaliśmy, wniosek ten jest w zupełnej sprzeczności z danymi doświadczalnymi. Aby sformułować obraz właściwy, należy podkreślić, że w kwantowej mechanice elektrony mają charakter podwójny: zachowują się jak cząstki materii, ale równocześnie jak fale. Zgodnie z tym obrazem elektron możemy traktować jako falę płaską, której długość (tzw. długość fali de Broglie'a) wyraża się wzorem $\lambda = h/p$, gdzie $p = mv$ jest wielkością pędu elektronu, a h — stałą Plancka.

W oddziaływaniu z jonami metalu, uporządkowanymi periodycznie w sieci kryształu, falowy charakter elektronów odgrywa decydującą rolę. Można bowiem pokazać, że w doskonale periodycznym układzie (jakim jest kryształ idealny) fala płaska rozchodzi się bez tłumienia. W zastosowaniu do elektronu w krysztale oznacza to, że dzięki swej falowej naturze, może on poruszać się na przestrzeni idealnego kryształu zupełnie bez rozpraszania!

Rozpraszanie więc odbywa się nie na jonach tworzących periodyczny układ kryształu, lecz na zaburzeniach idealnie periodycznej sieci. Elektron może się zderzać z napotykanymi na swej drodze obcymi atomami (domieszkami), z defektami struktury kryształu oraz z termicznymi drganiami sieci, fononami (→ Dynamika sieci krystalicznej), powodującymi chwilowe odchylenia jonów od położenia równowagi. Obraz falowo-korpuskularny elektronu od razu więc tłumaczy, dlaczego droga swobodna elektronów, mierzona w niskich temperaturach w bardzo czystych próbkach o wysokim stopniu doskonałości krystalicznej, jest tak długa.

rozpraszanie na zaburzeniach sieci

Ruch elektronów w kryształach rzeczywistych

Chcąc podkreślić podstawowe własności kwantowe gazu elektronowego przedstawiliśmy, jak dotychczas, obraz nieco wyidealizowany, nie uwzględniliśmy bowiem wpływu dodatnich jonów metalu na ruch elektronów (poza wzmianką, że w periodycznej sieci kryształu jony nie powodują rozpraszania). W rzeczywistości, jak pokażemy, wpływ sieci kryształu na dynamikę elektronów jest bardzo istotny.

Ustaliliśmy już, że na każde pasmo energetyczne przypada tyle oddzielnych, różnych od siebie stanów elektronowych, ile mamy komórek elementarnych w objętości kryształu (co u większości metali, dzięki ich prostej strukturze, równa się po prostu liczbie atomów). Stany te wynikają z oddziaływania między atomami tworzącymi sieć kryształu. Oczywiście w oddzia-

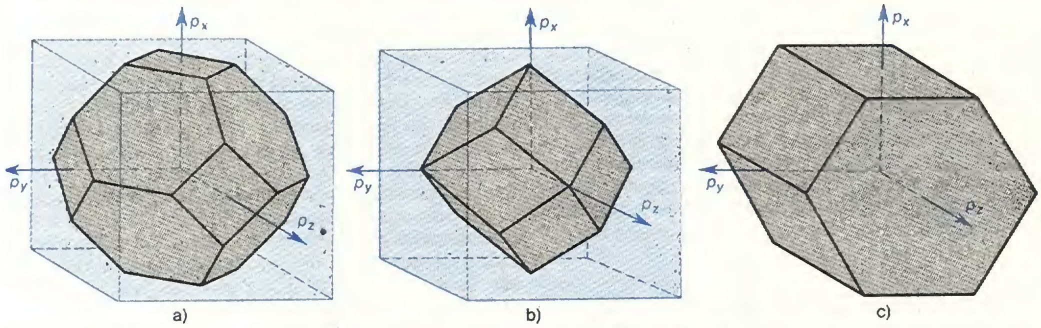
ływaniu tym decydującą rolę odgrywa wzajemne położenie atomów, a więc geometria sieci.

strefy Brillouina

Po uwzględnieniu szczegółów krystalograficznych otrzymujemy następujący obraz: stany dozwolone pasma (powiedzmy, że jest ich *N*) tworzą siatkę składającą się z *N* punktów w przestrzeni prędkości (jak na rys. 9). Punkty te jednorodnie wypełniają bryłę (w przestrzeni prędkości) o kształcie określonym przez symetrię danego kryształu (np. zbiór punktów na rys. 9 odpowiada dwuwymiarowej sieci kwadratowej). Bryłę tę, zawierającą wszystkie stany pasma (bez względu na to, czy są one obsadzone czy nie) nazywamy strefą Brillouina.

przestrzeń pędu

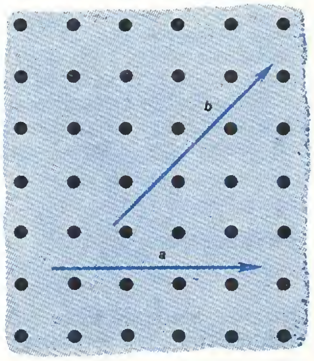
W praktyce, ze względów formalnych, stosuje się tu nie przestrzeń prędkości, lecz analogiczną przestrzeń pędu, gdzie wzdłuż osi x, y, z odkłada się składowe pędu p_x, p_y, p_z, odpowiadające poszczególnym stanom elektronowym. Przykłady stref Brillouina w przestrzeni pędu dla najczęściej spotykanych struktur krystalicznych metali pokazaliśmy na rys. 13. Pędy p w bezpośrednim sąsiedztwie granic strefy odpowiadają falom elektronowym o długości $\lambda = h/p$ porównywalnej ze stałą sieci. Stany odpowiadające takim pędom są przez to dużo bardziej czułe na wpływ sieci kryształu niż stany odpowiadające mniejszym wartościom pędu.



Rys. 13. Strefy Brillouina w przestrzeni pędu dla kryształów o strukturze: a) regularnej płasko centrowanej; b) regularnej przestrzennie centrowanej; c) heksagonalnej

energia kinetyczna elektronu w krysztale

Rozważmy następnie wpływ struktury kryształu na ruch danego elektronu, poruszającego się np. w dwuwymiarowym układzie pokazanym na rys. 14. Elektron taki porusza się w na przemian przyśpieszającym go i hamującym polu elektrostatycznym jonów, co wpływa na jego energię kinetyczną. Elektron poruszający się wzdłuż linii poziomej (*a* na rys. 14) napotyka jony w mniejszych odstępach czasu niż np. elektron poruszający się z tą samą prędkością wzdłuż linii *b*. Mimo identycznej prędkości elektrony te, ze względu na różną częstość oddziaływań z jonami, mają różne energie kinetyczne. Podobnie elektron poruszający się z dużą prędkością odczuwa wpływ jonów inaczej niż elektron poruszający się wolniej, chociaż w tym samym kierunku.



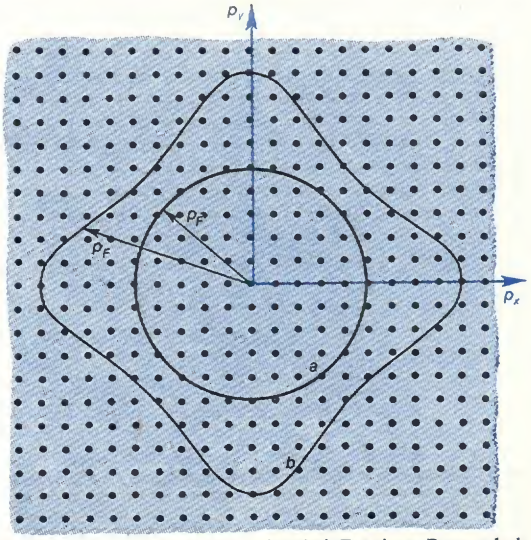
Rys. 14. Ruch elektronu (strzałki) w dwóch różnych kierunkach w krysztale

Wynikające w ten sposób komplikacje w zależności energii kinetycznej od wielkości i kierunku pędu mają bezpośredni wpływ na kształt powierzchni Fermiego.

powierzchnia Fermiego w kryształach rzeczywistych

Powierzchnię tę otrzymaliśmy pierwotnie w kształcie kuli, opierając się na założeniu, że energia kinetyczna zależy tylko od wielkości pędu, $E = \frac{1}{2}mv^2 = p^2/2\,m$. W rzeczywistości, jak się okazuje, ta sama wartość energii może odpowiadać dwu różnym wielkościom prędkości (względnie pędu) w dwu różnych kierunkach. Zatem, wskutek anizotropii kryształu, powierzchnia Fermiego przedstawiająca zbiór wszystkich pędów, odpowiadających tej samej wartości energii E_F, ulega zniekształceniu.

Wpływ struktury kryształu, zniekształcający kulę Fermiego, możemy zilustrować za pomocą nieco uproszczonego obrazu. Na rys. 15 układ punktów przedstawia stany kryształu wypełniające strefę Brillouina. Przypominamy, że stany leżące w pobliżu granicy strefy są dużo bardziej podatne na wpływ sieci kryształu niż stany w głębi strefy (odpowiadające długim falom de Broglie'a). Jeżeli energia Fermiego E_F jest dużo niższa od górnej granicy pasma (np. pasmo wypełnione ao połowy, jak w wypadku sodu), wartości pędu odpowiadające tej wartości E_F leżą z dala od granicy strefy Brillouina. Wpływ potencjału sieci na stany obsadzone jest wtedy względnie mały i powierzchnia Fermiego jest w dobrym przybliżeniu kulą (*a* na rys. 15). Jeżeli natomiast energia Fermiego jest wyższa (jak np. dla bardziej skomplikowanych dwu- lub więcej wartościowych metali), tak że pęd odpowiadający E_F zbliża się do granicy strefy, wpływ



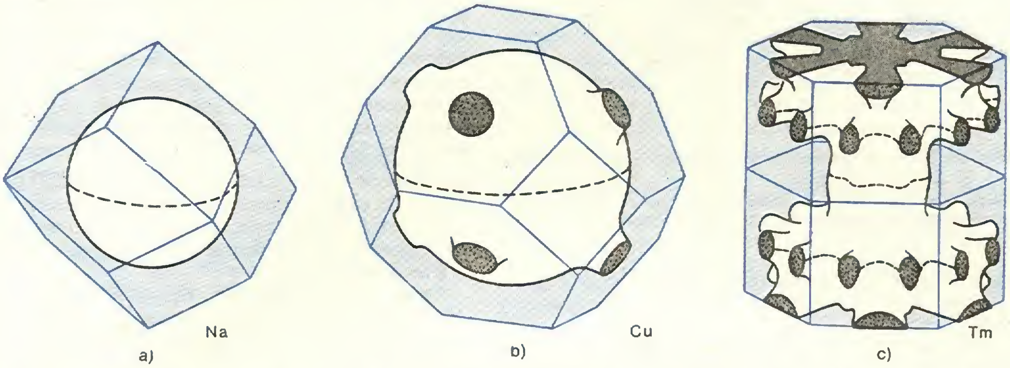
Rys. 15. Ewolucja kształtu powierzchni Fermiego. Przy małych koncentracjach elektronów przestrzeń pędu obsadzona przez elektrony leży z dala od granic strefy Brillouina, wpływ sieci kryształu na kształt powierzchni Fermiego jest zatem mały i powierzchnia pozostaje w przybliżeniu kulą (*a*). Przy wyższych koncentracjach powierzchnia Fermiego leży bliżej granic strefy, silnie ulega więc wpływowi sieci kryształu i znacznie odbiega kształtem od kuli (*b*)

anizotropia powierzchni Fermiego

sieci jest znacznie większy i prowadzi do wyraźnej anizotropii powierzchni Fermiego (*b* na rys. 15). Co więcej, w przypadku wielowartościowych metali mamy do czynienia z nakładającymi się na siebie pasmami energetycznymi (do czego jeszcze wrócimy) i skutkiem tego z kilkoma strefami Brillouina równocześnie. Jednakże ten wypadek ze względu na dodatkowe komplikacje, przekracza zakres niniejszego artykułu.

runku osi z. Obecność pola nie wpływa na składową ruchu wzdłuż z, natomiast w płaszczyźnie xy elektron wchodzi na orbitę, po której krąży z tzw. częstością cyklotronową. Jego ruch w płaszczyźnie poprzecznej do $\vec{H}$ jest zatem ruchem periodycznym. Jak wiadomo, wg zasad mechaniki kwantowej energia w ruchu periodycznym ulega kwantyzacji. Toteż energia kinetyczna elektronu związana z jego ruchem poprzecznym do

częstość cyklotronowa



Rys. 16. Przykłady powierzchni Fermiego: a) sód — sieć regularna przestrzennie centrowana, b) miedź — sieć regularna płasko centrowana, c) tul (ziemie rzadkie) — sieć heksagonalna o najgęstszym upakowaniu. Powierzchnie Fermiego pokazane są wewnątrz odpowiadających im stref Brillouina

powierzchnie Fermiego sodu, miedzi i tulu

Na rys. 16 pokazaliśmy przykłady powierzchni Fermiego kilku rzeczywistych metali. U góry przedstawiony jest przypadek sodu. Struktura krystaliczna sodu odpowiada sieci regularnej centrowanej przestrzennie. Powierzchnia Fermiego (wewnątrz której znajduje się dokładnie połowa stanów pasma, a więc połowa objętości strefy Brillouina) jest względnie odległa od granic strefy, toteż prawie zupełnie nie ulega zniekształceniu i zachowuje kształt kuli (według pomiarów doświadczalnych z dokładnością $^1/_{1000}$). Miedź krystalizuje w strukturze regularnej płasko centrowanej, co odpowiada strefie Brillouina pokazanej na rys. 13a. Podobnie jak sód, miedź ma jeden elektron walencyjny, a więc i tutaj powierzchnia Fermiego wypełnia dokładnie połowę objętości strefy. Jeżeli wpiszemy kulę o tej objętości w strefę Brillouina miedzi, zauważymy, że w niektórych punktach powierzchnia kuli (ze względu na specyficzną geometrię strefy) bardzo się zbliża do granic strefy. Zgodnie z wyżej opisanym rozumowaniem prowadzi to w miejscach zbliżenia do zniekształcenia pierwotnej kuli Fermiego. Powierzchnia Fermiego miedzi, skonstruowana na podstawie danych doświadczalnych i pokazująca odchylenia od sferyczności (tzw. szyjki) w miejscach największego zbliżenia do granic strefy Brillouina, pokazana jest na rys. 16b. Również na rys. 16 pokazaliśmy powierzchnię Fermiego dla bardziej egzotycznego przypadku tulu (dwuwartościowy metal z grupy ziem rzadkich, struktura heksagonalna), ilustrując nader złożoną strukturę, na jaką napotykamy w bardziej skomplikowanych (a więc wielowartościowych) materiałach.

Poziomy Landaua i oscylacje kwantowe

Powierzchnia Fermiego danego metalu stanowi zwarty (choć nieco abstrakcyjny) opis jego właściwości elektronowych, np. przewodnictwa elektrycznego, własności magnetycznych. Topologia powierzchni Fermiego jest więc obiektem intensywnych badań doświadczalnych współczesnej fizyki. Bardzo pomocne w ustalaniu szczegółów powierzchni Fermiego jest m.in. zjawisko de Haasa-van Alphena (i pokrewne mu zjawisko Szubnikowa-de Haasa), wynikające z zachowania się elektronów w silnym polu magnetycznym.

zjawisko de Haasa–van Alphena

Rozpatrzmy ruch swobodnego elektronu w obecności pola magnetycznego $\vec{H}$, przyłożonego np. w kie-

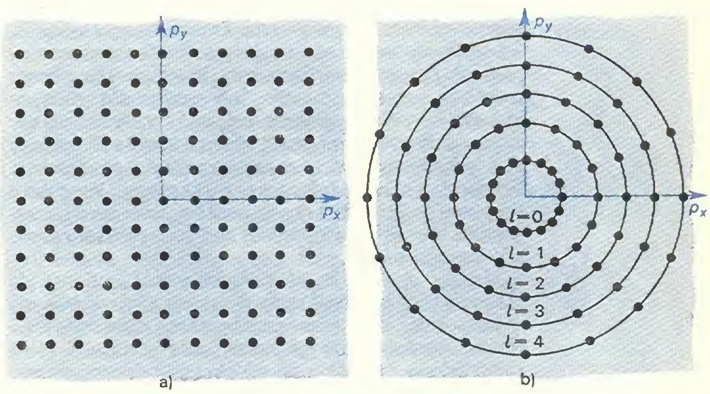
$\vec{H}$ (tzn. ze składowymi pędu p_x i p_y) może mieć tylko pewne ściśle określone, kwantowe wartości. Te dozwolone poziomy energii, nazywane poziomami Landaua, wyrażone są wzorem:

poziomy Landaua

$$E_l = (l+\tfrac{1}{2})\hbar\omega_c = \frac{p_x^2+p_y^2}{2m},$$

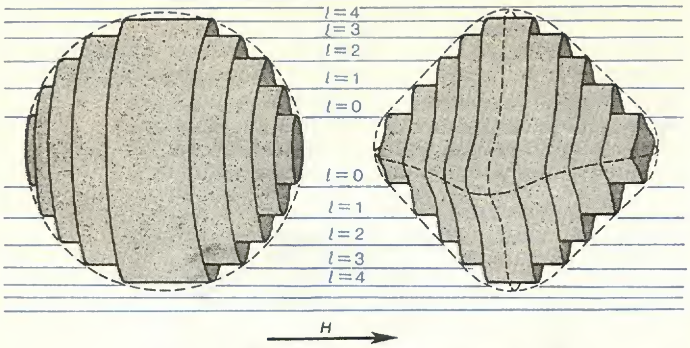
gdzie l jest dowolną liczbą całkowitą, ω_c — częstością cyklotronową, a m — masą efektywną elektronu; $\hbar = h/2\pi$.

Kwantyzacja Landaua wpływa drastycznie na rozkład stanów w przestrzeni pędu, oznacza bowiem, że jedynie niektóre składowe pędu w płaszczyźnie prostopadłej do pola są dozwolone. Pokazaliśmy to na rys. 17, przedstawiającym przekrój przestrzeni pędu w płaszczyźnie xy. Punkty na rys. 17a, oznaczające dozwolone stany w nieobecności pola, muszą się przemieścić i „osiąść" na stanach odpowiadających energiom E_l (rys. 17b). Stany obsadzone, które w nieobecności



Rys. 17. Wpływ pola magnetycznego na rozkład stanów w przestrzeni pędu

pola wypełniały przestrzeń wewnątrz powierzchni Fermiego jednorodną, gęstą siatką punktów, pod wpływem pola przemieszczają się (pozostając jednakże wewnątrz powierzchni Fermiego) na stany leżące na powierzchniach cylindrycznych („rurach") o stałym przekroju, z osią skierowaną wzdłuż pola magnetycznego $\vec{H}$, jak to widać na rys. 18. Obecność pola magnetycznego nie wprowadza ograniczeń wartości



Rys. 18. Rozkład elektronów w przestrzeni pędu w obecności pola $\vec{H}$ w wypadku sferycznej i anizotropowej powierzchni Fermiego. Elektrony mogą obsadzać jedynie stany zawarte wewnątrz powierzchni Fermiego (linia przerywana) i położone na rurach $l = 0, 1, 2, 3 \ldots$

pędu wzdłuż $\vec{H}$. Na rys. 18b przedstawiliśmy rozkład stanów, gdy powierzchnia Fermiego jest bardziej skomplikowana.

Przekrój „rury" stanów odpowiadającej poziomowi E_l ma pewną stałą wartość A_l, proporcjonalną do wielkości pola $\vec{H}$. Przy zwiększaniu pola, z chwilą gdy A_l przekroczy maksymalną wartość przekroju powierzchni Fermiego w płaszczyźnie prostopadłej do pola, elektrony obsadzające l-ty poziom Landaua „przelewają" się na poziomy niższe ($l=1$, $l=2$, ... itd.) w taki sposób, aby pozostać wewnątrz powierzchni Fermiego. Takiemu „przelaniu" się elektronów towarzyszą mierzalne zmiany w przewodnictwie, namagnesowaniu i innych zjawiskach elektronowych. W typowych polach magnetycznych rzędu jednej tesli mamy wiele tysięcy poziomów Landaua poniżej energii E_F, które w miarę zwiększania pola jeden po drugim „przerastają" powierzchnię Fermiego. Takie kolejne „przelewanie" się elektronów na coraz to niższe poziomy objawia się w bardzo niskich temperaturach w formie oscylacji zjawisk elektronowych jako funkcji pola, tzw. oscylacji kwantowych. Oscylacje kwantowe wartości namagnesowania nazywamy efektem de Haasa–van Alphena, zaś oscylacje przewodności elektrycznej — efektem Szubnikowa–de Haasa. Można pokazać, że oscylacje kwantowe danych wielkości są periodyczną funkcją $1/H$, a okres oscylacji jest proporcjonalny do maksymalnego przekroju powierzchni Fermiego w płaszczyźnie prostopadłej do $\vec{H}$. Zjawiska te nadają się idealnie do badania nieizotropowych powierzchni Fermiego. Przez pomiar okresu oscylacji namagnesowania lub prądu dla różnych kierunków pola względem kryształu możemy ustalić maksymalne wartości przekroju powierzchni Fermiego dla tych kierunków, a zatem topologię powierzchni Fermiego z wielką dokładnością. Oscylacje kwantowe stanowią więc jedną z najbardziej owocnych, najczęściej stosowanych metod badania podstawowych właściwości metali.

oscylacje kwantowe namagnesowania i przewodnictwa

Należy dodać, że do obserwowania efektu de Haasa–van Alphena i innych oscylacji kwantowych muszą być spełnione następujące dwa warunki eksperymentalne. Po pierwsze, czas relaksacji elektronu winien być dłuższy od okresu cyklotronowego ($\omega_c \tau > 1$), tak aby droga elektronu między zderzeniami równała się przynajmniej jednej orbicie cyklotronowej. Po drugie, różnica energii między poziomami Landaua powinna przekraczać energię termiczną elektronu ($\hbar\omega_c > kT$). Warunki te mogą być spełnione jedynie w bardzo czystych próbkach, w niskich temperaturach (około temperatury ciekłego helu) oraz w wysokich polach magnetycznych (rzędu 1 T i więcej).

Efekt de Haasa-van Alphena, a więc oscylacje kwantowe namagnesowania w funkcji pola magnetycznego, mierzy się przeważnie jednym z dwu sposobów. W pierwszym z nich wykorzystujemy fakt, że w obecności pola magnetycznego $\vec{H}$ próbka „odczuwa" moment siły proporcjonalny do jej namagnesowania $\vec{M}$ (tzn. do jej momentu magnetycznego), jeżeli $\vec{M}$ i $\vec{H}$ nie są równoległe. Sytuacja taka występuje, gdy powierzchnia Fermiego jest niesferyczna, a pole $\vec{H}$ skierowane jest dowolnie (tzn. nie wzdłuż osi symetrii). Jeżeli zawiesimy próbkę na cienkim a sztywnym pręciku, ulegnie ona skręceniu pod działaniem momentu siły wywieranym przez pole. Mierząc położenie kątowe tak zawieszonej próbki w funkcji pola otrzymujemy bezpośredni pomiar oscylacji kwantowych namagnesowania. Okres oscylacji daje nam, jak już wspomniano, wartość ekstremalną przekroju powierzchni Fermiego w płaszczyźnie prostopadłej do $\vec{H}$. Mierząc zatem zależność okresu oscylacji (a więc zależność ekstremalnego przekroju powierzchni Fermiego) od kierunku pola możemy w wielu wypadkach wnioskować o kształcie powierzchni Fermiego w trzech wymiarach.

pomiar oscylacji kwantowych namagnesowania

Druga metoda badania oscylacji de Haasa–van Alphena polega na pomiarze napięcia indukowanego zmiennym polem magnetycznym dH/dt w cewce indukcyjnej otaczającej próbkę metalu. Można pokazać, że napięcie to będzie proporcjonalne do $dM/dt = (dM/dH)(dH/dt)$, gdzie M jest namagnesowaniem próbki. Jeżeli namagnesowanie M oscyluje w funkcji pola H, to przy równomiernej zmianie pola dH/dt oscylacje te obserwować będziemy bezpośrednio jako oscylacje w indukowanym napięciu cewki w funkcji czasu. Metoda indukcyjna nadaje się szczególnie do zastosowania w wypadku magnesów impulsowych, dzięki którym otrzymujemy wzrost pola od zera do kilkudziesięciu tesli w przeciągu kilku tysięcznych sekundy. Oscylujący sygnał indukowany w cewce można obserwować bezpośrednio na ekranie oscyloskopu.

oscylacje kwantowe w zjawiskach transportu

Wspominaliśmy już, że oscylacje kwantowe występują również w zjawiskach transportu w metalach (tzw. zjawisko Szubnikowa–de Haasa), które możemy badać np. przez pomiar oporu próbki w funkcji pola magnetycznego. Należy dodać, że oscylacje kwantowe występują właściwie w każdym zjawisku, w którym biorą udział elektrony przewodnictwa. Możemy je zatem również obserwować w takich zjawiskach elektronowych, jak rozchodzenie i tłumienie ultradźwięku w metalach, siła termoelektryczna, przewodnictwo cieplne, tłumienie fal o częstościach radiowych (tzw. fal helikonowych), magnetostrykcja (zmiany objętości próbki pod wpływem pola) i wiele innych.

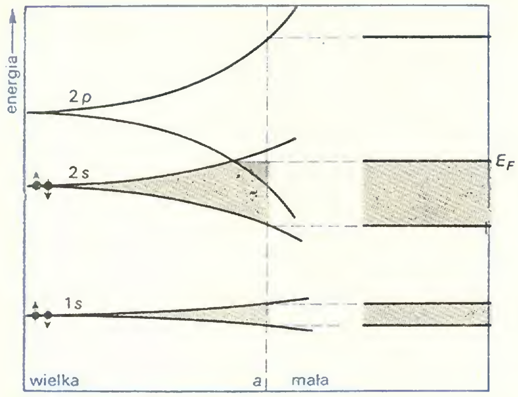
Metale jedno- i wielowartościowe

Podstawową cechą charakterystyczną metalu, odróżniającą go od innych materiałów, jest posiadanie pasm energetycznych, w których tylko część stanów jest obsadzona przez elektrony nawet przy $T = 0$ K (tzn. w stanie podstawowym metalu). W krysztale składającym się z N komórek elementarnych (co w większości metali odpowiada po prostu N atomom), na każde pasmo przypada N różniących się od siebie, oddzielnych stanów. Każde pasmo może zatem pomieścić, zgodnie z zakazem Pauliego, $2N$ elektronów.

Widać z tego, że pierwiastki jednowartościowe o prostych strukturach krystalicznych, wnoszące do pasma przewodnictwa tylko po jednym elektronie na komórkę elementarną (jak np. sód lub miedź), wypełniają pasmo dokładnie do połowy i muszą być metalami.

Czym jednak wytłumaczyć, że pierwiastki dwuwartościowe (np. beryl, cynk, żelazo) lub nawet niektóre czterowartościowe (cyna, ołów) są również metalami?

beryl Rozpatrzmy przykład berylu, którego atom ma 2 elektrony w stanie $2s$, a więc pełną powłokę, co powinno odpowiadać całkowicie obsadzonemu pasmu w krysztale. Wydawałoby się więc, że beryl powinien być izolatorem. Gdy jednak zbliżamy atomy berylu do siebie, tworząc kryształ (jak na rys. 19), oddziaływanie



Rys. 19. Nakładanie się pasm w wypadku berylu (a odpowiada stałej sieci kryształu rzeczywistego)

między atomami rozszczepia całkowicie obsadzoną powłokę elektronową $2s$ oraz nieobsadzoną powłokę $2p$ tak dalece, że odpowiadające im pasma nakładają się na siebie. Ponieważ układ jako całość dąży do stanu o najniższej energii, elektrony z najwyżej położonych poziomów pasma $2s$ „przelewają" się do niżej od nich leżących poziomów w pasmie $2p$. W wyniku tego mamy dwa niecałkowicie obsadzone pasma. Obydwa zatem pasma biorą udział w przewodnictwie elektrycznym i innych zjawiskach elektronowych. Jak widać z rys. 19, stopień nałożenia się pasm jest określony przez wartość stałej sieci: gdyby odległość między atomami kryształu była nieco większa, nałożenie się pasm nie nastąpiłoby i beryl byłby półprzewodnikiem.

wpływ ciśnienia Mała różnica w stałej sieci może zatem spowodować ogromne zmiany w fizycznych właściwościach materiału. Warto wspomnieć, że przy zastosowaniu wysokich ciśnień do niektórych niemetali powodujemy zmiany w strukturze krystalicznej i w odległościach międzyatomowych, w wyniku których możemy otrzymać nakładanie się pasm i materiał staje się wtedy metalem. Pod bardzo wysokim ciśnieniem obserwowano przejścia w fazę metaliczną takich pierwiastków jak german, krzem i nawet (pod ciśnieniem powyżej 10^5 MPa) diament!

Nasuwa się również pytanie, dlaczego nie wszystkie pierwiastki o nieparzystej wartościowości są metalami. Mogłoby się wydawać, że w przypadku takich pierwiastków pasma odpowiadające zewnętrznym powłokom atomów są z konieczności tylko częściowo obsadzone. Rozważmy to na przykładzie wodoru.

wodór Wodór jest pierwiastkiem jednowartościowym, w stanie stałym (wodór stały otrzymujemy pod wysokim ciśnieniem w niskich temperaturach) jest izolatorem. Wynika to stąd, że wiązanie atomów wodoru w cząsteczce H_2 jest bardzo silne i w stanie stałym wodór jest siecią cząsteczek H_2, a nie atomów. Wodór tworzy zatem tzw. kryształ molekularny, w którym (w odróżnieniu od prostych układów krystalicznych) na komórkę elementarną przypada cząsteczka, a więc dwa elektrony. W krysztale wodoru mamy więc dokładnie tyle elektronów ($2N$), ile mieści całkowicie obsadzone pasmo energetyczne. Teoria jednak przewiduje, że pod niezwykle wysokim ciśnieniem (ponad $2 \cdot 10^5$ MPa) wodór stały zmienia swą strukturę krystaliczną i rzeczywiście staje się metalem. Doświadczenia nad otrzymaniem fazy metalicznej wodoru są obecnie w toku. Informacje na ten temat znajdzie Czytelnik w artykule „Wysokie ciśnienia".

Optyczne właściwości metali

Metale zwracają na siebie uwagę przez swój piękny, metaliczny połysk. Optyczne właściwości metalu podobnie jak i inne ich cechy charakterystyczne, wynikają również w głównej mierze z istnienia niezupełnie obsadzonych pasm energetycznych (a więc z istnienia gazu elektronowego). Rozważmy falę świetlną o częstości ω padającą na gaz swobodnych elektronów. Fala poprzez swoje pole elektryczne zmusza elektrony do drgań o tej samej częstości ω, co z kolei daje zmienną w czasie polaryzację ośrodka. Faza, z którą swobodne elektrony oscylują pod wpływem fali, jest przy tym taka, że wynikająca z ruchu elektronów polaryzacja jest ujemna (tzn. odwrotna w fazie do elektrycznego pola fali). Jak się okaże, z tego właśnie faktu wynikają główne cechy optyczne metali. **polaryzacja ośrodka wywołana przez falę świetlną**

Polaryzacja ośrodka może być formalnie wyrażona przez tzw. efektywną przenikalność elektryczną (przenikalność względna) $\varkappa$, która z kolei stanowi punkt wyjściowy przy omówieniu właściwości optycznych materiałów. Dla metalu $\varkappa$ można wyrazić z dobrym przybliżeniem jako:

$$\varkappa = 1-(ne^2/m\omega^2\varepsilon_0) = 1-(\omega_p^2/\omega^2),$$

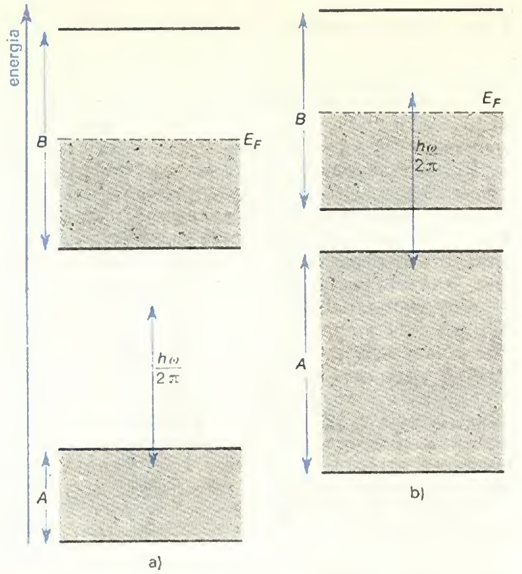
gdzie n jest koncentracją elektronów na 1 cm^3, e — ładunkiem, m — masą elektronu, ε_0 — przenikalnością próżni, a ω_p — tzw. częstością plazmową. Człon $ne^2/m\omega^2\varepsilon_0$ przedstawia wspomniany wyżej ujemny przyczynek swobodnych elektronów do polaryzacji ośrodka. Ze względu na dużą koncentrację elektronów w metalach, wartość tego członu w obszarze widzialnym, podczerwonym i w niższych częstościach jest bardzo duża w porównaniu z jednością, a więc przenikalność elektryczna $\varkappa$ dla tych obszarów częstości ma również wartość ujemną. **ujemna przenikalność elektryczna**

Z równań Maxwella wiemy, że w ośrodku o ujemnej przenikalności elektrycznej fale elektromagnetyczne nie mogą się rozchodzić. W tych warunkach fale świetlne nie mogąc przeniknąć do wnętrza ośrodka, zostają całkowicie odbite, niezależnie od częstości. To tłumaczy kolor, połysk i brak przezroczystości „białych" metali, takich jak srebro lub aluminium. Gładkie powierzchnie tych metali, odbijając światło w całym obszarze widzialnym, stanowią idealne lustra. Patrząc na błyszczący metal, możemy więc powiedzieć że w zasadzie „widzimy" swobodne elektrony. **całkowite odbicie światła**

W powyższym opisie pominęliśmy wpływ absorpcji światła. Absorpcja światła przez gaz elektronowy spowodowana jest rozpraszaniem elektronów i wyraża się przez poprawkę do $\varkappa$ rzędu ν/ω, gdzie ν jest średnią częstością zderzeń. Dla typowych wartości ν (10^9–10^{14} Hz) i dla częstości optycznych ($5 \cdot 10^{15}$ Hz) poprawka ta jest tak mała, że możemy ją pominąć. **absorpcja światła**

W naszych rozważaniach nie uwzględniliśmy również istnienia niżej położonych, całkowicie obsadzonych pasm energetycznych, takich jak np. pasmo $2p$ na rys. 8. W większości metali pasma te odpowiadają tak niskim energiom w stosunku do pasma elektronów swobodnych, że w zakresie widzialnym energia kwantu światła (fotonu) $h\omega/2\pi$ jest za mała, aby spowodować przejścia elektronowe z tych głęboko położonych pasm (A na rys. 20a) do poziomów nieobsadzonych (pasmo B). Fale świetlne nie wywołują więc żadnych zmian w takich całkowicie obsadzonych pasmach. W rezultacie pasma te nie wnoszą nic nowego do naszego obrazu właściwości optycznych metali.

Jednakże w wypadku niektórych metali (np. złota lub miedzi) sprawa przedstawia się inaczej. Energia dzieląca pasmo przewodnictwa od niższych pasm jest w tych metalach względnie mała, mniej więcej tego samego rzędu co energia kwantu $h\omega/2\pi$ dla krótszych fal widzialnego widma. Zatem fale poniżej pewnej długości mogą spowodować przejścia między pasmami, tak jak na rys. 20b. W złocie i miedzi energie



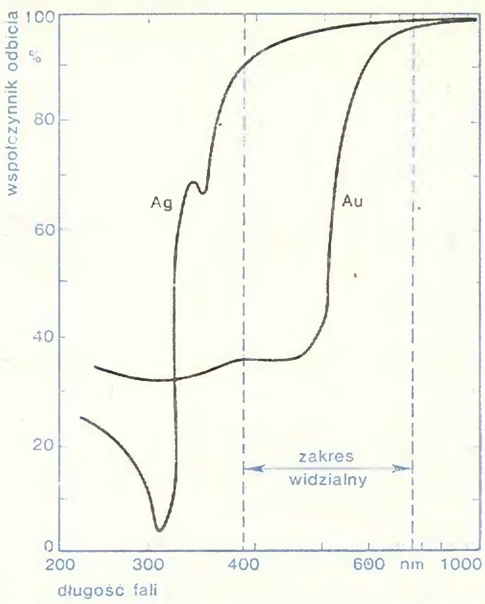
Rys. 20. Mechanizm absorpcji przez przejścia międzypasmowe: a) energia fotonu jest zbyt mała, aby wywołać przejścia z pasma A do stanów nieobsadzonych w pasmie B, b) różnica energii między pasmami jest mniejsza, co umożliwia przejścia międzypasmowe i związaną z nimi absorpcję w obszarze widzialnym

barwa złota i miedzi

powodujące takie przejścia odpowiadają falom krótszym od 500 nm, a więc obszarowi niebieskiemu i fioletowemu widma. Fale z tego obszaru widma będą więc pochłaniane i częściowo usunięte z odbitego przez metal światła. W widmie odbitego światła przeważać zatem będą fale dłuższe, a więc światło czerwone, pomarańczowe i żółte, co tłumaczy ciepły kolor tych metali. Na rys. 21 porównujemy współczynniki odbicia dla srebra i złota, odróżniające metal kolorowy i biały.

Pokazaliśmy poprzednio, że przenikalność elektryczna metali $\varkappa$ ma wartość ujemną ze względu na wysoką wartość członu $ne^2/m\omega^2\varepsilon_0 = \omega_p^2/\omega^2$. Człon ten zmniejsza się jednak z wzrastaniem częstości i dla $\omega = \omega_p$ w obszarze nadfioletu osiąga wartość 1. Dla częstości wyższych od ω_p przenikalność elek-

dodatnia przenikalność elektryczna



Rys. 21. Zależność współczynnika odbicia od długości fali dla srebra i dla złota

tryczna $\varkappa$ staje się więc dodatnia. Toteż w obszarze bardzo krótkich fal (nadfiolet i krótsze) metal staje się przeźroczysty i zachowuje się nie inaczej niż diament lub kwarc w widzialnym zakresie widma.

M. YA. AZBEL, M. I. KAGANOV, I. M. LIFSHITZ *Conduction Electrons in Metals*, Sc. Am. 228, 88 (1973); A. H. COTTRELL *The Nature of Metals*. Sc. Am. 217, 90 (1967); L. KALINOWSKI *Fizyka metali*, Warszawa 1970; A. R. MACKINTOSH *The Fermi Surface of Metals*, Sc.Am. 209, 110 (1963).

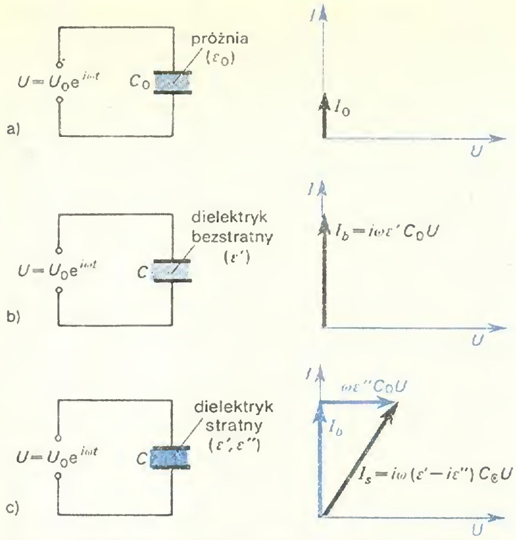
Dielektryki

Bożena Hilczer

gromadzenie ładunku

Dielektrykami (izolatorami) przyjęto nazywać materiały, które w danych warunkach termodynamicznych (określonej temperaturze i ciśnieniu) mają przerwę energetyczną (→ Struktura elektronowa ciał stałych) większą od 3 eV, stałoprądową przewodność elektryczną w polach poniżej 10^7 V/m mniejszą od 10^{-6} Ω^{-1}m^{-1} oraz tangens kąta strat w zakresie częstości od 50 Hz do 10^6 Hz mniejszy od 0,5. Są to kryteria formalne. Najważniejszą własnością fizyczną dielektryków jest zdolność do gromadzenia ładunku elektrycznego. Własność ta jest znana już bardzo dawno. Wyniki eksperymentów z ciałami naładowanymi elektrycznie pozwoliły ustalić prawidłowości powszechnie znane jako prawo Coulomba. Stała materiałowa charakteryzująca ośrodek ze względu na oddziaływanie ładunków zwana jest przenikalnością elektryczną ε. Z elektrostatyki wiemy również, że umieszczenie dielektryka w jednorodnym polu elektrycznym powoduje zmianę gęstości linii sił pola elektrycznego, zależną od wielkości przenikalności elektrycznej dielektryka. Najprostszym przykładem jednorodnego pola elektrycznego jest pole panujące w kondensatorze płaskim (o powierzchniach okładek a odległych o d) umieszczonym w próżni, do którego przyłożono napięcie U. Na okładkach kondensatora zgromadzony jest ładunek Q_0, linie sił pola elektrycznego są prostopadłe do okładek (gdy pominiemy efekty brzegowe), a natężenie pola elektrycznego $E = U/d$. Znając gęstość D strumienia linii sił pola elektrycznego (tzw. przesunięcie elektryczne) oraz natężenie E pola elektrycznego możemy wyznaczyć przenikalność elektryczną próżni jako stosunek przesunięcia elektrycznego w próżni do natężenia pola elektrycznego: $\varepsilon_0 = D/E = C_0 d/a$. Wartość przenikalności elektrycznej próżni można doświadczalnie wyznaczyć mierząc za pomocą galwanometru balistycznego ładunek Q_0; wynosi ona $\varepsilon_0 = 8{,}85\cdot 10^{-12}$ F/m. Gdy między okładkami kondensatora płaskiego mającego w próżni pojemność C_0 umieścimy dielektryk, to pojemność jego wzrośnie do wartości $C = \varepsilon C_0$. Liczbę określającą ile razy pojemność kondensatora zawierającego dielektryk jest większa od pojemności takiego samego kondensatora próżniowego nazywamy przenikalnością elektryczną ε. Jeżeli do kondensatora z dielektrykiem przyłożymy przemienne napięcie $U = U_0 e^{i\omega t}$, to w obwodzie popłynie prąd, który w wypadku dielektryka idealnego (bezstratnego) będzie wyprzedzał napięcie w fazie o $\pi/2$ (rys. 1b). Ogólnie, w dielektryku realnym (stratnym) oprócz prądu przesunięcia popłynie prąd przewodzenia o fazie zgodnej z przyłożonym napięciem (rys. 1c) i przenikalność elektryczną wyraża się jako wielkość zespoloną $\varepsilon^* = \varepsilon' - i\varepsilon''$, gdzie ε' oznacza rzeczywistą składową przenikalności elektrycznej,

przenikalność elektryczna



Rys. 1. Relacje fazowe między natężeniem prądu I i napięciem U: a) kondensator próżniowy, b) kondensator z dielektrykiem bezstratnym, c) kondensator z dielektrykiem stratnym

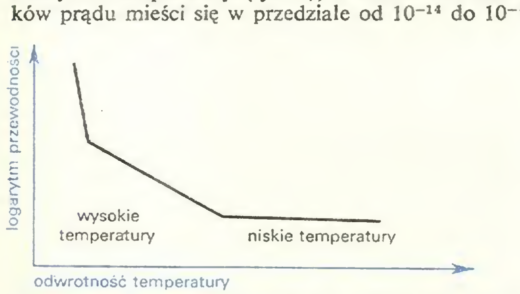
natomiast ε'' — składową urojoną, charakteryzującą straty dielektryczne. Straty dielektryczne przyjęto określać wielkością tg δ, czyli stosunku natężenia prądu zgodnego w fazie z napięciem do natężenia prądu przesunięcia: $\text{tg}\,\delta = \omega\varepsilon'' C_0 U/\omega\varepsilon' C_0 U = \varepsilon''/\varepsilon'$. Często do opisu własności dielektryków stratnych stosuje się również wielkość G, zwaną przewodnością elektryczną, określoną jako iloczyn częstości kołowej i urojonej części przenikalności elektrycznej: $G = \omega\varepsilon_0\varepsilon''$. Wielkość ta opisuje straty energii związane z różnymi procesami zachodzącymi w dielektryku.

przewodność elektryczna

Przewodnictwo elektryczne

Prąd płynący po przyłożeniu do dielektryka stałego pola elektrycznego zanika w czasie. Jest to konsekwencja czasowej zmiany przewodności elektrycznej dielektryka, określonej stosunkiem gęstości prądu do natężenia pola elektrycznego. Od wartości przewodności elektrycznej σ w stanie ustalonym (gdy czasowe zmiany są mniejsze niż kilka procent na godzinę) zależą własności elektryczne dielektryków. Jak wspomnieliśmy na wstępie, przewodność elektryczna dielektryków jest mała; w dobrych dielektrykach w warunkach normalnych jest mniejsza nawet od $10^{-18}\Omega^{-1}\text{m}^{-1}$. Przy niskich temperaturach przewodność elektryczna dielektryków bardzo słabo zależy od temperatury (rys. 2), a ruchliwość nośników prądu mieści się w przedziale od 10^{-14} do 10^{-4}

zależność od temperatury

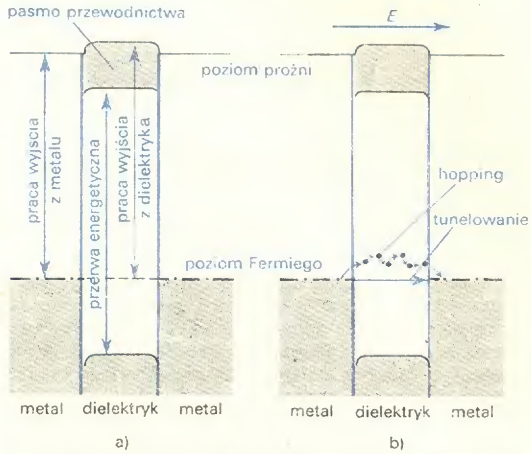


Rys. 2. Schematyczne przedstawienie temperaturowej zależności przewodnictwa elektrycznego dielektryków

m^2/Vs. Ze wzrostem temperatury ruchliwość nośników prądu rośnie, wiemy jednak, że rośnie również liczba nośników prądu (→ Półprzewodniki). Ponieważ przewodność elektryczna zależy zarówno od liczby nośników jak i od ich ruchliwości, a obie te wielkości zależą od temperatury w sposób aktywacyjny, to temperaturową zależność przewodności elektrycznej dielektryków σ opisuje dobrze prawo Arrheniusa (rys. 2): $\sigma = \sigma_0 e^{-W/kT}$, gdzie W oznacza energię aktywacji.

prawo Arrheniusa

Zastanówmy się teraz, jakie są możliwe sposoby transportu elektrycznego w dielektrykach. Aby zmierzyć przewodność elektryczną dielektryka, musimy nałożyć na niego elektrody metalowe i przyłożyć zewnętrzne pole elektryczne. Diagram struktury pasmowej dielektryka z nałożonymi elektrodami jest przedstawiony schematycznie na rys. 3a, a zagięcie pasm w pobliżu zetknięcia powierzchni metalu i dielektryka zależy od pracy wyjścia obydwu materiałów.



Rys. 3. Struktura pasmowa układu metal–dielektryk–metal (a) i schematyczne przedstawienie tunelowania oraz przewodnictwa hoppingowego (b)

W niezbyt silnych polach elektrycznych (mniejszych niż 10^7 V/m) jednym z możliwych mechanizmów transportu ładunku w takim układzie jest ruch elektronu lub dziury w pasmach dozwolonych. Poruszający się w dielektryku elektron (lub dziura) wywołuje jednak lokalną polaryzację sieci, przemieszczającą się wraz z nim. Obiekt taki, zachowujący się zupełnie inaczej niż swobodny elektron, nazywamy polaronem. Drugim możliwym mechanizmem jest transport elektronów (lub dziur) przez pasmo wzbronione. W idealnym dielektryku jest to możliwe tylko w wyniku kwantowego procesu tunelowania w bardzo cienkich kryształach (o grubościach mniejszych niż 3 nm). W rzeczywistych dielektrykach defekty sieci krystalicznej powodują powstanie pasmie wzbronionym lokalnych, dozwolonych poziomów energetycznych, zwanych poziomami pułapkowymi. Jeżeli gęstość lokalnych poziomów pułapkowych jest duża i odległości między nimi są małe (mniejsze niż 3 nm), to mogą wystąpić zjawiska kwantowe i elektrony w przyłożonym polu elektrycznym będą się poruszać skacząc od jednej pułapki do drugiej (rys. 3b). Przewodnictwo związane z tym zjawiskiem nazywa się przewodnictwem hoppingowym. Przewodnictwo elektryczne dielektryków może mieć również charakter jonowy. W tym wypadku transport ładunku w polu elektrycznym jest związany z ruchem jonów między położeniami międzywęzłowymi lub lukami w węzłach sieci, podobnie jak w kryształach jonowych. Ogólnie w dielektrykach można wyróżnić kilka możliwych mechanizmów transportu elektrycznego. Przewodnictwo elektryczne może być związane z ruchem elektronów (lub dziur) w obrębie dozwolonych pasm energetycznych i proces ten opisuje formalizm ruchu polaronów z ruchem elektronów (lub dziur) przez pasmo wzbronione, co prowadzi do procesu tunelowania lub hoppingu, oraz z ruchem ujemnie lub dodatnio naładowanych jonów między defektami punktowymi sieci.

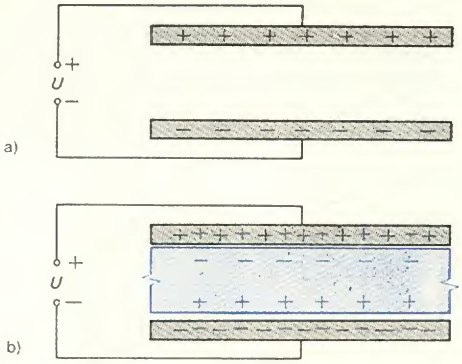
polaron

przewodnictwo hoppingowe

przewodnictwo jonowe

Polaryzacja dielektryczna

Przyłożenie stałego napięcia U do płaskiego kondensatora próżniowego spowoduje naładowanie go do pojemności C_0, takiej, aby swobodne ładunki na każdej okładce wytworzyły różnicę potencjałów równą co do wielkości U, lecz o przeciwnej polarności



Rys. 4. Wpływ dielektryka na gęstość ładunku na okładkach kondensatora

(rys. 4a). Dielektryk o przenikalności elektrycznej ε umieszczony między okładkami takiego kondensatora zwiększy jego pojemność, gdyż na okładki kondensatora musi dopłynąć ze źródła ładunek kompensujący ładunek polaryzujący dielektryka (rys. 4b). Efekt polaryzacji związany jest z orientacją dipoli elektrycznych (indukowanych i trwałych) pod wpływem przyłożonego pola elektrycznego. Powstała polaryzacja dielektryczna P jest określona wielkością momentu dipolowego jednostki objętości dielektryka. Z rys. 4b widać, że polaryzację dielektryka można określić również jednoznacznie gęstością ładunków na powierzchni dielektryka i definicje te są używane zamiennie. Gęstość strumienia linii sił pola elektrycznego jest wtedy określona związkiem $D = \varepsilon_0 E + P$, a ponieważ $D = \varepsilon_0 \varepsilon' E$, wyrażenie wiążące polaryzację z natężeniem pola elektrycznego ma postać $P = \varepsilon_0(\varepsilon' - 1)E$. Wielkość $(\varepsilon' - 1)$ nazywamy również podatnością dielektryczną ośrodka; określa ona stosunek gęstości ładunku związanego do gęstości ładunku swobodnego.

podatność dielektryczna

Przy wprowadzaniu wielkości ε', ε'', G oraz P, traktowaliśmy dielektryk jako ciało izotropowe. Kryształy dielektryczne są jednak ciałami wykazującymi anizotropię własności fizycznych, przy czym wielkości te są zależne od kierunku w krysztale. Dokładny opis własności dielektrycznych, uwzględniający ich anizotropowy charakter, jest możliwy jedynie za pomocą rachunku tensorowego.

molekularny mechanizm polaryzacji

Dotychczas nie wnikaliśmy w molekularny mechanizm powstawania polaryzacji dielektrycznej, ani nie wyróżnialiśmy rodzaju momentów dipolowych odpowiedzialnych za polaryzację dielektryka, związaliśmy jedynie polaryzację dielektryczną z wielkością natężenia pola elektrycznego. Obecnie omówimy molekularne mechanizmy polaryzacji dielektrycznej. Na wstępie musimy zdać sobie sprawę, że wewnętrzne (lokalne) pole elektryczne F panujące w dielektryku różni się dość istotnie od przyłożonego zewnętrznego pola E. Zagadnienie pola wewnętrznego jest ciągle jednym z głównych problemów teorii dielektryków i w ogólnym wypadku nie jest rozwiązane. R. Clausius związał polaryzację dielektryczną z polem wewnętrznym F równaniem:

$$P = (\varepsilon' - 1)\varepsilon_0 E = N\alpha F,$$

gdzie α — polaryzowalność, czyli zdolność do tworzenia dipoli przez atomy lub molekuły ośrodka, natomiast N jest liczbą elementarnych dipoli w jednostce objętości. Pole lokalne w dielektrykach niedipolowych i dielektrykach z całkowicie nieuporządkowanymi momentami dipolowymi lub dipolami uporządkowanymi w sieć o najwyższej symetrii, można opisać w sposób przybliżony wzorem Lorentza:

wzór Lorentza

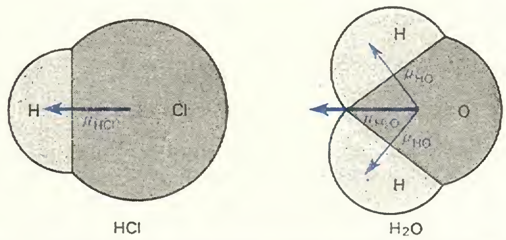
$$F = E + \frac{P}{3\varepsilon_0} = \frac{\varepsilon' + 2}{3} E \quad \text{(pole Lorentza).}$$

Inne wyrażenia na pole wewnętrzne podali L. Onsager (w ujęciu molekularnym) oraz H. Fröhlich i J. G. Kirkwood (na podstawie rozważań statystycznych). Wyrażenia te są bardzo skomplikowane i na ogół zastępuje się je wyrażeniami otrzymanymi w sposób półempiryczny.

Ogólnie polaryzacja jednorodnego dielektryka składa się z polaryzacji dipolowej P_d (wywołanej orientacją trwałych molekularnych dipoli w polu elektrycznym) oraz z polaryzacji deformacyjnej (związanej z indukowaną przez pole elektryczne polaryzacją atomową P_a i polaryzacją elektronową P_e). Istnieją oczywiście materiały dielektryczne nie mające trwałych molekularnych momentów dipolowych i wtedy ich polaryzacja dielektryczna jest tylko polaryzacją deformacyjną. Można wyróżnić również dielektryki, w których występuje jedynie polaryzacja elektronowa.

trwałe momenty dipolowe

Trwałe momenty dipolowe mają molekuły o niesymetrycznym rozkładzie ładunku (rys. 5). Moment dipolowy dużych molekuł organicznych związany jest



Rys. 5. Trwałe momenty dipolowe związane z asymetrią rozkładu ładunku w molekułach

na ogół z takimi grupami dipolowymi jak grupa hydroksylowa O—H lub grupa ketonowa C=O. Całkowity moment dipolowy molekuły jest wtedy w dobrym przybliżeniu równy wektorowej sumie wszystkich składowych momentów pochodzących od grup dipolowych.

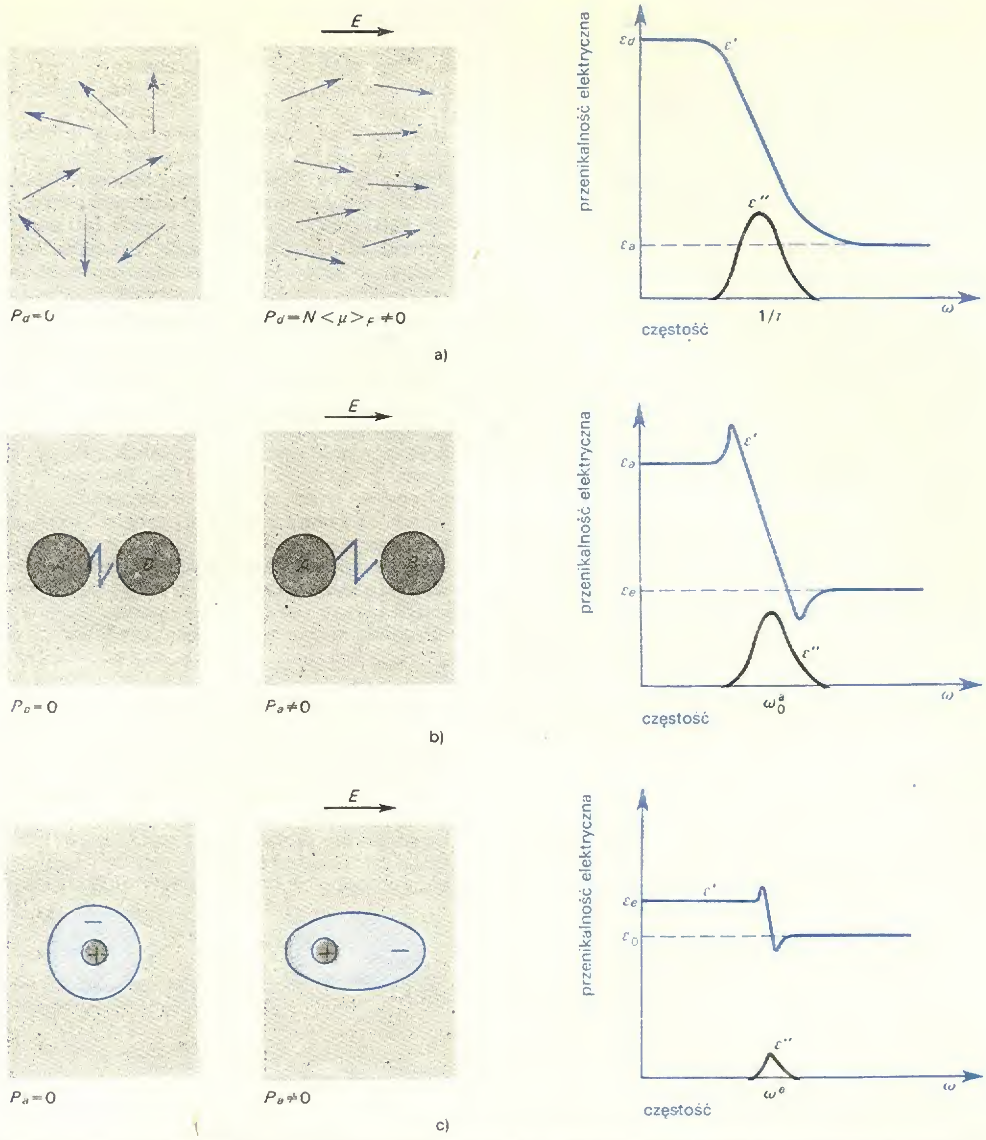
polaryzacja orientacyjna

Gdy w dielektryku występują trwałe dipole molekularne μ, to w zewnętrznym polu elektrycznym orientują się one w kierunku pola i polaryzacja dipolowa, zwana również polaryzacją orientacyjną, jest określona średnim rzutem momentów dipolowych na kierunek pola wewnętrznego w dielektryku (rys. 6a). W przemiennym polu elektrycznym, w najprostszym wypadku, gdy w dielektryku występuje tylko jeden rodzaj trwałych dipoli molekularnych, zespoloną przenikalność elektryczną można przedstawić wzorem:

$$\varepsilon^* = \varepsilon_a + \int_0^\infty \beta(t) e^{i\omega t} dt,$$

gdzie ε_a oznacza wartość przenikalności elektrycznej przy wysokich częstościach, natomiast $\beta(t)$ ma charakter współczynnika zanikania, określającego opóźnienie zmian polaryzacji względem zmian pola elektrycznego. W roku 1912 P. Debye zaproponował wykładniczą formę współczynnika zaniku $\beta(t) = \beta(0)e^{t/\tau}$, z czasem relaksacji τ, charakteryzującym dielektryk dipolowy i zależnym od temperatury w sposób aktywacyjny. Prowadzi to do dyspersji przenikalności elektrycznej przedstawionej na rys. 6a. Debye'owski model dielektryka, w którym występują dipole molekularne jednego rodzaju, mające jeden czas relaksacji, jest modelem uproszczonym. W rzeczywistych dielektrykach występuje na ogół rozszerzenie obszaru relaksacji wywołane zarówno różnicami naturalnej częstości drgań dipoli związane

polaryzacja dielektryka jednorodnego



Rys. 6. Schematyczne przedstawienie polaryzacji orientacyjnej oraz dyspersja i absorpcja debyeowska (a), polaryzacji atomowej (b) i elektronowej (c) oraz związana z tymi zjawiskami anomalna dyspersja i absorpcja

z różnicami w rozmiarach i kształtach trwałych dipoli molekularnych jak i różnicami energii aktywacji dipoli, a więc różną konfiguracją najbliższego sąsiedztwa takich samych dipoli molekularnych. Opis matematyczny tego zjawiska jest bardziej skomplikowany, gdyż trzeba wprowadzić funkcje rozkładu czasów relaksacji. Wprowadza się je na ogół w postaci podanej przez Cole–Cole, Fröhlicha lub Fuossa–Kirkwooda. Ogólna zależność przenikalności elektrycznej od częstości dla układów, w których występuje rozpraszanie energii, została podana przez Kramersa i Kroniga. Dyspersja przenikalności elektrycznej związana z relaksacją polaryzacji dipolowej występuje w zakresie częstości mikrofalowych.

polaryzacja deformacyjna

Wróćmy teraz do polaryzacji deformacyjnej i najpierw rozpatrzmy polaryzację atomową i jonową. Dla uproszczenia rozważmy molekułę złożoną z dwu atomów (lub jonów) *A* oraz *B* (rys. 6b), między którymi występuje oddziaływanie. Przyłożenie pola elektrycznego spowoduje rozsunięcie rdzeni atomów (jonów), zwiększając odległość między nimi, co powoduje powstanie momentu dipolowego (lub jeżeli molekuła miała bez pola moment dipolowy — jego zwiększenie). Prowadzi to do polaryzacji ośrodka. Wiemy jednak, że atomy można traktować również jako obiekty składające się z naładowanego dodatnio rdzenia, otoczonego chmurami elektronów (rys. 6c), których symetria jest określona ich stanami kwantowymi (→ Chemia kwantowa, rozdz. „Atomy wieloelektronowe"). Przyłożenie pola elektrycznego powoduje przesunięcie chmury elektronów względem rdzenia, co prowadzi również do powstania indukowanego momentu dipolowego i w konsekwencji do indukowania polaryzacji elektronowej.

model oscylatora harmonicznego

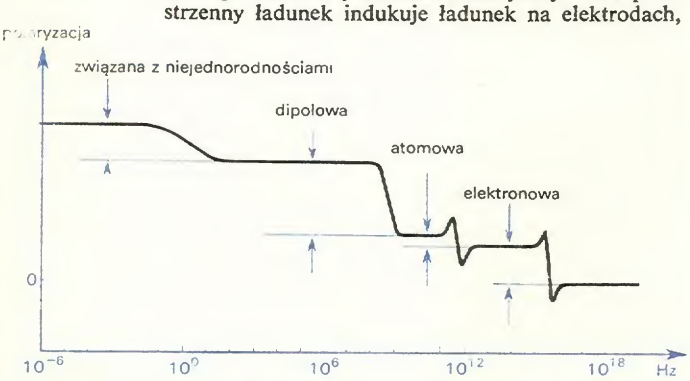
Do polaryzacji atomowej, jonowej, a także elektronowej można zastosować model oscylatora harmonicznego. Przesunięcie przez pole elektryczne ładunków przeciwnych znaków, związanych ze sobą sprężyście, powoduje polaryzację ośrodka. Po usunięciu zaburzenia wywołanego polem ładunki wracają do swoich pierwotnych położeń równowagi wykonując drgania, które zanikają z szybkością określoną tłumieniem. Gdy pole wywołujące polaryzację deformacyjną jest polem przemiennym, układ złożony z takich oscylatorów może zaabsorbować energię przy pewnej charakterystycznej częstości ω_0. Zjawisko to jest zupełnie analogiczne do absorpcji rezonansowej obwodu elektrycznego zawierającego opór omowy, pojemność oraz indukcyjność. Do opisu zależności zespolonej przenikalności elektrycznej od częstości

pola elektrycznego, gdy zachodzi absorpcja rezonansowa, H. Fröhlich wprowadził czynnik (analogiczny do współczynnika Debye'a) określający zanikanie polaryzacji: $\beta(t) = \beta(0)e^{-t/\tau}\cos(\omega_0 t + \varphi)$, gdzie φ oznacza opóźnienie fazowe polaryzacji względem pola elektrycznego. Na rys. 6b i 6c przedstawiono zależność ε' oraz ε'' od częstości, w wypadku polaryzacji atomowej i elektronowej. Rezonansowa absorpcja i dyspersja związana z drganiami rdzeni atomowych w molekułach leży w obszarze podczerwieni, a związana z drganiami dipola wytworzonego przesunięciem chmury elektronowej względem dodatniego rdzenia — w obszarze częstości optycznych. W zakresie częstości optycznych zamiast zespolonej przenikalności elektrycznej wprowadza się zwykle zespolony współczynnik załamania n^*, przy czym zespolona przenikalność elektryczna $\varepsilon_e^* = (n^*)^2$ (zależność Maxwella). Polaryzacja deformacyjna może być wywołana także polem elektrycznym fali świetlnej. Zjawisko to odgrywa istotną rolę, gdy dielektryk jest poddany działaniu silnych wiązek światła laserowego.

zanikanie polaryzacji

polaryzacja dielektryka niejednorodnego

Polaryzacja całkowita dielektryka jednorodnego składa się z polaryzacji deformacyjnej i polaryzacji orientacyjnej. Materiały dielektryczne są jednak na ogół układami niejednorodnymi. Występujące w ciałach stałych defekty sieci krystalicznej (defekty punktowe, dyslokacje lub granice ziaren) stanowią pułapki dla poruszających się pod wpływem przyłożonego pola elektrycznego nośników prądu i w efekcie na nich gromadzi się ładunek elektryczny. Ten przestrzenny ładunek indukuje ładunek na elektrodach,



Rys. 7. Zależność całkowitej polaryzacji dielektryka niejednorodnego od częstości

co powoduje powstanie dodatkowej polaryzacji P_i. Dyspersja i absorpcja dielektryczna związana z ładunkiem zgromadzonym na niejednorodnościach dielektryka występuje w obszarze częstości od 10^{-5} Hz do 10^2 Hz (częstości infraakustyczne i akustyczne). Zależność całkowitej polaryzacji dielektryka niejednorodnego od częstości przedstawiono na rys. 7.

Ferroelektryki

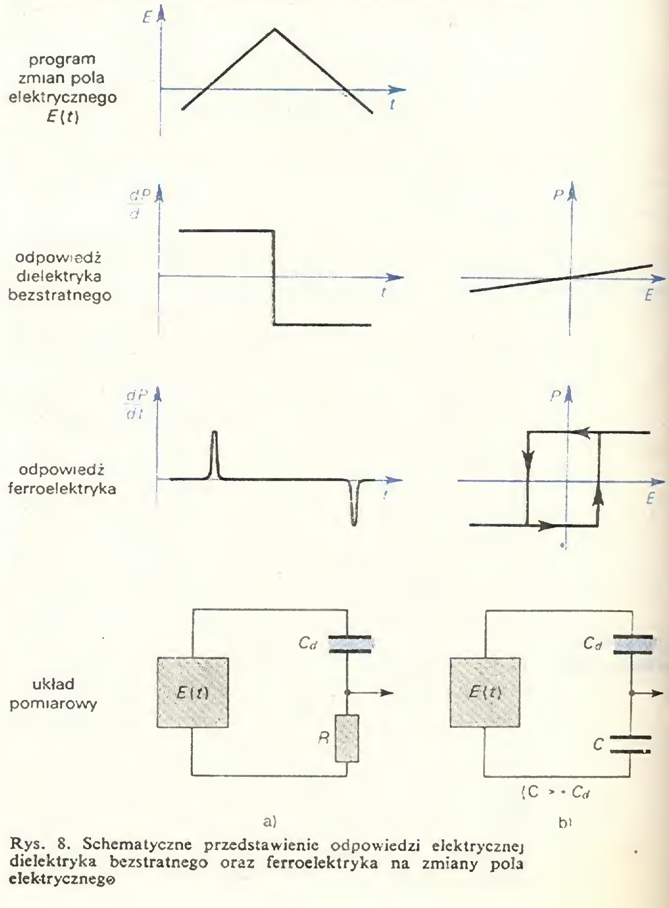
Omówione powyżej własności dielektryków pozwalają na wyciągnięcie wniosku, że najważniejszą cechą charakteryzującą dielektryki jest możliwość zmiany ich polaryzacji za pomocą małych zmian zewnętrznego pola elektrycznego. Polaryzacja kryształów dielektrycznych może się jednak zmieniać również wskutek działania innych czynników zewnętrznych, takich jak naprężenia mechaniczne lub temperatura. Kryształy, w których małe zmiany naprężeń zewnętrznych wywołują zmianę polaryzacji (niezależnie od zmian wywołanych polem elektrycznym), nazywamy kryształami piezoelektrycznymi. Niektóre spośród piezoelektryków wykazują polaryzację spontaniczną, tzn. polaryzację dielektryczną występującą w zerowym polu elektrycznym i przy zerowych naprężeniach mechanicznych, a ponadto polaryzacja ta zmienia się wskutek małych zmian temperatury. Kryształy takie to piroelektryki. Ze względu na swoje własności kryształy te znajdują szerokie zastosowanie jako detektory promieniowania cieplnego, szczególnie podczerwonego, np. w noktowizorach. Niektóre spośród kryształów piroelektrycznych mają jeszcze dodatkowo niezwykle interesującą własność, mianowicie kierunek ich polaryzacji spontanicznej może być zmieniony za pomocą zewnętrznego pola elektrycznego o natężeniu małym w porównaniu z natężeniem pola wewnętrznego. Kryształy te nazywamy ferroelektrykami, ich najważniejszą cechą jest możliwość przepolaryzowania. Wystąpienie własności piezo- i piroelektrycznych dielektryków można w zasadzie przewidzieć na podstawie znajomości klasy symetrii kryształu, nie można jednak oczywiście ocenić wielkości występującego efektu. Piezoelektrykami mogą być kryształy należące do 20 klas nie mających środka symetrii (spośród 32 klas; → Budowa kryształów), natomiast tylko kryształy należące do 10 klas mających oś symetrii mogą wykazywać własności piroelektryczne. Warunkiem koniecznym wystąpienia własności ferroelektrycznych jest przynależność kryształu do klasy polarnej (tzn. mającej oś symetrii), nie jest to jednak warunek dostateczny. Jedynie eksperymentalnie można rozstrzygnąć, czy kierunek polaryzacji może być zmieniony za pomocą zewnętrznego pola elektrycznego. W czasie przepolaryzowania przez kryształ ferroelektryczny płynie prąd przesunięcia, którego gęstość jest w każdej chwili miarą szybkości zmian polaryzacji. Szybkość zmiany polaryzacji po przyłożeniu pola elektrycznego można badać mierząc rzeczywisty prąd płynący przez opór połączony szeregowo z kryształem (rys. 8a), natomiast polaryzację spontaniczną można wyznaczyć stosując układ podany na rys. 8b. Na rysunku tym przedstawiono również różnicę odpowiedzi kryształów dielektrycznych bez-

piezo-elektryki

piroelektryki

ferroelektryki

przepola-ryzowanie

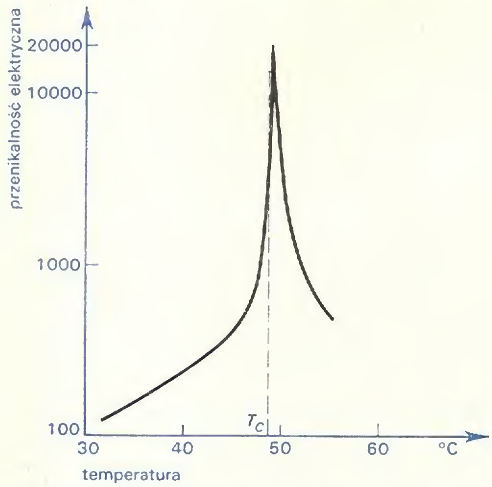


Rys. 8. Schematyczne przedstawienie odpowiedzi elektrycznej dielektryka bezstratnego oraz ferroelektryka na zmiany pola elektrycznego

stratnych i kryształów ferroelektrycznych na zmiany zewnętrznego pola elektrycznego. Z rysunku widać, że polaryzacja dielektryczna ferroelektryków nie jest liniową funkcją natężenia pola elektrycznego; zależność $P(E)$ ma kształt typowej pętli histerezy. Nieliniowość dielektryczna ferroelektryków jest ich bardzo ważną własnością; przejawia się ona również w zależności przenikalności elektrycznej od natężenia pola elektrycznego. Pojemność kondensatora zawierającego ferroelektryk zależy bardzo silnie od natężenia mierzącego pola — mierzona za pomocą pola przemiennego, o bardzo małym natężeniu, zależy również od natężenia stałego pola polaryzującego. Trzecią bardzo ważną własnością ferroelektryków są anomalie własności fizycznych w temperaturze Curie T_C (temperatura przejścia ze stanu ferroelektrycznego do stanu paraelektrycznego). W temperaturze tej przenikalność elektryczna osiąga wartość maksymalną, a następnie w fazie paraelektrycznej maleje hiperbolicznie ze wzrostem temperatury.

nieliniowość dielektryczna

anomalie w temperaturze Curie



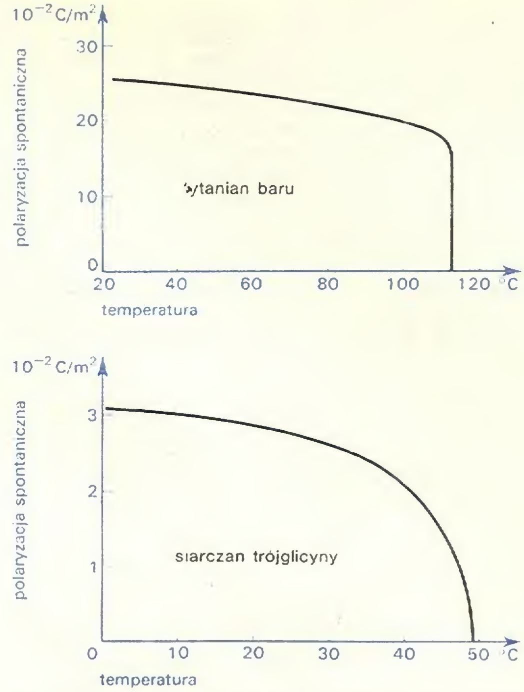
Rys. 9. Zależność od temperatury przenikalności elektrycznej siarczanu trójglicyny (przy częstościach radiowych)

Jej temperaturową zależność opisuje dobrze prawo Curie–Weissa (rys. 9), podobnie jak dla ciał ferromagnetycznych (→ Teoria magnetyzmu). Niektóre spośród ferroelektryków są w fazie paraelektrycznej również piezoelektrykami (np. sól Seignette'a, kwaśny fosforan potasu *KDP*) i dla tych kryształów pewne, wyróżnione przez symetrię kryształu, moduły piezoelektryczne oraz moduły sprężystości osiągają w temperaturze Curie również anomalnie duże wartości. Na rys. 10 pokazano charakterystyczną temperaturową zależność polaryzacji spontanicznej dla tytanianu baru oraz siarczanu trójglicyny. W pierwszym wypadku przejście ze stanu ferroelektrycznego do stanu paraelektrycznego jest przemianą fazową I rodzaju i jest związane z zanikiem polaryzacji pochodzącej od przesunięcia jonów i elektronów w komórce elementarnej typu perowskitu. Drugi wypadek (siarczan trójglicyny) jest przykładem przemiany fazowej II rodzaju i to tzw. przemiany fazowej typu porządek–nieporządek. Model przemiany fazowej typu porządek–nieporządek przyjmuje istnienie w sieci krystalicznej elektrycznym momentów dipolowych, odwracalnych zewnętrznym polem elektrycznym, przy czym w fazie ferroelektrycznej równoległe uporządkowanie tych dipoli w dużej objętości jest przyczyną polaryzacji spontanicznej kryształu. Powyżej temperatury Curie znika uporządkowanie dipoli, a więc znika polaryzacja spontaniczna. Aby określić stopień uporządkowania ferroelektrycznego wprowadza się parametr uporządkowania dalekiego zasięgu $S = (N_+ - N_-)/(N_+ + N_-)$, gdzie N_+ oznacza liczbę dipoli elementarnych zorientowanych w wybranym kierunku dodatnim, a N_- —

prawo Curie–Weissa

zależność polaryzacji spontanicznej od temperatury

stopień uporządkowania

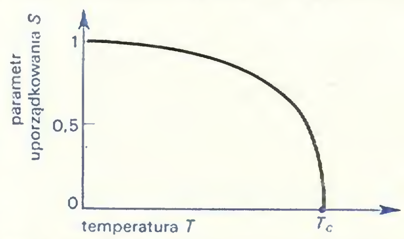


Rys. 10. Zależność od temperatury polaryzacji spontanicznej: a) tytanianu baru i b) siarczanu trójglicyny

liczbę dipoli zorientowanych w kierunku przeciwnym. Parametr uporządkowania ma wartości: $S = 0$ w fazie paraelektrycznej, $S = +1$ w fazie ferroelektrycznej (oznacza to, że wszystkie dipole są zorientowane w kierunku dodatnim), $S = -1$, gdy wszystkie dipole są zorientowane w kierunku ujemnym, oraz $S < 1$ w wypadku niecałkowitego uporządkowania. W tym prostym modelu polaryzacja spontaniczna $P_s = \mu(N_+ + N_-)S$, a temperaturowa zależność parametru uporządkowania dalekiego zasięgu (rys. 11) określa temperaturową zależność polaryzacji spontanicznej. Istnieją również kryształy, które mają dwie identyczne, antyrównolegle spolaryzowane podsieci, tzw. antyferroelektryki. Wypadkowa polaryzacja

parametr uporządkowania

antyferroelektryki



Rys. 11. Zależność od temperatury parametru uporządkowania dla ferroelektryków wykazujących przejście fazowe typu porządek–nieporządek

spontaniczna takich kryształów jest równa zeru. Gdy antyrównolegle spolaryzowane podsieci nie są identyczne, mamy kryształ ferrielektryczny o małej wypadkowej wartości polaryzacji spontanicznej.

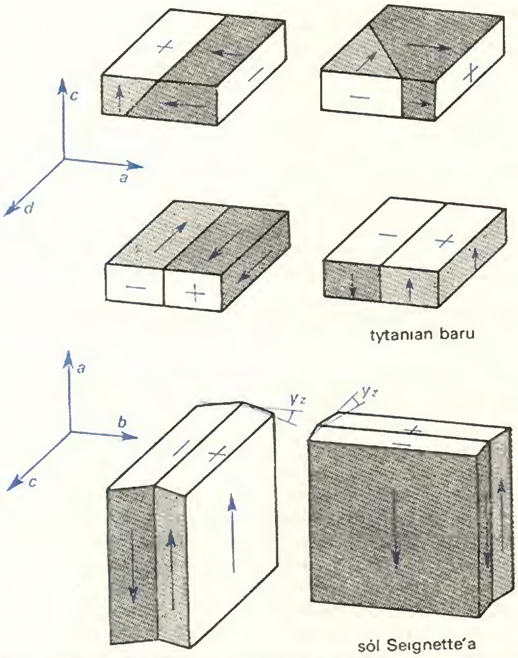
ferrielektryki

Temperaturowa zależność przenikalności elektrycznej dla modelu porządek–nieporządek jest określona głównie temperaturową zależnością funkcji korelacji (→ Przejścia fazowe i zjawiska krytyczne) opisującej, jak silny jest wpływ dipola o dodatniej (wyróżnionej) polarności na orientację innych dipoli. Funkcja ta ma maksimum w temperaturze Curie. Oznacza to, że przemiana fazowa II rodzaju w ferroelektrykach jest zjawiskiem związanym z wystąpieniem bardzo silnych oddziaływań między dipolami ferroelektrycznymi (których orientację można zmienić za pomocą zewnętrznego pola elektrycznego).

W krysztale ferroelektrycznym polaryzacja spontaniczna nie ma w całej objętości takiego samego kierunku. Kryształ jest podzielony na obszary elektrycznie uporządkowane (w których dipole ferroelektryczne są zorientowane w tym samym kierunku), zwane domenami ferroelektrycznymi. Domeny są obszarami makroskopowymi, a polaryzacja spontaniczna w sąsiednich różnych domenach ma taką samą wartość, lecz różni się zwrotem, przy czym kierunek polaryzacji musi się pokrywać z kierunkiem odpowiedniej osi polarnej. Znając symetrię kryształu w fazie paraelektrycznej oraz w fazie polarnej można przewidzieć dozwolone orientacje ściany domenowej w krysztale. Są one określone warunkami ciągłości składowej polaryzacji. W tytanianie baru w fazie

struktura domenowa

ściany domenowe



Rys. 12. Możliwe orientacje ściany domowej: a) w tytanianie baru, b) w soli Seignette'a (deformacja y_z przesadnie powiększona)

tetragonalnej możliwe są tzw. ściany 180°, oddzielające domeny o antyrównoległych kierunkach polaryzacji, oraz ściany 90°, oddzielające domeny, w których kierunki polaryzacji są prostopadłe (rys. 12). W siarczanie trójglicyny, który w fazie paraelektrycznej należy do klasy centrosymetrycznej, a w którym w fazie ferroelektrycznej polaryzacja spontaniczna jest równoległa do osi *b* układu jednoskośnego, możliwe są jedynie ściany 180°, równoległe do osi *b*. W soli Seignette'a, która również w fazie paraelektrycznej jest piezoelektrykiem, a polaryzacja spontaniczna w fazie ferroelektrycznej jest równoległa do osi *a* układu jednoskośnego, występuje piezoelektryczna deformacja y_z. W tym krysztale możliwe są dwie orientacje ścian domenowych; występują ściany domenowe równoległe do osi *a* oraz ściany domenowe równoległe do osi *c* (rys. 12). Podział kryształu ferroelektrycznego na domeny jest uwarunkowany względami energetycznymi. Kryształ dzieli się na domeny w taki sposób, aby ich energia wewnętrzna, składająca się z energii objętościowej (elektrycznej proporcjonalnej do kwadratu polaryzacji oraz sprężystej proporcjonalnej do kwadratu deformacji), energii depolaryzacji oraz energii ścian domenowych miała wartość minimalną. Rozmiary domen w różnych ferroelektrykach są więc różne, ale zależą również od wielkości kryształu. Ściany domenowe w ferroelektrykach są cienkie w porównaniu ze ścianami domenowymi w ferromagnetykach. Mają one grubość

rozmiary domen

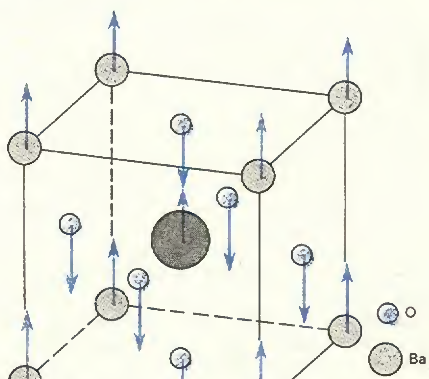
rzędu kilku stałych sieci, podczas gdy w ferromagnetykach grubość jest rzędu kilkuset stałych. Duża grubość ścian domenowych w ferromagnetykach (→ Struktura domenowa i procesy magnesowania) jest związana z ciągłą zmianą kierunku namagnesowania w ściance, podczas gdy w ferroelektrykach moment dipolowy komórki elementarnej może mieć jedynie kierunek jednej z osi polarnych. Dla przykładu można podać, że grubość ściany domenowej w siarczanie trójglicyny, wyznaczona metodą mikroskopii elektronowej, wynosi około 12 nm. Do badania struktury domenowej ferroelektryków stosuje się różne metody, których wybór jest uzależniony własnościami badanych kryształów. Na il. 93 (tabl. 23) podano zdjęcia struktury domenowej ferroelektryków otrzymane różnymi metodami.

grubość ścian domen

Fenomenologiczna teoria ferroelektryków podana przez A. F. Devonshire'a oraz W. L. Ginzburga jest rozwinięciem teorii przemian fazowych L. Landaua. Teoria Landaua była oparta na założeniu, że w pobliżu punktu przemiany potencjał termodynamiczny można rozwinąć w szereg potęgowy względem parametru uporządkowania, a współczynniki rozwinięcia są funkcjami temperatury oraz ciśnienia. Devonshire przyjął, że parametrem uporządkowania ferroelektrycznego przejścia fazowego jest polaryzacja. Opracowana przez niego teoria opisuje dobrze niektóre własności ferroelektryków. W 1959 W. Cochran i P. W. Anderson wysunęli koncepcję, że ferroelektryczna przemiana fazowa jest wynikiem niestabilności jednego z normalnych modów drgań sieci krystalicznej (→ Dynamika sieci krystalicznej) i ten rodzaj drgań został nazwany miękkim modem. Oznacza to, że jedno z elementarnych wzbudzeń układu staje się niestabilne, gdy temperatura $T \to T_C$ (temperatura Curie). Częstość tego wzbudzenia dąży do 0, gdy wektor falowy dąży do zera. Z teorii dynamiki sieci ferroelektryków wynika, że parametrem ferroelektrycznej przemiany fazowej są wartości wektora falowego niestabilnych fononów. Ponieważ stan ferroelektryczny jest stanem polarnym, zatem „miękkie" fonony muszą być polarne, są więc aktywne w podczerwieni. Upraszczając można powiedzieć, że wartości wektora falowego miękkiego modu opisują przemieszczenie atomów wewnątrz komórki elementarnej oraz różnice fazowe między odpowiednimi przesunięciami. Struktura nowej fazy (fazy ferroelektrycznej) poniżej temperatury przemiany jest superpozycją zamrożonych przesunięć wywołanych miękkim modem oraz struktury fazy stabilnej powyżej T_C (rys. 13). Teoria miękkich fononów została najpierw rozwinięta dla ferroelektryków typu przesunięcia, jednak tego samego rodzaju koncepcja została później zastosowana do ferroelektryków typu porządek–nieporządek.

teorie stanu ferroelektrycznego

„miękkie" fonony



Rys. 13. Obrazowe przedstawienie wektora falowego miękkiego modu ferroelektrycznego w $BaTiO_3$ powodującego przemianę fazową $O_h \to C_{4v}$

ferroelektryki niewłaściwe

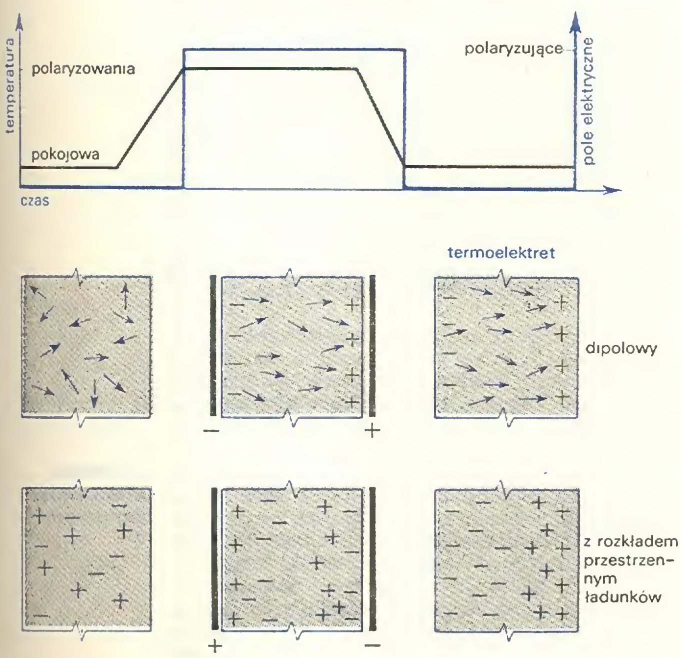
Istnieją ferroelektryki, w których polaryzacja spontaniczna nie jest parametrem uporządkowania. Takie ferroelektryki nazwano ferroelektrykami niewłaściwymi. Parametr uporządkowania może mieć taką samą symetrię jak polaryzacja spontaniczna, tak jak w kryształach KH_2PO_4, albo może mieć inne własności symetrii niż P_s (np. w molibdenianie gadolinu $Gd_2(MoO_4)_3$).

Elektrety

Wspomnieliśmy już, że najważniejszą własnością dielektryków jest zdolność do gromadzenia ładunku elektrycznego. Dielektryk, w którym po usunięciu pola elektrycznego utrzymuje się polaryzacja dipolowa lub rzeczywiste ładunki elektryczne, nazywamy elektretem. Elektret wytwarza zewnętrzne pole elektryczne i w tym sensie jest elektrostatycznym odpowiednikiem magnesu trwałego. Elektrety znane są już od dawna (w 1919 r. fizyk japoński M. Eguchi wyprodukował pierwszy elektret z wosku karnauba). Bardzo długo stan elektretowy wiązano z trwałym uporządkowaniem orientacji dipoli elektrycznych i oddziaływaniem rzeczywistych ładunków wstrzykniętych w procesie ładowania z polaryzacją dipolową, a wytwarzane elektrety, głównie z dielektryków dipolowych, miały grubość rzędu centymetrów. Dopiero pod koniec lat sześćdziesiątych zaczęto wytwarzać elektrety z folii polimerowych wstrzykując z zewnętrznych źródeł rzeczywisty ładunek elektryczny.

wytwarzanie elektretów

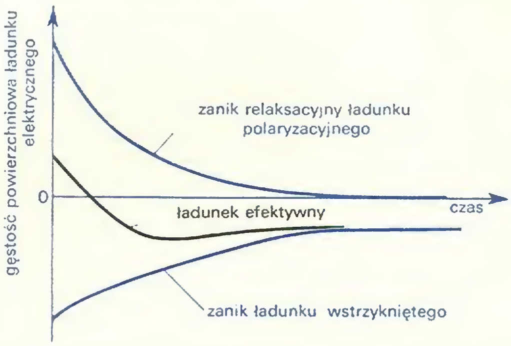
Najpierw omówimy sposób wytwarzania klasycznych termoelektretów, których pole elektryczne jest związane z zamrożoną polaryzacją dipolową (lub pochodzi od uporządkowanego ładunku przestrzennego) i ładunkami wstrzykniętymi w procesie wytwarzania z przestrzeni pomiędzy dielektrykiem a elektrodą. Materiałem na termoelektrety dipolowe są dielektryki, których moment dipolowy pochodzi na ogół od grup karboksylowych lub grup C—Cl dużych molekuł organicznych. Klasycznym materiałem termoelektretowym jest wosk karnauba. Termoelektrety dipolowe otrzymuje się przez polaryzowanie dielektryków



Rys. 14. Obrazowe przedstawienie procesu wytwarzania termoelektretów dipolowych oraz termoelektretów z rozkładem przestrzennym ładunku (ładunki wstrzyknięte zaznaczono kolorem czarnym)

termoelektrety

w silnym zewnętrznym polu elektrycznym w podwyższonej temperaturze (w której możliwy jest obrót dipoli molekularnych) i „zamrożenie" uporządkowania dipoli elektrycznych przez ochłodzenie dielektryka (rys. 14). Ładunek na powierzchni elektretu zaraz po jego wytworzeniu pochodzi przede wszystkim od polaryzacji dipolowej, ma więc znak przeciwny do znaku polarności elektrod (tzw. heteroładunek). Maleje on w czasie w sposób relaksacyjny. W procesie wytwarzania elektretów zostaje jednak również do dielektryka wstrzyknięty ładunek rzeczywisty z przestrzeni pomiędzy elektrodami a dielektrykiem (homoładunek), który zmniejsza się bardzo powoli w czasie. Efektywny ładunek powierzchniowy elektretu jest więc sumą ładunku pochodzącego od polaryzacji dipolowej i ładunku wstrzykniętego. Ponieważ czasowe zmiany gęstości tych ładunków są różne, termoelektret dipolowy wykazuje zaraz po wytworzeniu na powierzchniach heteroładunek, który maleje do zera, a następnie pojawia się homoładunek, który narasta do wartości maksymalnej i dalej zmienia się już w czasie bardzo powoli. Czasowe zmiany efektywnego ładunku termoelektretów obrazuje rys. 15.



Rys. 15. Zmiany w czasie ładunku termoelektretów dipolowych

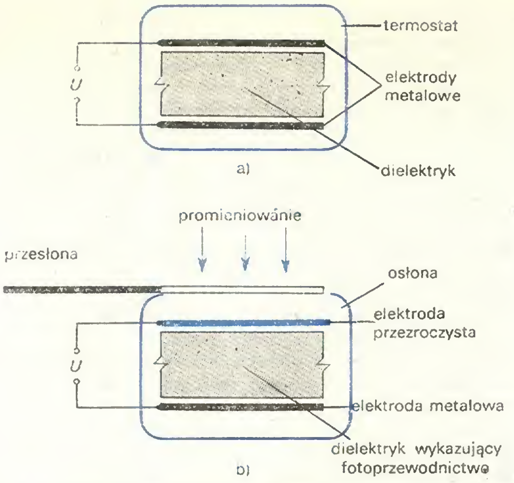
czas zmiany polarności

Czas zmiany polarności elektretu oraz wielkość ustabilizowanego ładunku i czas życia termoelektretu zależą od rodzaju materiału, warunków polaryzowania oraz warunków przechowywania. Polaryzując np. wosk karnauba w słabych polach (mniejszych od 10^6 V/m) otrzymuje się elektrety wykazujące wyłącznie polaryzację dipolową, malejącą w ciągu 10–20 dni do pewnej wartości, która utrzymuje się następnie w czasie rzędu kilku miesięcy. Elektrety z wosku karnauba wytworzone w polach większych od 10^6 V/m wykazują na powierzchni heteroładunek, który w czasie (rzędu minut lub godzin, w zależności od natężenia pola polaryzującego i temperatury przechowywania elektretu) zmniejsza się do wartości zerowej i następnie pojawia się homoładunek, którego ustabilizowana wartość nie zmienia się w czasie rzędu kilkunastu lat. Termoelektrety można otrzymać również z dielektryków niedipolowych, gdy rozłożone izotropowo w objętości ładunki elektryczne zostaną uporządkowane przez pole polaryzujące w podwyższonej temperaturze i ten stan uporządkowania zostaje zamrożony.

fotoelektrety

Porządkujące działanie pola elektrycznego wykorzystuje się również przy produkcji fotoelektretów i radioelektretów, których zasadę otrzymywania pokazuje rys. 16b. W wypadku fotoelektretów, które są wytwarzane z dielektryków wykazujących fotoprzewodnictwo, nośniki prądu zostają wzbudzone do pasma przewodnictwa pod wpływem światła i następnie spułapkowane w pobliżu elektrod o odpowiedniej polarności. Prowadzi to do pojawienia się warstwy ładunku przestrzennego. Depolaryzacja takich elektretów jest związana z aktywacją nośników z pułapek wskutek wzbudzenia termicznego lub wzbudzenia optycznego. Podobny jest mechanizm wytwarzania polaryzacji w radioelektretach, z tym że wzbudzenie jest wywołane promieniowaniem jonizującym.

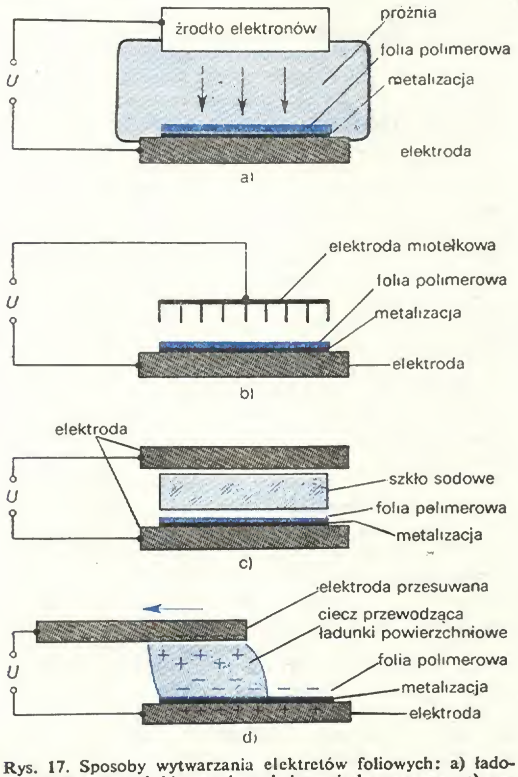
radioelektrety



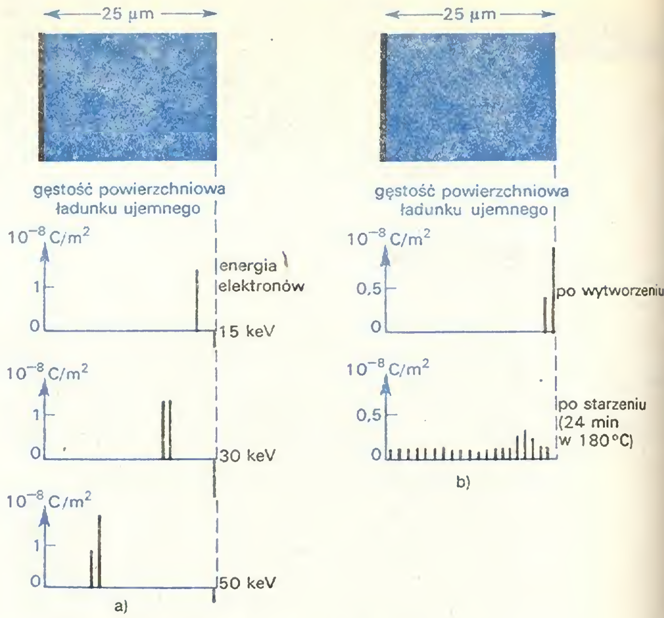
Rys. 16. Układ do wytwarzania: a) termoelektretów, b) fotoelektretów

elektrety foliowe

Największe zastosowanie praktyczne znalazły jednak elektrety z cienkich folii polimerowych, w których zgromadzono duże gęstości ładunku przez implantowanie go z zewnętrznych źródeł. Najprostszą metodą jest ładowanie folii w próżni wiązką elektronów (rys. 17a) o energiach rzędu kilkudziesięciu keV. Efektywny ładunek na powierzchni elektretu zależy od energii elektronów, natężenia wiązki i czasu bombardowania. Metodą tą można zgromadzić w dielektryku bardzo dużą gęstość ładunku; jedynym ograniczeniem gęstości są przebicia. Ładunek wprowadzony przez napromienienie wiązką monoenergetycznych elektronów jest zgromadzony w dielektryku na określonej głębokości, zdeterminowanej energią elektronów. Obrazuje to rys. 18a, na którym pokazano przestrzenny rozkład ładunku w elektretach wytwarzanych przez ładowanie elektronami o różnych energiach.



Rys. 17. Sposoby wytwarzania elektretów foliowych: a) ładowanie elektronami, b) metoda wyładowania koronowego, c) metoda przebiciowa, d) metoda cieczowa



Rys. 18. Rozkład ładunku we wnętrzu elektretu foliowego (wyznaczony przez Collinsa): a) elektret ładowany wiązką elektronów o różnych energiach (ładunek dodatni na powierzchni jest wywołany emisją wtórnych elektronów z dielektryka), b) elekret ładowany metodą cieczową

Elektrety foliowe o dużej gęstości ładunku ujemnego można otrzymać również wykorzystując wyładowanie koronowe w powietrzu (rys. 17b), podczas którego powierzchnia dielektryka jest bombardowana elektronami oraz różnego rodzaju jonami ujemnymi. Zjawisko to jest nieco bardziej skomplikowane, gdyż elektrony wnikają do dielektryka na pewną głębokość, natomiast jony ujemne rekombinują, wywołując zmiany chemiczne na powierzchni dielektryka. W efekcie otrzymuje się jednak elektrety o dużej gęstości ładunku ujemnego stabilnego w czasie. Należy jednak zaznaczyć, że rozkład efektywnego ładunku na powierzchni elektretów wytwarzanych metodą wyładowania koronowego wykazuje pewną niejednorodność.

Do produkcji elektretów foliowych wykorzystuje się również procesy przebiciowe na dużej powierzchni (rys. 17c). W metodzie tej folia dielektryka jest umieszczona pomiędzy elektrodami, do których przykłada się duże napięcie elektryczne, przy czym w celu stabilizacji procesów przebiciowych jedna z elektrod jest elektrodą o dużym oporze (np. szklaną). Gęstość ładunku wytworzonego w ten sposób elektretu zależy zarówno od przyłożonego napięcia jak i czasu polaryzacji, a stabilność jest porównywalna ze stabilnością elektretów wytworzonych przez bombardowanie elektronami.

Elektrety foliowe można również otrzymać za pomocą metody, w której wykorzystano przekazywanie ładunku do dielektryka z elektrody cieczowej o dużej przewodności elektrycznej (rys. 17d). Na powierzchni folii gromadzi się duże gęstości ładunku usuwając w polu elektrycznym warstewkę cieczy.

starzenie elektretów foliowych

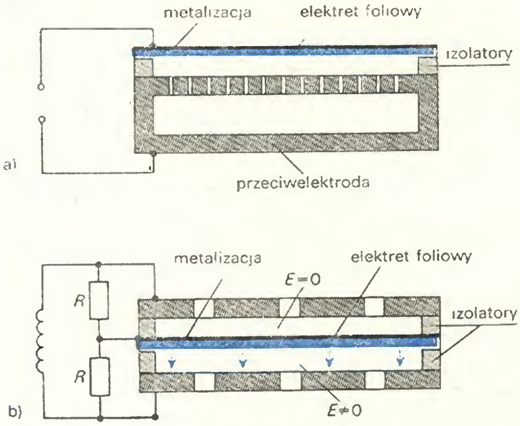
Stosując zaprogramowane starzenie przez wygrzanie elektretu w podwyższonej temperaturze można ładunek zgromadzony blisko powierzchni wprowadzić w głąb dielektryka. Obrazuje to rys. 18b. W procesie starzenia zmniejsza się efektywna gęstość powierzchniowa ładunku elektretu, natomiast ładunek pozostały wykazuje większą stabilność czasową. Jest to korzystne, ponieważ elektrety wyprodukowane omówionymi powyżej metodami mają napięcie równoważne rzędu 1000 V, natomiast do zastosowań praktycznych potrzebne są na ogół elektrety o napięciach około dziesięciokrotnie mniejszych.

elektretowe przetworniki elektro-akustyczne

Jak wspomniano poprzednio, elektrety wytwarzają zewnętrzne pole elektryczne i z tego względu znalazły szerokie zastosowanie w urządzeniach, w których pole to jest wykorzystywane w sposób zachowawczy, mianowicie w elektretowych przetwornikach elektroakustycznych. Możliwość wykorzystania elektretów foliowych jako membrany w mikrofonie pojemnościowym zmniejszyła wyraźnie koszty produkcji i umożliwiła miniaturyzację, a więc uczyniła mikrofon elektretowy szczególnie atrakcyjnym.

mikrofon elektretowy

Na rys. 19 pokazano schemat budowy mikrofonu elektretowego, w którym elektretem jest folia teflonowa jednostronnie metalizowana (grubości 13 μm lub 25 μm). Gęstość powierzchniowa ładunku takiego



Rys. 19. Schemat budowy: a) mikrofonu i b) głośnika w asymetrycznym układzie przeciwsobnym

elektretu jest rzędu 10 nC/cm² i w mikrofonie jest on oddalony od przeciwelektrody o ok. 10 μm. Napięcie wyjściowe V_m mikrofonu elektretowego, przy danym wychyleniu membrany x, zależy od natężenia pola elektrycznego w szczelinie powietrznej przetwornika. Jest ono określone gęstością powierzchniową ładunku elektretu A, grubością elektretu d_e, grubością warstwy powietrza d_0 oraz przenikalnościami elektrycznymi ε_0 oraz ε:

$$V_m = \frac{A d_0 x}{\varepsilon_0(d_0 + d_e\varepsilon)}.$$

W mikrofonach pojemnościowych (a takim jest właśnie przedstawiony mikrofon elektretowy) wychylenie membrany zależy liniowo od ciśnienia fali akustycznej w szerokim zakresie częstości, począwszy od częstości odcięcia do ok. 20 000 Hz. Odpowiedź napięciowa mikrofonu elektretowego przy stałym natężeniu dźwięku jest więc niezależna od częstości.

głośnik elektretowy

Elektret foliowy może być również zastosowany w głośniku; na rys. 19b przedstawiono schemat budowy głośnika elektretowego pracującego w asymetrycznym układzie przeciwsobnym. Przedstawione schematy budowy elektretowych przetworników elektroakustycznych są schematami najprostszymi. Produkcję ich rozpoczęto w 1970 r. Obecnie znane są konstrukcje elektretowych przetworników elektroakustycznych o specjalnych kierunkowych charakterystykach oraz konstrukcje przetworników pracujących w zakresie częstości infraakustycznych oraz ultraakustycznych (od 10^{-3} Hz do 10^8 Hz).

badanie starzenia elektretów

Na zakończenie należy wspomnieć, że szerokie zastosowanie różnego rodzaju elektretów (np. fotoelektretów w kserografii, elektretów foliowych w przetwornikach elektroakustycznych, różnego rodzaju przetwornikach elektromechanicznych oraz filtrach elektrostatycznych) wyprzedziły znacznie stan wiedzy o podstawowych procesach fizycznych gromadzenia ładunku we wnętrzu dielektryka i jego reorganizacji w procesie starzenia. Wydaje się, że postęp ten nastąpi w najbliższym czasie, gdyż w 1977 r. opracowano dwie interesujące metody badania przestrzennego rozkładu ładunku w elektretach foliowych. Jedna z nich opracowana przez G. M. Sesslera, pozwalająca otrzymać informacje o rozkładzie gęstości ładunku na grubości elektretu z dokładnością 1 μm, wykorzystuje jako sondę wirtualną elektrodę, którą wytwarza się za pomocą monoenergetycznej wiązki elektronowej. Druga metoda, opracowana przez R. E. Collinsa, jest metodą nieniszczącą i polega na badaniu odpowiedzi elektrycznej po przyłożeniu do jednej strony elektretu bardzo krótkiego impulsu cieplnego. Stosując tę metodę można otrzymać bezpośrednio informacje o wielkości ładunku elektretowego i jego średnim położeniu, a po opracowaniu matematycznym — informacje o przestrzennym rozkładzie ładunku w elektrecie i ogólnie w dielektryku. Za pomocą wspomnianej metody można badać również zmiany przestrzennego rozkładu gęstości ładunku w procesie samorzutnego oraz zaprogramowanego starzenia elektretów, co rokuje nadzieję na wyjaśnienie mechanizmu tego procesu.

J. C. Anderson *Dielectrics*, London 1967; R. Blinc, B. Žekš *Soft Modes in Ferroelectrics and Antiferroelectrics*, Amsterdam 1974; A. Chełkowski *Fizyka dielektryków*, Warszawa 1979; B. Hilczer, J. Małecki *Elektrety*, Warszawa 1980; S. Kielich *Molekularna optyka nieliniowa*, Warszawa 1977; T. Krajewski (red.) *Zagadnienia fizyki dielektryków*, Warszawa 1970; A. Piekara *Polaryzacja materii*. Materiały Konferencji Chemików w Spale, 1959, str. 268; A. Piekara *Mikrofale i spektroskopia mikrofalowa*, Warszawa 1953.

Półprzewodniki

Tadeusz Figielski

Spośród olbrzymiej grupy ciał stałych, występujących w sposób naturalny w przyrodzie lub wytworzonych przez człowieka, wyróżniamy tzw. półprzewodniki. Są to substancje krystaliczne, przewodzące prąd elektryczny o szczególnie interesujących właściwościach. Zaliczamy do nich materiały wielce różnorodne pod względem budowy chemicznej, a więc zarówno czyste pierwiastki german i krzem, tlenki i siarczki niektórych metali, np. Cu_2O, ZnO, Pbs, związki międzymetaliczne, jak InSb, GaAs, HgTe i wiele innych. Poznanie i wykorzystanie tych materiałów stanowi wielki krok naprzód w rozwoju technologicznym naszej epoki. Przed niespełna trzydziestu laty został wynaleziony tranzystor. Od tej pory rozpoczął się, ciągle jeszcze trwający, niezwykle intensywny rozwój fizyki półprzewodników.

O przewodzeniu prądu elektrycznego

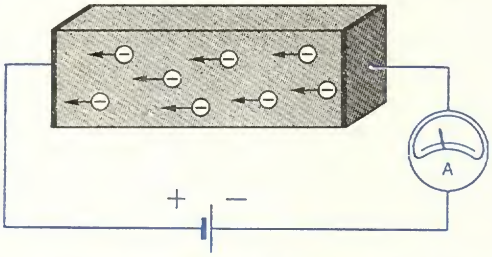
Przewodnictwo elektryczne półprzewodników, podobnie jak metali, ma naturę elektronową, to znaczy elementarnymi nośnikami prądu elektrycznego są tu elektrony, obdarzone najmniejszą możliwą porcją ujemnego ładunku elektryczności. Wchodzą one w skład poszczególnych atomów, z których zbudowany jest kryształ. Jeśli kryształ przewodzi prąd, oznacza to, że przynajmniej część tych elektronów straciła więź ze swoimi macierzystymi atomami i może stosunkowo swobodnie podróżować po całym krysztale. Tak więc możemy wyobrażać sobie, że kryształ metalu czy półprzewodnika zbudowany jest z regularnie rozmieszczonych jonów — sieci krystalicznej — i czegoś w rodzaju gazu prawie swobodnych elektronów

gaz elektronowy

wypełniających tę sieć. W gazie elektronowym poszczególne cząstki poruszają się bezładnie, a ruch ten jest tym intensywniejszy, im wyższa jest temperatura. Z chwilą przyłożenia do kryształu różnicy potencjałów elektrostatycznych na elektrony działa siła wprowadzająca dodatkowy, uporządkowany ruch ładunków, który stanowi właśnie prąd elektryczny.

uporządkowany ruch ładunków elektrycznych

Wygodnie jest wyobrażać sobie kryształ półprzewodnika w postaci prostopadłościennej próbki, której jeden wymiar (długość d) jest znacznie większy od dwóch pozostałych. Taki kształt próbek jest



Rys. 1. Próbka półprzewodnikowa włączona w obwód prądu

najczęściej stosowany w badaniach fizycznych. Chcąc zmierzyć, powiedzmy, przewodność elektryczną próbki, musimy na jej końcach umieścić elektrody zapewniające dobry kontakt elektryczny z zewnętrznym obwodem prądu. Jeśli do naszej próbki przyłożymy różnicę potencjałów U, to wytwarza ona wewnątrz kryształu pole elektryczne o natężeniu $\mathcal{E} = U/d$. Na elektron o ładunku $-e$ działa teraz siła elektrostatyczna, której wartość wynosi $F = -e\mathcal{E}$. Każdy prawie swobodny elektron w krysztale doznaje zatem przyspieszenia o wartości $a = F/m$ w kierunku przeciwnym do pola elektrycznego; m oznacza masę elektronu. Gdyby elektrony były całkowicie swobodne, poruszałyby się ruchem jednostajnie przyspieszonym osiągając coraz większe prędkości. Nie uwzględniliśmy w tym jeszcze bezładnego ruchu cieplnego elektronów. Ponieważ jednak prędkości dodają się tak jak wektory, możemy bez popełnienia błędu wyodrębnić z wypadkowego ruchu składową prędkość wywołaną unoszeniem elektronów przez pole elektryczne i rozpatrywać ją niezależnie.

prędkość unoszenia

Prędkość unoszenia elektronu nie może wzrastać nieskończenie, ponieważ elektron co pewien czas zderza się z jonami sieci krystalicznej i ulega rozproszeniu. Jeśli uwzględnimy, że przyspieszenie elektronu zachodzi w czasie (τ) jego swobodnego przebiegu między dwoma zderzeniami, to otrzymamy na maksymalną prędkość unoszenia wartość $v = e\mathcal{E}\tau/m$. Przy prawidłowym uwzględnieniu statystycznego rozkładu przedziałów czasowych τ dla różnych elektronów w różnych chwilach wzór ten zostaje słuszny, jeśli przez v i τ będziemy rozumieć średnie wartości tych parametrów w ciągu ruchu.

gęstość prądu

Jeśli liczba „swobodnych" elektronów w jednostce objętości kryształu, czyli tzw. koncentracja elektronów, wynosi n, to gęstość prądu (natężenie prądu na jednostkę powierzchni przekroju) płynącego przez kryształ wyniesie $j = env = en(e\tau/m)\mathcal{E}$. Wielkość $e\tau/m$, oznaczana literą μ, jest miarą średniej prędkości, jaką nabywa elektron w jednostkowym polu elektrycznym i nazywa się ruchliwością elektronu. Wyprowadzony tu związek między j i $\mathcal{E}$ jest po prostu prawem Ohma. Współczynnik proporcjonalności pomiędzy j i $\mathcal{E}$ nazywa się przewodnością właściwą σ materiału, zaś jego odwrotność oporem właściwym ϱ. Tak więc $\sigma = 1/\varrho = j/\mathcal{E}$. Korzystając z uprzednich wyprowadzeń można teraz zapisać wyrażenie na σ w postaci:

ruchliwość elektronu

$$\sigma = en\mu.$$

Jest to podstawowy wzór — słuszny nie tylko dla półprzewodników — który wiąże przewodność elektryczną materiału z koncentracją i ruchliwością nośników prądu. Przewodność właściwa ciał zaliczanych do półprzewodników zawiera się w bardzo szerokim zakresie wartości od 10^4 do 10^{-10} $(\Omega\text{cm})^{-1}$.

Trzeba od razu zdać sobie sprawę z dwóch istotnych rzeczy. Po pierwsze, zarówno koncentracja, jak i ruchliwość w konkretnym półprzewodniku mogą drastycznie zmieniać się wraz ze zmianą temperatury. Po drugie, w półprzewodniku mogą występować dwa rodzaje nośników prądu. Wówczas przyczynki do przewodnictwa pochodzące od każdego z nich będą się sumować.

Kwantowy opis półprzewodnika

Traktowanie w dalszym ciągu elektronu jako klasycznej cząstki materii nie zaprowadziłoby nas zbyt daleko. Musimy uwzględnić fakt, że w mikroświecie obowiązują specyficzne prawa, prawa mechaniki kwantowej. Zgodnie z teorią kwantową elektron należy rozpatrywać jako falę, zwaną czasem falą de Broglie'a. Elektronowi o pędzie $p = mv$ musimy przypisać falę o długości $\lambda = h/p$, gdzie h jest stałą Plancka. Jeślibyśmy zapytali się w duchu fizyki klasycznej (tzn. niekwantowej), gdzie dokładnie znajduje się jakiś konkretny elektron, to na gruncie mechaniki kwantowej nie dostalibyśmy jednoznacznej odpowiedzi. Jeśli bowiem elektron jest opisany prostą falą „sinusoidalną", a zatem ma ściśle określony pęd, to jego umiejscowienie, jako fali, jest zupełnie nieokreślone; elektron jest wszędzie. Można, co prawda, skonstruować taką falę, której różna od zera amplituda jest zawarta w pewnym skończonym obszarze przestrzeni, ale będzie ona wówczas złożeniem kilku fal prostych o różnych długościach. Nie będzie ona zatem opisywać elektronu o ściśle określonym pędzie. Mechanika kwantowa podaje dokładny związek między przedziałami nieokreśloności położenia i pędu cząstki w swej słynnej zasadzie — zasadzie nieokreśloności Heisenberga.

elektron jako fala

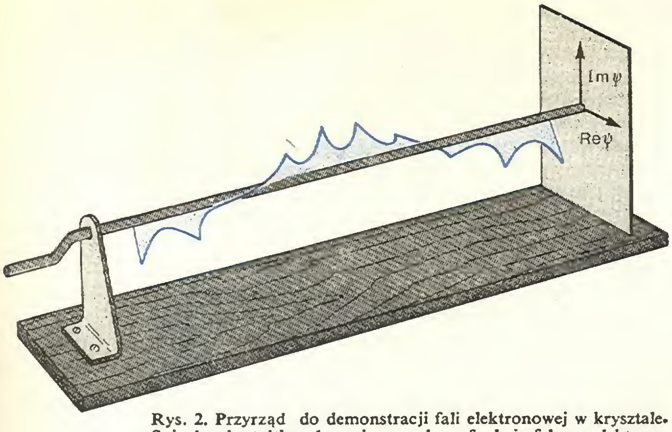
Jedną z uderzających konsekwencji tej zasady jest coś, co można by nazwać ograniczonością miejsca dla elektronów. Mianowicie, dla każdego przedziału pędu (lub energii) istnieje w danej objętości ściśle określona liczba stanów kwantowych, w których mogą znajdować się elektrony. Ponadto, zgodnie z tzw. zakazem Pauliego, w każdym stanie mogą przebywać co najwyżej dwa elektrony. Różne stany kwantowe swobodnego elektronu możemy opisać różnymi wartościami jego energii kinetycznej.

W teorii kwantowej zostały wyprowadzone przez Fermiego i Diraca ogólne prawa statystyczne opisujące prawdopodobieństwo obsadzenia poszczególnych stanów elektronami w zależności od temperatury. I tak w temperaturze zera bezwzględnego będą obsadzone wszystkie możliwe stany o najniższej energii. Przy podwyższaniu temperatury następuje stopniowe „rozluźnianie" zapełnienia; część elektronów przechodzi do stanów o wyższej energii. Jeśli jednak porównujemy dwa dowolne stany kwantowe, to w równowadze prawdopodobieństwo obsadzenia stanu o niższej energii zawsze pozostaje większe.

Prześledźmy teraz, w jaki sposób mechanika kwantowa modyfikuje nasz prosty obraz przewodnictwa elektrycznego. Podstawy teorii kwantowej ciał stałych zostały sformułowane jeszcze w latach trzydziestych, dzięki pracom licznych badaczy, m.in. A. H. Wilsona i F. Blocha. Średnia długość fali elektronu w krysztale półprzewodnikowym jest tysiąc razy większa od odległości między sąsiednimi jonami sieci krystalicznej. Taka fala rozchodzi się w idealnej sieci prawie bez zakłóceń. Można też na to spojrzeć z innego punktu widzenia, traktując samą sieć atomową jako nośnik fali, tj. jako ośrodek, w którym fala elektronowa została wzbudzona.

długość fali elektronowej

Dla uproszczenia będziemy dalej rozpatrywać zachowanie się elektronu w jednowymiarowym, nieskończenie długim krysztale reprezentowanym przez

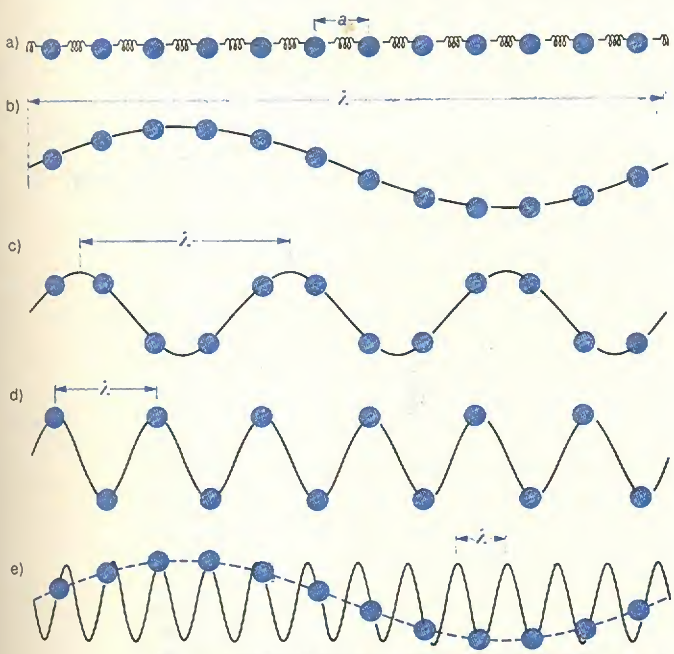


Rys. 2. Przyrząd do demonstracji fali elektronowej w krysztale. Spiralna kształtka obrazuje zespoloną funkcję falową elektronu ψ, której kwadrat modułu w każdym punkcie wzdłuż osi określa prawdopodobieństwo znalezienia tam elektronu. Przez obrót korbki wytwarzamy falę biegnącą, a mimo to rozkład prawdopodobieństwa pozostaje zawsze taki sam

łańcuch równoodległych od siebie atomów. Oczywiście kryształ taki w rzeczywistości nie istnieje. Stanowi on jednak wygodny model, na którym można prześledzić pewne ogólne właściwości trójwymiarowego kryształu. Weźmy najpierw pod uwagę falę sprężystą rozchodzącą się w naszym łańcuchu — falę tego typu, jaka powstałaby, gdybyśmy mocno napięty sznur gumowy uderzyli z góry dłonią. W łańcuchu mogą rozchodzić się fale o najrozmaitszych długościach w dwu przeciwnych kierunkach. W wielu działach fizyki przyjął się wygodny zwyczaj opisywania fali o określonej długości i kierunku za pomocą tzw. wektora falowego $\vec{k}$. Wektor $\vec{k}$ jest to wektor, którego kierunek jest zgodny z kierunkiem rozchodzenia się fali, natomiast jego długość równa się odwrotności fali, pomnożonej przez 2π, tzn. $k = 2\pi/\lambda$.

wektor falowy

Chwilowe wychylenia poszczególnych atomów łańcucha dla prostych fal sinusoidalnych o różnych długościach są przedstawione na rys. 3. Kolejne wykresy odpowiadają malejącym wartościom λ, czyli wzrastającym długościom wektora falowego. Krzywa na rys. 3d przedstawia falę o długości równej podwójnej odległości między sąsiednimi atomami, a więc taką, w której para sąsiednich atomów drga przeciwsobnie. Dla tej fali $\lambda = 2a$ i $k = \pi/a$. Każdej fali o wektorze falowym leżącym w przedziale od 0 do π/a odpowiada jeden, ściśle określony rozkład chwilowych wychyleń atomów. I tu dochodzimy do bardzo osobliwej własności ruchu falowego w ośrodku nieciągłym. Jeśli w sposób analogiczny do poprzedniego przedstawimy fale jeszcze krótsze, to okaże się, że nie tworzą one wcale nowych konfiguracji atomów, które nie byłyby już uwzględnione w przedziale 0, $\pm\pi/a$. Fale, których wektory falowe różnią się o całkowitą wielokrotność $2\pi/a$ są fizycznie nierozróżnialne, a zatem są identyczne. Ta właściwość, która odnosi się równie dobrze do fali elektronowej w krysztale, pozwala zrozumieć różnicę w zachowaniu się swobodnego elektronu i elektronu w sieci krystalicznej.



Rys. 3. Poprzeczne fale sprężyste o malejącej długości λ w jednowymiarowym łańcuchu atomowym (przedstawionym na rys. a): b) $\lambda = 12a, k = \frac{1}{6}\frac{\pi}{a}$; c) $\lambda = 4a, k = \frac{1}{2}\frac{\pi}{a}$; d) $\lambda = 2a, k = \frac{\pi}{a}$; e) $\lambda = \frac{12}{13}a, k = \frac{13}{6}\frac{\pi}{a} = 2\frac{\pi}{a} + \frac{1}{6}\frac{\pi}{a}$; dla fali (e) wychylenia atomów są identyczne jak dla fali (b)

Zauważmy przede wszystkim, że wartość pędu elektronu — odwrotnie proporcjonalną do długości fali de Broglie'a — można wyrazić wprost przez wektor falowy $\vec{k}$. Związek ten ma postać:

$$\vec{p} = h\vec{k}/2\pi = \hbar\vec{k} \ (\hbar = h/2\pi).$$

A zatem wektor falowy i pęd elektronu można uważać za wielkości równoważne, tyle że wyrażone w różnych jednostkach. Związek ten pozostaje słuszny również dla elektronu poruszającego się w krysztale, z tym że wówczas przez p należy rozumieć średni w czasie ruchu pęd elektronu (zwany kwazipędem lub pędem krystalicznym).

kwazipęd elektronu

Energia kinetyczna swobodnego elektronu $E = mv^2/2 = p^2/2m$, co można zapisać teraz jako

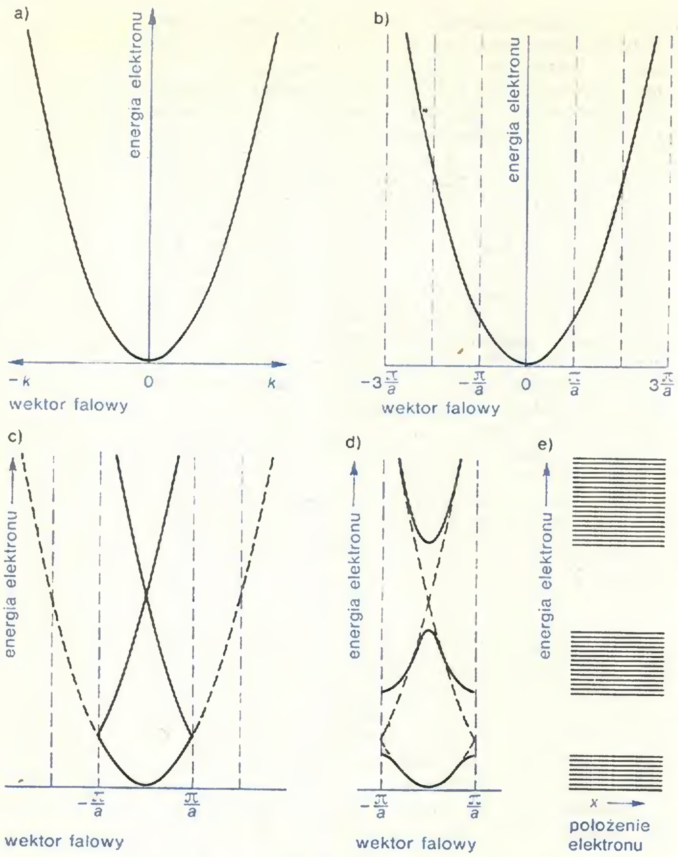
$$E = (1/2\pi)^2(h^2k^2/2m) = \hbar^2k^2/2m.$$

energia elektronu

Wykres energii kinetycznej w funkcji wektora falowego (pędu) jest parabolą. Relacja między energią elektronu a jego wektorem falowym, tj. funkcja $E(\vec{k})$ ma w fizyce ciała stałego fundamentalne znaczenie. Wnikliwy Czytelnik zauważy, że znając zależność funkcyjną E od $\vec{k}$ dla swobodnego elektronu, można wyrazić z niej prędkość elektronu jako pierwszą pochodną $E(\vec{k})$ po $\vec{k}$, i podobnie masę elektronu jako odwrotność drugiej pochodnej $E(\vec{k})$ po $\vec{k}$, oczywiście z pewnymi współczynnikami. W przypadku swobodnego elektronu takie przedstawienie jest tylko formalną sztuczką, natomiast jest ono niezmiernie wygodne przy rozpatrywaniu elektronów w krysztale (→ Dynamika elektronu w ciałach stałych). Jeśli wyjdziemy z założenia, że jony sieci krystalicznej w ogóle nie oddziałują na ruch elektronu, to wykres energii kinetycznej elektronu w naszym jednowymiarowym krysztale będzie taką samą parabolą, jak dla elektronu swobodnego. Ale w sieci wszystkie wektory $\vec{k}$ różniące się o całkowitą wielokrotność $2\pi/a$ opisują elektrony z tym samym pędem krystalicznym. Aby uwzględnić to na naszym wykresie, musimy podzielić go na strefy odpowiadające różnym przedziałom wartości $\vec{k}$ i przenieść odpowiednie odcinki krzywej $E(\vec{k})$ do pierwszej strefy, tzn. do $\vec{k}$ leżących w przedziale $-\pi/a$, π/a. Taki sposób postępowania pokazany jest na rys. 4c. Elektron o określonym pędzie nie ma teraz jednoznacznie określonej energii, gdyż może on znajdować się na jednej z wielu gałęzi krzywej $E(\vec{k})$.

pasma energetyczne

Chcielibyśmy jeszcze wiedzieć, czy i w jaki sposób samo istnienie sieci krystalicznej wpływa na kształt zależności $E(\vec{k})$. Ścisły rachunek kwantowy pokazuje, że wpływ ten jest najsilniejszy tam, gdzie krzywe odpowiadające różnym gałęziom $E(\vec{k})$ najbardziej się do siebie zbliżają, a więc na krawędzi strefy bądź w jej środku. Rezultat jest taki, jak gdyby krzywe w tych



Rys. 4. Jak tworzą się pasma energii dozwolonej i wzbronionej: a) wykres energii kinetycznej swobodnego elektronu w funkcji wektora $\vec{k}$ (pędu), b) ten sam wykres dla prawie swobodnego elektronu w krysztale; zaznaczone są strefy zmienności wektora $\vec{k}$, c) ten sam wykres sprowadzony do pierwszej strefy, d) uwzględniono modyfikację wywołaną jonami sieci, e) symboliczne przedstawienie pasm energetycznych

struktura elektronowa kryształu

miejscach wzajemnie się odpychały, odkształcając się w sposób pokazany na rys. 4d. Wykres ten przedstawia to, co nazywamy strukturą elektronową kryształu (→ Struktura elektronowa ciał stałych). Widzimy, że wytworzyły się teraz pewne obszary energii, jakiej elektron w ogóle nie może posiadać. Mamy więc w krysztale występujące na przemian pasma energii dozwolonej i wzbronionej dla elektronu. Ciekawe, że do takiego samego wniosku doszlibyśmy wychodząc z całkiem odmiennych założeń, a mianowicie biorąc za punkt wyjścia dozwolone poziomy energii w izolowanych atomach, i śledząc co nastąpi, jeśli te atomy zbliżymy do siebie aż do utworzenia kryształu.

Elektrony i dziury

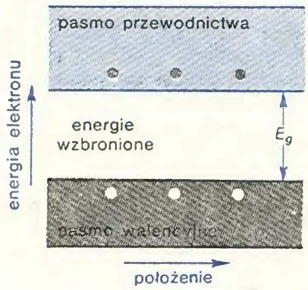
W rzeczywistych, trójwymiarowych kryształach struktura energetyczna jest znacznie bardziej skomplikowana niż ta, którą przedstawiliśmy na rys. 4d. W szczególności krzywizna krzywej $E(\vec{k})$, w każdym jej punkcie może być zupełnie różna niż dla elektronu swobodnego. Miarą krzywizny jest wartość drugiej pochodnej E po $\vec{k}$, a ta dla swobodnego elektronu jest odwrotnie proporcjonalna do jego masy. Przenosząc tego typu argumentację na przypadek elektronu w krysztale, powinniśmy przypisać mu masę całkiem różną od masy elektronu swobodnego! Ten na pozór paradoksalny wniosek ma w istocie rzeczy rozsądne uzasadnienie. Elektron w swoim ruchu poprzez kryształ oddziałuje z ładunkami jonów sieci. Zewnętrzne pole elektryczne, potrzebne do nadania elektronowi w krysztale pewnego średniego przyspieszenia, będzie zatem w ogólności inne niż dla elektronu swobodnego. Nie znając dokładnie rozkładu i wartości wewnętrznych pól elektrycznych w krysztale i chcąc być w zgodzie z prawami fizyki, musimy przypisać elektronowi pewną fikcyjną masę, różną od jego masy rzeczywistej. Nazywamy ją masą efektywną. Nie jest ona wielkością stałą, ale zależy od pędu i energii elektronu. Jeżeli spojrzymy jeszcze raz na rys. 4d, to zauważymy, że w pewnych obszarach, na krawędziach bądź w środku strefy, krzywizna $E(\vec{k})$ jest ujemna, co odpowiada ujemnej masie efektywnej! Nie powinno to nas już jednak specjalnie niepokoić.

masa efektywna

Ze względu na skończoną liczbę stanów kwantowych pojemność każdego z pasm dozwolonej energii jest ograniczona. W pasmie może zmieścić się jedynie ściśle określona liczba elektronów, która, jak można pokazać, jest zawsze pewną wielokrotnością całkowitej liczby atomów w krysztale. Jeżeli niższe pasmo energetyczne jest całkowicie zapełnione, natomiast wyższe jest zupełnie puste, to kryształ będzie wówczas doskonałym izolatorem prądu. Elektron, ażeby przewodzić, musi uzyskać pod wpływem zewnętrznego pola elektrycznego dodatkową prędkość, a zatem przejść do innego stanu kwantowego. Tymczasem wszystkie możliwe stany w pasmie są już zajęte.

izolator

Taka sytuacja może zaistnieć jedynie w temperaturze zera bezwzględnego. W każdej wyższej temperaturze pewna część elektronów jest termicznie wzbudzona do wyższego pustego pasma, gdzie może już uczestniczyć w przewodzeniu prądu. To wyższe pasmo będziemy zatem nazywać pasmem przewodnictwa, natomiast pasmo dolne, zapełnione — pasmem walencyjnym (rys. 5).



Rys. 5. Model pasmowy półprzewodnika. Zaznaczone są wzbudzone termicznie elektrony i dziury

Elektron wzbudzony do pasma przewodnictwa pozostawia po sobie „dziurę", czyli pusty stan w pasmie walencyjnym. A zatem również w tym pasmie może odbywać się teraz przewodzenie. Sytuacja tutaj jest jednak nieco osobliwa. Zgodnie ze statystyką obsadzenia stanów, większość „dziur" będzie się znajdować w obszarze bliskim wierzchołka pasma walencyjnego. Tu jednak elektrony mają ujemną masę efektywną (rys. 4d). Kierunek przyspieszenia takich elektronów będzie przeciwny niż „normalnych" elektronów. Dla obserwatora, który nie wie nic o masie efektywnej, wygląda to tak, jak gdyby elektron był teraz obdarzony ładunkiem dodatnim. W istocie rzeczy ścisły rachunek pokazuje, że zbiór wszystkich elektronów w prawie całkowicie zapełnionym pasmie zachowuje się tak, jak niewielka liczba cząstek obdarzonych dodatnią masą i dodatnim ładunkiem elektrycznym. Liczba tych cząstek jest dokładnie równa liczbie „dziur". Możemy przeto nadać „dziurze" obywatelstwo, uwolnić ją od cudzysłowu i traktować ją jako pełnoprawny dodatni nośnik prądu w krysztale, na równi z ujemnym elektronem. Ściślejsze omówienie tych spraw znajdzie Czytelnik w artykule pt. „Dynamika elektronów w ciałach stałych".

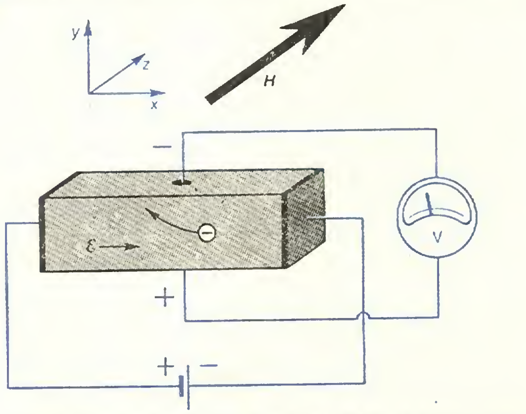
dziury

Ze wzrostem temperatury rośnie koncentracja termicznie wzbudzonych elektronów i dziur i kryształ staje się coraz bardziej przewodzący. Takie przewodnictwo nazywamy samoistnym. Im mniejsza jest szerokość przerwy energetycznej, tj. różnica energii mię-

dzy dnem pasma przewodnictwa i wierzchołkiem pasma walencyjnego, tym większe jest samoistne przewodnictwo kryształu w określonej temperaturze. Sytuacja taka jest typowa zarówno dla izolatorów, jak i półprzewodników. Różnica między tymi dwiema grupami materiałów jest czysto ilościowa i wskutek tego w pewnym sensie umowna. Za półprzewodniki uważamy zazwyczaj te materiały, których przerwa energetyczna jest mniejsza od 5 eV; np. dla InSb wynosi ona 0,18 eV, dla ZnO 3,2 eV. A metale? W metalach mamy sytuację dwojakiego rodzaju; bądź zapełnione pasmo walencyjne i puste pasmo przewodnictwa zachodzą na siebie, bądź też pasmo przewodnictwa jest tylko częściowo zapełnione elektronami (→ Metale).

Możemy się spytać, czy można eksperymentalnie odróżnić przewodnictwo wywołane dziurami od przewodnictwa elektronowego. Otóż tak. Umożliwia to znane od dawna zjawisko Halla, które jest wynikiem

zjawisko Halla



Rys. 6. Powstawanie w próbce napięcia Halla, w wyniku działania pola magnetycznego $\vec{H}$

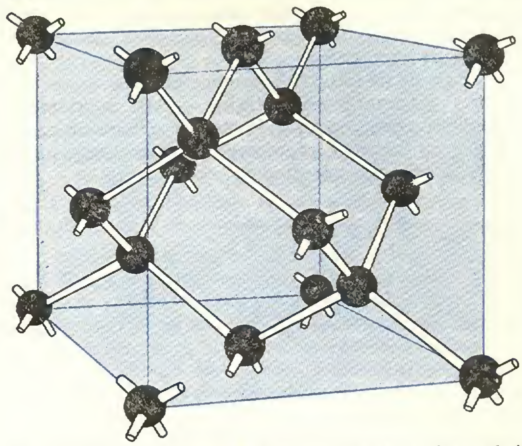
działania siły Lorentza na poruszający się ładunek elektryczny w polu magnetycznym. Jeśli pole elektryczne jest skierowane w kierunku osi x a pole magnetyczne w kierunku osi z i nośnikami prądu są wyłącznie elektrony to pod wpływem siły Lorentza odchylą się one w dodatnim kierunku osi y (rys. 6). Wskutek tego na górnej ściance kryształu, prostopadłej do osi y, wytworzy się w stosunku do ścianki przeciwległej nadmiar ładunku ujemnego i związana z tym różnica potencjałów, zwana napięciem Halla. Jeśli nośnikami prądu są dziury, to na tej samej ściance powstanie nadmiar ładunku dodatniego i znak napięcia Halla jest przeciwny. Wielkość napięcia Halla jest odwrotnie proporcjonalna do koncentracji jednoimiennych nośników prądu. Gdy nośnikami prądu są wzbudzone termicznie elektrony i dziury w ilościach dokładnie sobie równych, to pochodzące od nich przyczynki do napięcia Halla praktycznie się zniosą.

Domieszki w półprzewodnikach

Badania efektu Halla w różnych półprzewodnikach dość wcześnie doprowadziły do wykrycia zdumiewającego faktu. Ten sam pod względem chemicznym półprzewodnik może wykazywać bądź przewodnictwo elektronowe bądź dziurowe, w zależności od rodzaju wprowadzonych doń minimalnych ilości domieszek. Wielkość przewodnictwa wzrasta z domieszkowaniem i może być w ten sposób dokładnie regulowana. Te wyniki świadczą dobitnie o tym, że przedstawiony dotychczas model przewodnictwa nie jest kompletny i wymaga istotnego uzupełnienia. Przeprowadzimy je na przykładzie germanu.

German, podobnie jak krzem, krystalizuje, tworząc tzw. strukturę typu diamentu, w której każdy z atomów jest otoczony czterema najbliższymi sąsiadami (rys. 7). Atom germanu ma w swej zewnętrznej powłoce 4 elektrony, które całkowicie wypełniają pasmo walencyjne. To właśnie powoduje, że kryształ germanu nie jest metalem. Rozpatrując tę sprawę

struktura germanu

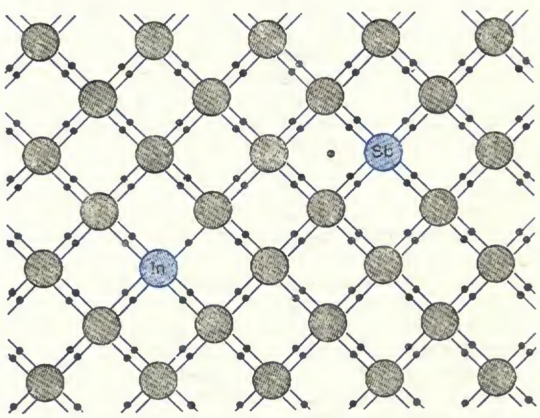


Rys. 7. Struktura krystaliczna germanu i krzemu. Czarne kulki obrazują atomy, a łączące je pręty — wiązania chemiczne

z punktu widzenia krystalochemii powiedzielibyśmy, że cztery zewnętrzne elektrony wchodzą w skład wiązań chemicznych, zapewniających spójność całego kryształu. Para elektronów — po jednym z dwu sąsiednich atomów — wytwarza tzw. wiązanie kowalencyjne. Ten chemiczny punkt widzenia jest pozornie całkowicie sprzeczny z uprzednio wprowadzonym obrazem prawie swobodnych elektronów. Ale tak nie jest. Prawdopodobieństwo znalezienia elektronu walencyjnego jest rzeczywiście największe w „mostku" łączącym dwa sąsiednie atomy. Jednak elektrony te, rozpatrywane w dużych obszarach kryształu, zachowują się właśnie tak jak fale — fale prawdopodobieństwa. Ten zadziwiający fakt, że to samo zjawisko można opisać na różne, często pozornie przeciwstawne sposoby, jest bardzo powszechny w fizyce.

domieszki w germanie

Jeśli teraz w miejsce jednego z czterowartościowych atomów Ge w krysztale umieścimy jakiś obcy atom o pięciu elektronach w zewnętrznej powłoce (takimi są atomy pierwiastków piątej grupy tablicy Mendelejewa, np. antymon), to wbuduje on się w sieć krystaliczną wykorzystując do wiązań z sąsiadami tylko cztery ze swoich elektronów (rys. 8). Pozostały,



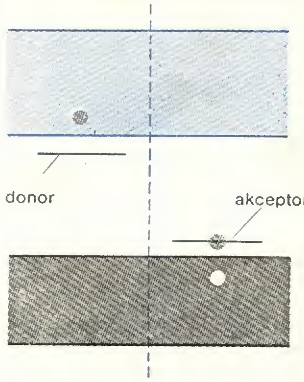
Rys. 8. Dwuwymiarowy schemat kryształu germanu (Ge) z wbudowanymi obcymi atomami indu (In) i antymonu (Sb). Czarne kółka oznaczają jony germanu, kropki — elektrony walencyjne

piąty elektron, pozostaje bardzo słabo związany z atomem domieszki i może być stosunkowo łatwo od niej oderwany i przeniesiony do pasma przewodnictwa, gdzie zachowuje się oczywiście jak zwykły elektron

donory przewodnictwa. Taką domieszkę nazywamy donorem. Analogicznie, jeśli domieszką jest atom trzeciej grupy, mający trzy zewnętrzne elektrony, np. ind, to czwarty elektron kosztem niewielkiej energii może być — dla skompletowania wiązań — uzupełniony z pasma walencyjnego. W pasmie walencyjnym pozostaje

akceptory wówczas przewodząca dziura. Tego typu domieszka nazywa się akceptorem.

Po oderwaniu lub przyłączeniu elektronu, atom domieszki — początkowo neutralny — staje się dodatnio lub ujemnie naładowanym jonem. W schemacie pasm energetycznych stan kwantowy zlokalizowanego elektronu domieszki obrazujemy rysując kreskę — po-



Rys. 9. Donory i akceptory w półprzewodniku

ziom energetyczny — wewnątrz przerwy wzbronionej. Dla atomu antymonu poziom ten leży bardzo blisko pasma przewodnictwa, ok. 0,01 eV poniżej jego dna. Nawet w temperaturze −200°C prawie wszystkie donory Sb ulegają zjonizowaniu, oddając swoje elektrony do pasma przewodnictwa. German domieszkowany antymonem wykazuje przewodnictwo elektronowe lub, jak mówimy, jest materiałem typu *n*. Poziom typowego akceptora, indu, leży 0,01 eV ponad dnem pasma walencyjnego. Elektrony z pasma walencyjnego obsadzając poziomy akceptorowe, pozostawiają po sobie dziury. Materiał domieszkowany indem ma zatem przewodnictwo dziurowe (lub inaczej przewodnictwo typu *p*), jest więc materiałem typu *p*.

przewodnictwo elektronowe

przewodnictwo dziurowe

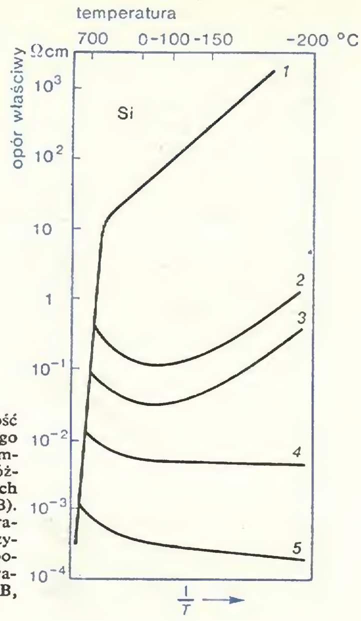
Wnioski, które tu przedstawiliśmy są na ogół słuszne dla wszystkich półprzewodników, z tym że w różnych materiałach rolę donorów i akceptorów mogą spełniać różne domieszki. Co więcej, tę samą rolę mogą odgrywać również istniejące zawsze niedoskonałości w budowie krystalicznej półprzewodnika, czyli defekty sieci (→ Budowa kryształów). Nie zawsze jednak poziomy energetyczne domieszek lub defektów leżą tak blisko pasm dozwolonej energii. Wewnątrz przerwy energetycznej mogą także występować głębokie poziomy. Są one nazywane czasem pułapkami elektronowymi, dla podkreślenia faktu, że utrudnione jest termiczne wyswobodzenie się z nich elektronu.

pułapki elektronowe

Jak czułe są własności elektryczne półprzewodników na zawartość domieszek obrazuje fakt, że dodanie np. 1 mg antymonu na 1 tonę germanu zmienia przewodnictwo elektryczne kryształu na tyle, że decyduje to o jego dalszych zastosowaniach technicznych. Kariera półprzewodników opiera się przede wszystkim na dwu wielkich osiągnięciach technologicznych. Pierwszym z nich było wynalezienie przez W. Pfanna sposobu oczyszczania kryształów półprzewodnikowych do takiego stopnia czystości, jaki nigdy przedtem nie był osiągalny dla żadnych materiałów. Sposób ten polega na miejscowym stopieniu pałeczki półprzewodnikowej i następnym wielokrotnym przeciąganiu stopionej strefy wzdłuż całej długości kryształu. Domieszki, które wolą gromadzić się w fazie ciekłej, zostają w ten sposób wyprowadzone na brzeg kryształu.

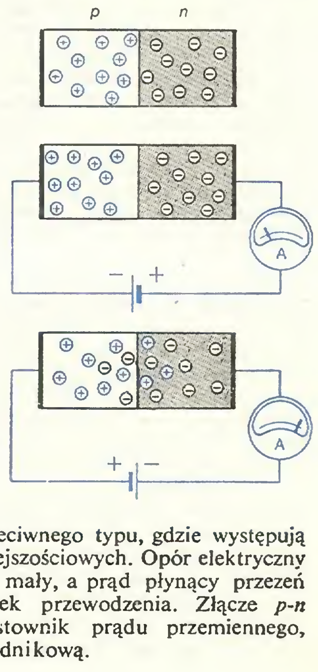
Drugim osiągnięciem było wytworzenie w tak oczyszczonym materiale dwóch sąsiadujących ze sobą obszarów o różnym typie przewodnictwa, czyli otrzymanie złącza *p-n*. Tak więc w jednym kawałku pół-

złącze *p-n*



Rys. 10. Zależność oporu właściwego krzemu (Si) od temperatury, przy różnych zawartościach domieszki boru (B). Ilość domieszki wzrasta z numerem krzywej. Krzywa *1* odpowiada 10^{-7}% wprowadzonych atomów B, krzywa *5* — 10^{-2}%

przewodnika możemy mieć dwa względnie ostro rozgraniczone między sobą obszary, w jednym z nich dominującymi, czyli większościowymi nośnikami prądu są elektrony, w drugim zaś — dziury. Natężenie prądu, który płynie przez tego rodzaju złącze, zależy w sposób zasadniczy od kierunku przyłożonego napięcia. Jeśli kierunek polaryzacji jest taki, że ujemny biegun baterii połączony jest z obszarem typu *p*, to zarówno dziury, jak i elektrony w swoich macierzystych obszarach są odciągane przez pole elektryczne z obszaru złącza. W pobliżu granicy *p-n* wytwarza się warstwa zubożona w nośniki prądu i prąd płynący przez złącze jest bardzo słaby. Jest to zatem kierunek zaporowy. Natomiast przy przeciwnej polaryzacji dziury i elektrony są spychane w kierunku złącza; niewielka ich część zostaje nawet „wstrzyknię-



Rys. 11. Złącze *p-n* przy dwóch kierunkach polaryzacji

ta" do obszarów przeciwnego typu, gdzie występują w roli nośników mniejszościowych. Opór elektryczny złącza jest wówczas mały, a prąd płynący przezeń duży; jest to kierunek przewodzenia. Złącze *p-n* stanowi zatem prostownik prądu przemiennego, tzn. diodę półprzewodnikową.

Ten mocno uproszczony model złącza *p-n*, który tu przedstawiliśmy, ma wartość raczej mnemotechniczną. Opisuje on prawidłowo kierunek prostowania, a nawet uwidacznia wystąpienie nowego ważnego zjawiska wstrzykiwania nośników mniejszościowych. W rzeczywistości warstwa zubożona w nośniki istnieje na złączu nawet bez przyłożenia doń zewnętrznego napięcia, co jest wynikiem pojawienia się pewnej samoistnej różnicy potencjałów zapewniającej wewnętrzną równowagę tego, bądź co bądź, złożonego układu.

Światłoczułość półprzewodników

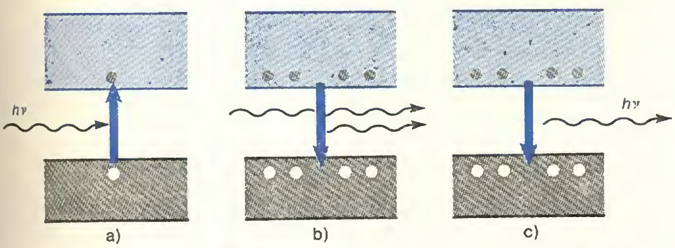
Wstrzykiwanie za pomocą złącza *p-n* jest jednym ze sposobów chwilowego zwiększenia koncentracji nośników prądu w półprzewodniku. Spośród kilku innych znanych sposobów najważniejszym jest oświetlenie kryształu. Jaki jest mechanizm oddziaływania światła, lub ogólniej promieniowania elektromagnetycznego, na kryształ półprzewodnikowy? Najbardziej podstawowym procesem jest wymuszanie przez falę elektromagnetyczną przejść elektronu z jednego pasma dozwolonej energii do innego. Zjawisko to ma charakter rezonansowy i zachodzi tylko wtedy, gdy częstość ν padającej fali jest dokładnie równa różnicy energii E_2-E_1 stanów kwantowych, między którymi następuje przejście, podzielonej przez stałą Plancka h. Warunek ten zapisuje się w postaci

$$h\nu = E_2-E_1.$$

$h\nu$ jest tu energią najmniejszej porcji promieniowania o danej częstości, tj. fotonu.

absorpcja

Jeśli elektron znajduje się początkowo w stanie kwantowym o niższej energii — taka sytuacja jest charakterystyczna dla równowagi cieplnej — wówczas wynikiem oddziaływania z falą elektromagnetyczną będzie wzbudzenie go do stanu o wyższej energii. Mówimy wtedy, że foton został pochłonięty (zaabsorbowany) a jego energia przekazana elektronowi (rys. 12a). Jeśliby zaś elektron znajdował się początkowo



Rys. 12. Rodzaje przejść elektronowych z udziałem fali elektromagnetycznej: a) absorpcja fotonu, b) emisja wymuszona, c) emisja spontaniczna

emisja wymuszona

na wyższym poziomie energetycznym (a tego rodzaju nietrwały stan można wytworzyć), to wówczas następowałoby przejście rezonansowe na poziom niższy, a uwolniona przy tym energia byłaby wypromieniowana w postaci fotonu (rys. 12b). Zjawisko to, zwane emisją wymuszoną, jest podstawą działania laserów. Oczywiście, tego rodzaju przejście elektronu z góry na dół może zajść także samorzutnie, bez obecności zewnętrznego promieniowania (emisja spontaniczna, rys. 12c).

Jeśli chodzi o półprzewodniki, to przejścia rezonansowe zachodzą przede wszystkim między stanami kwantowymi pasma walencyjnego i pasma przewodnictwa. Wystąpią one zatem nie przy ściśle określonej długości fali, ale w pewnym pasmie długości fal. Największa długość fali, przy której pojawi się jeszcze rezonansowa absorpcja promieniowania, zależy oczywiście od szerokości przerwy wzbronionej danego półprzewodnika. Jeśli zapamiętamy prostą relację między długością fali w μm a energią fotonu w eV: $\lambda = 1{,}24/h\nu$, to dla każdego materiału o znanej przerwie wzbronionej E_g, będziemy mogli obliczyć, kładąc

progowa długość fali

$h\nu = E_g$, jaka jest progowa długość fali, przy której rozpoczyna się silna absorpcja światła. Dla krzemu na przykład, dla którego $E_g = 1{,}1$ eV, wynosi ona 1,1 μm. Krzem jest więc zupełnie nieprzezroczysty dla światła widzialnego. Rzeczywiście, płytka monokrystaliczna Si ma wygląd metaliczny. Jest rzeczą zdumiewającą, że taka płytka nie stanowi prawie żadnej przeszkody dla promieniowania cieplnego, tzn. podczerwonego o długości fali większej od 1,1 μm. Natomiast czysty ZnO, który ma przerwę 3,2 eV, jest przezroczysty niemal jak szyba i silnie absorbuje dopiero promieniowanie nadfioletowe.

fonon

W niektórych półprzewodnikach, takich jak Ge i Si, stany kwantowe odpowiadające dnu pasma przewodnictwa oraz wierzchołkowi pasma walencyjnego odpowiadają zupełnie różnym wartościom wektora $\vec{k}$. Przejścia kwantowe między tymi stanami muszą wówczas zachodzić ze zmianą pędu krystalicznego elektronu. Ale we wszystkich zjawiskach fizycznych prawa zachowania pędu i energii muszą być dokładnie spełnione. Foton padającego promieniowania niesie ze sobą tak mały pęd, że można go w ogóle nie uwzględniać w tym procesie. Jak wobec tego zbilansować tę różnicę pędów? Na pomoc przychodzi tu dodatkowa kwazicząstka, tzw. fonon, czyli kwant drgań sprężystych sieci krystalicznej (→ Dynamika sieci krystalicznej). W procesie trójcząstkowym, w którym biorą udział foton, elektron i emitowany bądź absorbowany fonon, prawa zachowania mogą być już spełnione. Jednakże taki proces jest mniej prawdopodobny niż proces dwucząstkowy i związana z nim absorpcja, a także emisja promieniowania są dużo słabsze.

półprzewodniki z wąską przerwą

Jest rzeczą zupełnie oczywistą, że drgania sieci — fonony — będą odgrywać zasadniczą rolę również w zjawisku przewodnictwa elektrycznego i innych zjawiskach transportu elektronowego w półprzewodnikach. Ograniczają one przede wszystkim drogę swobodnego przebiegu elektronu, a tym samym wartość ruchliwości nośników prądu. W bardzo niskich temperaturach, gdy drgania sieci są mało intensywne, ruchliwość nośników ograniczona jest raczej rozpraszaniem na domieszkach i defektach. Dla większości półprzewodników ruchliwość nie przekracza zazwyczaj w normalnych warunkach wartości kilku tysięcy cm^2/V s. W pewnych jednak materiałach z bardzo wąską przerwą energetyczną ruchliwość elektronów osiąga monstrualnie duże wielkości.

Weźmy przykład kryształu będącego stopem dwóch związków półprzewodnikowych CdTe i HgTe. Szerokość przerwy wzbronionej tego materiału zależy silnie od procentowej zawartości obu składników i dla stopu, w którym 10% całego materiału stanowi CdTe, przerwa energetyczna całkowicie się zamyka. Krzywizna pasma przewodnictwa w pobliżu jego dna staje się wówczas bardzo duża i masa efektywna elektronu osiąga wartość minimalną. Powoduje to anomalny wzrost ruchliwości elektronów, która w niskich temperaturach wyraża się liczbą sześciocyfrową.

poziomy Landaua

Materiały o tak wielkich ruchliwościach mają osobliwe własności w polu magnetycznym. Elektron, którego droga swobodnego przebiegu jest wówczas bardzo długa, porusza się na skutek działania siły Lorentza po orbitach kołowych leżących w płaszczyźnie prostopadłej do kierunku pola magnetycznego. W myśl mechaniki kwantowej orbity te muszą być skwantowane, czyli odpowiadać ściśle określonym nieciągłym wartościom energii. Następuje tu zatem pewnego rodzaju rozszczepienie pasma przewodnictwa na poszczególne poziomy kwantowe (tzw. poziomy Landaua). Stają się wtedy możliwe przejścia rezonansowe elektronu między rozszczepionymi poziomami tego samego pasma. Energie tych przejść można w szerokim zakresie regulować polem magnetycznym. Leżą już one w obszarze długości fal odpowiadających dalekiej podczerwieni. Stwarza to interesujące możliwości nowych zastosowań półprzewodników z wąską przerwą energetyczną, np. w przyrządach optoelektronicznych.

Wiemy już, że w wyniku absorpcji światła następuje wzbudzenie elektronów z pasma walencyjnego do pasma przewodnictwa; rezultatem jest pojawienie się dodatkowych nośników prądu — elektronów i dziur. Tak więc oświetlenie półprzewodnika odpowiednio krótkofalowym światłem zwiększa jego przewodnictwo elektryczne, co jest określane mianem fotoprzewodnictwa. Zjawisko to jest szeroko wykorzystywane do wykrywania i pomiaru promieniowania elektromagnetycznego.

fotoprzewodnictwo

Po wyłączeniu oświetlenia wzbudzone światłem nośniki prądu zanikają z czasem, aż do całkowitego przywrócenia stanu początkowego. Ten proces zaniku, zwany rekombinacją, może następować na przykład wskutek samorzutnych przejść elektronów z pasma przewodnictwa do pasma walencyjnego z emisją fotonów. W niektórych półprzewodnikach, takich jak GaAs, ten sposób powrotu do równowagi jest dość uprzywilejowany. Jeśli dodatkowe pary elektron-dziura są wprowadzane do materiału przez wstrzykiwanie złączem *p-n* a ich rekombinacja zachodzi z emisją promieniowania, to wówczas mamy do czynienia z bezpośrednią przemianą energii elektrycznej w światło, zwaną elektroluminescencją (→ Optoelektronika półprzewodnikowa). Wydajność takiego świecenia ograniczona jest konkurencyjnym mechanizmem rekombinacji, w którym uwolniona energia przekazywana jest drganiom sieci. Ten ostatni proces zachodzi szczególnie chętnie na wszelkiego rodzaju defektach sieci i domieszkach i on to zazwyczaj ogranicza długość czasu życia wzbudzonych nośników prądu.

rekombinacja

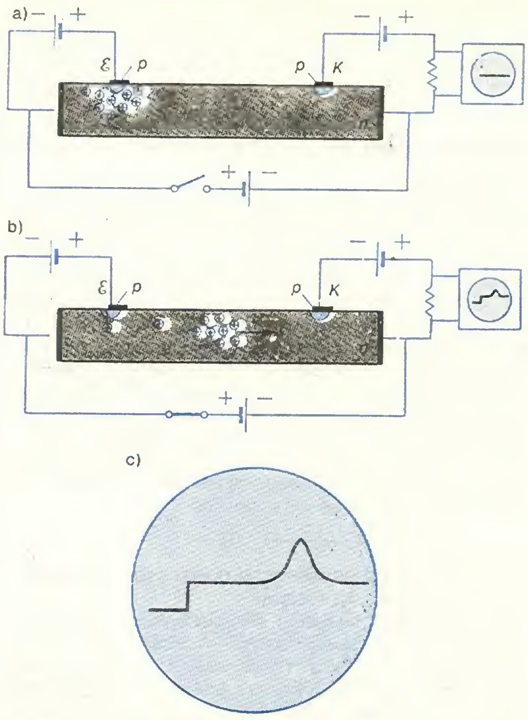
elektroluminescencja

Od tranzystora do generatora Gunna

Swoiste właściwości materiałów półprzewodnikowych uwidaczniają się wyraziście w pewnym, klasycznym już dziś doświadczeniu zaproponowanym przez Haynesa i W. Shockleya z laboratorium Bell Telephone. Opiszemy je tutaj w wersji różniącej się nieco od oryginału.

Na próbce germanowej — powiedzmy, że jest ona typu *n* — wytworzono w pobliżu jej końców dwa maleńkie obszary typu *p*. Jedno z utworzonych w ten sposób złącz spolaryzowano w kierunku przewodzenia, drugie w kierunku zaporowym. To pierwsze złącze nazywamy emiterem, ponieważ wstrzykuje ono do próbki nośniki mniejszościowe — dziury. Drugie złącze ma całkiem odmienną właściwość. Może ono „wysysać" dziury z obszaru kryształu bezpośrednio doń przylegającego. Dzieje się tak dlatego, że złącze spolaryzowane zaporowo nie stanowi przeszkody dla nośników mniejszościowych. Wstrzykiwane przez emiter dziury rozprzestrzeniają się w głąb próbki dzięki procesowi dyfuzji, zanikając po drodze na skutek rekombinacji. Ze względu na stosunkowo krótki czas życia nadmiarowych nośników, efektywny zasięg chmury dziurowej nie przekracza zazwyczaj milimetra. Jeśli odległość emiter–kolektor wynosi np. 1 cm, to wstrzyknięte dziury nie docierają do kolektora. Można im jednak w tym pomóc, przykładając w pewnej chwili do kontaktów na końcach próbki taką różnicę potencjałów, która będzie je ciągnąć w kierunku kolektora. Chmura nadmiarowych nośników zostaje teraz oderwana od emitera i uniesiona polem elektrycznym wzdłuż próbki. Jeśli natężenie pola elektrycznego w krysztale wynosi, powiedzmy, 50 V/cm, to biorąc pod uwagę ruchliwość dziur w Ge — 1800 $cm^2/V\,s$ — znajdujemy, że prędkość unoszenia wynosi $9\cdot10^4$ cm/s. Dziury przebędą zatem odległość emiter–kolektor w czasie równym w przybliżeniu 10^{-5} s, jest to czas znacznie krótszy od czasu życia nadmiarowych nośników prądu w dobrym krysztale. Tak więc prawie wszystkie dziury dotrą do obszaru kolektora i tu zostaną przezeń wessane, przez co zwiększy się chwilowy prąd w obwodzie kolektora. Jeśli na ekranie oscyloskopu będziemy rejestrować sygnał elektryczny kolektora, to zaobserwujemy jego przebieg w czasie, taki jak na rys. 13c. Pierwszy skok sygnału pojawia się w momencie włączenia pola ciągnącego i jest wywołany spadkiem napięcia wzdłuż długości kryształu. Po pewnym czasie pojawia się rozmyty sygnał związany z dotarciem chmury dziurowej do kolektora.

zasada działania tranzystora



Rys. 13. Doświadczenie Haynesa–Shockleya demonstrujące ruch chmury dziurowej w polu elektrycznym

Eksperyment ten wykazuje w sposób bezpośredni istnienie dodatnio naładowanych nośników prądu, tj. dziur i pozwala wyznaczyć ich ruchliwość a nawet czas życia. Demonstruje on również zjawisko wstrzykiwania nośników prądu oraz działanie kolektora. Stąd już tylko krok do zrozumienia zasady działania tranzystora. Jeśli umieścić emiter i kolektor w dostatecznie małej odległości od siebie, to wówczas, nawet bez pola unoszącego nośniki, znaczna część wstrzykniętych dziur dotrze do kolektora. Taki układ, który stanowi uproszczony model tranzystora, ma własność wzmacniania sygnałów elektrycznych. Sygnał wprowadzony do obwodu emitera jest przenoszony za pośrednictwem nośników mniejszościowych do obwodu kolektora. Ale napięcie przyłożone do kolektora może znacznie przewyższać napięcie na emiterze; stąd — wzmocnienie mocy przenoszonego sygnału.

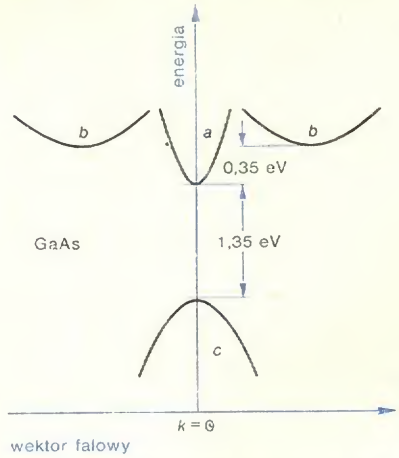
Przyjmowaliśmy dotychczas, że zależność między gęstością prądu i natężeniem pola elektrycznego w jednorodnym półprzewodniku jest liniowa, tzn. spełnia prawo Ohma. Jednakże, to prawo przestaje być słuszne dla bardzo silnych pól, gdy prędkość unoszenia elektronów staje się porównywalna ze średnią prędkością ich chaotycznego ruchu cieplnego. Występuje wówczas zjawisko grzania gazu elektronowego, któremu towarzyszy zazwyczaj wzrost rozpraszania i zmniejszenie ruchliwości.

gorące elektrony

Istnienie „gorących" elektronów może prowadzić w pewnych wypadkach do pojawienia się zupełnie nowych i niezwykłych zjawisk, np. pojawienie się ujemnej przewodności elektrycznej w GaAs. Struktura pasmowa tego półprzewodnika przedstawiona jest schematycznie na rys. 14. Pasmo przewodnictwa ma bezwzględne minimum (*a*) w środku strefy Brillouina, tzn. dla $\vec{k} = 0$. To minimum odległe jest od krawędzi pasma walencyjnego (*c*) o 1,35 eV. W pasmie przewodnictwa występują ponadto dodatkowe lokalne

ujemna przewodność elektryczna

minima (b), odpowiadające pewnym określonym kierunkom i wartościom wektora $\vec{k}$ elektronu. Te lokalne minima leżą 0,35 eV powyżej krawędzi pasma przewodnictwa.



Rys. 14. Schemat struktury elektronowej arsenku galu GaAs: a, b — minima pasma przewodnictwa, c — pasmo walencyjne

W domieszkowanym GaAs typu n elektrony, które biorą udział w przewodzeniu prądu, znajdują się, w normalnych warunkach, w najniższym, środkowym minimum pasma przewodnictwa. W silnych polach elektrycznych elektrony uzyskują dodatkową energię kinetyczną. Jeśli przyłożone pole zwiększy średnią energię kinetyczną elektronów powyżej wartości 0,35 eV, wówczas następuje obsadzenie elektronami bocznych minimów pasma przewodnictwa. Ale pojemność bocznych minimów lub, mówiąc ściślej, gęstość stanów kwantowych w tych minimach jest znacznie większa niż w minimum środkowym, wskutek różnicy ich krzywizn. A zatem, chociaż prawdopodobieństwo obsadzenia stanów w bocznych minimach pozostaje — ze względu na statystykę — mniejsze niż w środkowym minimum, to jednak w warunkach silnego pola elektrycznego może znaleźć się w nich przeważająca część elektronów przewodnictwa. Z drugiej strony mniejsza krzywizna bocznych minimów powoduje, że masa efektywna m^* znajdujących się tam elektronów jest duża, a zatem ich ruchliwość $\mu = e\tau/m^*$ jest mała. Mamy już wszystkie elementy niezbędne do tego, aby zrozumieć istotę pojawiającego się w GaAs zjawiska ujemnej przewodności elektrycznej. Mechanizm tego efektu można teraz lapidarnie objaśnić w następujący sposób. W zakresie słabych pól elektrycznych większość elektronów znajduje się w środkowym minimum i wtedy obserwuje się normalne, niezależne od pola przewodnictwo kryształu. Gdy pole jest dostatecznie silne, rozpoczyna się obsadzanie bocznych minimów, w których ruchliwość elektronów jest mała. Ze wzrostem pola elektrycznego coraz większa część elektronów przechodzi do minimów o małej ruchliwości. W tym zakresie prąd płynący przez kryształ maleje ze wzrostem przyłożonego do kryształu napięcia.

efekt Gunna

Taka sytuacja jest wielce nienormalna i prowadzi do tego, że pole elektryczne wewnątrz kryształu przestaje być jednorodne. Tworzą się obszary (domeny) silnego i słabego pola, ostro od siebie odgraniczone. Ich granice przemieszczają się wzdłuż kryształu zgodnie z unoszeniem elektronów przez pole elektryczne. Docierając do końców kryształu domeny wywołują w zewnętrznym obwodzie oscylacje prądu. Jeśli długość kryształu jest niewielka, to częstość tych oscylacji jest bardzo duża. Zjawisko to, zwane efektem Gunna, jest wykorzystywane do generacji fal elektromagnetycznych o częstości mikrofalowej (→ Generacja mikrofal).

materiały półprzewodnikowe

Dla wygody Czytelnika załączamy krótki wykaz podstawowych materiałów półprzewodnikowych wraz z podaniem ich najważniejszych zastosowań. O bardzo wielu z nich w ogóle nie wspominaliśmy w tym artykule.

Niektóre półprzewodniki

Materiał	Przerwa energetyczna, eV	Zastosowanie
InSb	0,18	przetworniki magnetoelektryczne, detektory podczerwieni
Ge	0,67	diody, tranzystory, fotodiody
Si	1,1	diody, tranzystory, baterie słoneczne
GaAs	1,35	diody elektroluminescencyjne, lasery, generatory mikrofal
CdTe	1,5	detektory promieniowania jądrowego
Cu_2O	2,1	prostowniki
ZnO	3,2	luminofory
CdTe-HgTe	0–1,5	detektory promieniowania podczerwonego

Fizyka półprzewodników osiągnęła obecnie swój okres dojrzałości. Pomimo tego, ciągle prowadzi się intensywne poszukiwania zarówno nowych zjawisk, jak i nowych materiałów. Istnieją obecnie trzy grupy substancji, które przyciągają szczególną uwagę badaczy. Są to tzw. półprzewodniki magnetyczne, półprzewodniki amorficzne (niekrystaliczne) oraz półprzewodnikowe związki organiczne.

N. B. Hannay *Półprzewodniki*, Warszawa 1962; P. S. Kiriejew *Fizyka półrzewodników*, Warszawa 1971; R. A. Smith *Półprzewodniki*, Warszawa 1966.

Półprzewodniki magnetyczne

Robert R. Gałązka

Wszystkie ciała stałe występujące w przyrodzie można podzielić na kilka grup, przy czym kryterium podziału stanowią ich własności fizyczne. Półprzewodniki i magnetyki tworzą dwie duże grupy ciał stałych o wyraźnie różnych własnościach fizycznych. W ostatnim dziesięcioleciu zaszła potrzeba wyróżnienia nowej grupy materiałów, które nazwano półprzewodnikami magnetycznymi. Półprzewodniki magnetyczne można zdefiniować w różny sposób, np. jako:

określenie półprzewodnika magnetycznego

1) magnetyki (materiały magnetyczne) o przewodności elektrycznej od 10^4 do 10^{-10} $(\Omega cm)^{-1}$;

2) materiały zawierające jednocześnie swobodne nośniki prądu oraz zlokalizowane atomowe momenty magnetyczne. Koncentracja nośników prądu może się zmieniać pod wpływem warunków zewnętrznych, takich jak oświetlenie, temperatura, ciśnienie, pole elektryczne i magnetyczne.

Materiałami magnetycznymi mogą być zarówno metale jak i dielektryki — izolatory. Istnieje również duża grupa magnetyków o przewodności pośredniej, pomiędzy metalami i izolatorami, i właśnie tę grupę materiałów nazwano półprzewodnikami magnetycznymi. Warto zwrócić uwagę na to, że pierwsza definicja charakteryzuje materiał niejako od zewnątrz, poprzez makroskopowy parametr, jakim jest przewodność właściwa.

Definicja druga jest głębsza — charakteryzuje materiał od strony mikroskopowej. Magnetyczny moment pędu atomu składa się z części orbitalnej (związanej z orbitalnym ruchem elektronów wokół jądra)

oraz części spinowej (→ Teoria magnetyzmu). Podobnie jak ładunek elektryczny, np. elektronu, charakteryzuje pole elektryczne wytwarzane przez elektron, tak spin charakteryzuje pole magnetyczne elektronu. Ponieważ moment orbitalny jest wielkością małą, o momencie magnetycznym atomu, o magnetycznych własnościach pierwiastków i związków chemicznych decyduje spin atomu i oddziaływania spinowe.

W układzie okresowym istnieją dwie grupy pierwiastków, których atomy mają szczególnie duży wypadkowy spin. Są to tzw. metale przejściowe (np. Mn, Fe, Ni, Cr) oraz pierwiastki ziem rzadkich (np. Eu, Gd, La). Duża wartość spinu tych atomów związana jest z istnieniem niezapełnionych powłok elektronowych *d* i *f* odpowiednio dla metali przejściowych i pierwiastków ziem rzadkich. Mówiąc w drugiej definicji o zlokalizowanych momentach magnetycznych, mieliśmy na myśli właśnie obecność tych pierwiastków w materiale.

Możemy teraz uzupełnić definicje półprzewodników magnetycznych dodając, że każdy półprzewodnik magnetyczny musi zawierać w postaci domieszki lub składnika co najmniej 1 pierwiastek z grupy metali przejściowych lub grupy pierwiastków ziem rzadkich.

własności fizyczne

Własności fizyczne półprzewodników magnetycznych w dużej części są nowe i nietypowe zarówno dla magnetyków, jak i półprzewodników. Dotychczas w półprzewodnikach nie uwzględniano oddziaływań spinowych (normalne półprzewodniki nie zawierają pierwiastków magnetycznych). Fizyka magnetyków nie brała pod uwagę wpływu swobodnych nośników prądu na oddziaływania spinowe (teorie opracowane dla metali magnetycznych ze względu na bardzo dużą i stałą koncentrację nośników prądu nie stosują się do półprzewodników magnetycznych). Półprzewodniki magnetyczne budzą duże zainteresowanie tak z punktu widzenia poznawczego (możliwość obserwacji nowych efektów fizycznych) jak i zastosowań (możliwość otrzymania półprzewodników o regulowanych własnościach magnetycznych — np. przez zmianę zawartości atomów magnetycznych, pole magnetyczne, temperaturę, oraz magnetyków o różnych koncentracjach swobodnych nośników — elektronów i dziur).

Ze względu na nowość problematyki, jak również duże trudności w matematycznym opisie półprzewodników magnetycznych, wiele istniejących teorii, które próbują wyjaśnić własności tych materiałów, ma raczej charakter roboczych hipotez. Zrozumienie zjawisk zachodzących w półprzewodnikach magnetycznych szybko się pogłębia, ale jeszcze nie może być porównywane z wiedzą i aparatem matematycznym, jakimi dysponuje fizyka półprzewodników i fizyka magnetyków.

Struktura energetyczna — gęstość stanów

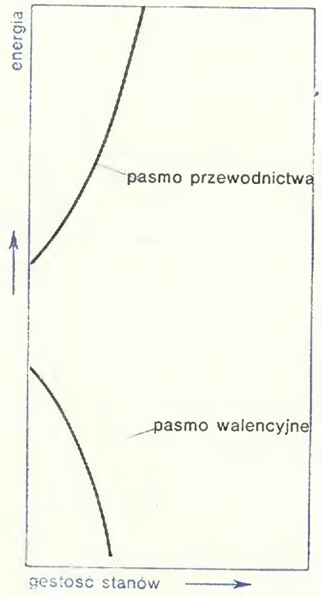
Elektrony w atomie zajmują określone stany energetyczne. Zmiana energii elektronu (np. przez oświetlenie atomu — doprowadzenie do elektronu energii fotonu fali elektromagnetycznej) nie może zachodzić w sposób ciągły. Elektron przeskakuje na wyższy stan energetyczny pochłaniając energię lub spada na niższy stan energetyczny oddając energię. Istnieje więc wiele dozwolonych dla elektronów stanów — poziomów energetycznych, a wszystkie stany pośrednich energii są dla elektronu zabronione (→ Struktura elektronowa ciał stałych).

pasma energetyczne w ciałach stałych

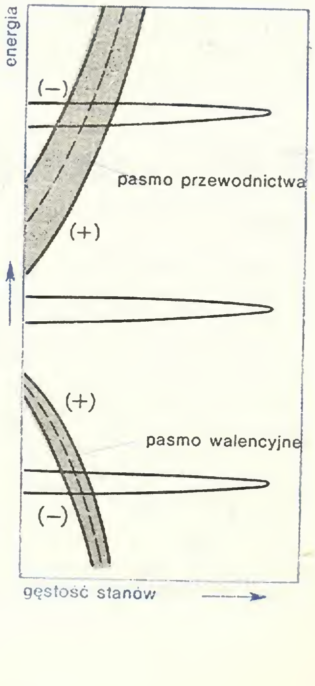
W krystalicznych ciałach stałych, np. w półprzewodnikach, rolę pojedynczych stanów spełniają pasma energetyczne. Trzy pasma decydują o własnościach półprzewodnika: pasmo podstawowe — walencyjne, pasmo energii wzbronionej (przerwa energetyczna) i pasmo przewodnictwa (→ Półprzewodniki). W rozważaniach nad półprzewodnikami magnetycznymi bardzo ważnym parametrem jest gęstość stanów, czyli ilość stanów energetycznych na jednostkę energii. Jeżeli elektron na skutek zderzeń z innymi elektronami, czy też rozpraszania przez drgające atomy sieci zmienia swoją energię, gęstość stanów dostarcza nam informacji, ile stanów energetycznych będzie miał do dyspozycji — jak łatwo lub trudno będzie się poddawał zmianom energii. W typowym półprzewodniku gęstość stanów energetycznych rośnie monotonicznie z energią elektronu lub dziury (rys. 1).

gęstość stanów

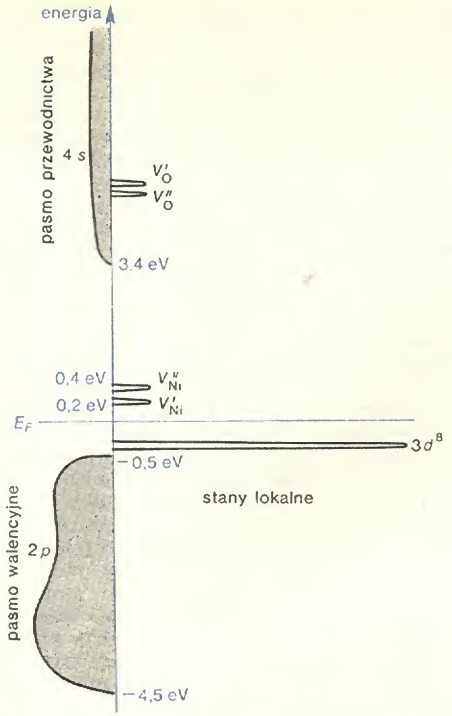
Popatrzmy teraz na rys. 2 i 3, pokazujące strukturę gęstości stanów w półprzewodnikach magnetycznych. Widzimy wyraźną różnicę. Pojawiają się dodatkowe piki gęstości stanów w pewnych wąskich przedziałach energii, co więcej — pojawiają się dozwolone stany w pasmie energii wzbronionej. Pasmo przewodnictwa i pasmo walencyjne rozdzielają się na dwa, komplikując dodatkowo sytuację (fakt rozdzielenia się pasm nie występuje zawsze — jest typowy tylko dla ferromagnetyków i tylko dla pewnych kierunków ruchu elektronów w krysztale). Za taką komplikację obrazu gęstości stanów odpowiedzialne są oczywiście atomy „magnetyczne". Dodatkowe piki gęstości stanów związane są właśnie z elektronami *d* lub *f* tych atomów. Rozszczepianie się pasm związane jest natomiast



Rys. 1. Zależność gęstości stanów energetycznych od energii w pasmie walencyjnym i pasmie przewodnictwa typowego półprzewodnika (gęstość stanów w pasmie energii wzbronionej równa jest oczywiście zeru)



Rys. 2. Schematyczny obraz gęstości stanów w zależności od energii dla półprzewodnika ferromagnetycznego; (+) i (−) oznaczają przeciwne kierunki spinów w rozszczepionych pasmach przewodnictwa i walencyjnym



Rys. 3. Postulowana struktura gęstości stanów w półprzewodniku magnetycznym NiO; E_F — poziom Fermiego

z sumarycznym polem magnetycznym wytworzonym przez te atomy. Zwróćmy uwagę na jeszcze jeden fakt komplikujący obraz. W dostatecznie niskiej temperaturze, gdy cieplne drgania atomów są małe, siły oddziaływań spinowych pomiędzy atomami powodują powstanie porządku magnetycznego. Spiny kolejnych atomów „magnetycznych" mogą się dodawać jak w ferromagnetykach lub odejmować ja w antyferromagnetykach. W ferromagnetycznym półprzewodniku powstaje silne wewnętrzne pole magnetyczne, którego skutkiem jest rozszczepienie się pasma przewodnictwa i pasma walencyjnego. Tak się dzieje, gdy energia cieplnych drgań sieci krystalicznej jest dostatecznie mała w porównaniu z energią oddziaływań spinowych. Dla każdego magnetyka istnieje jednak pewna ściśle określona temperatura (temperatura krytyczna), w której drgania cieplne sieci niszczą porządek magnetyczny i wektory momentów magnetycznych ustawiają się przypadkowo. Powyżej tej temperatury nie ma już wewnętrznego pola magnetycznego i tym samym znika rozszczepienie pasm.

wewnętrzne pole magnetyczne

temperatura krytyczna

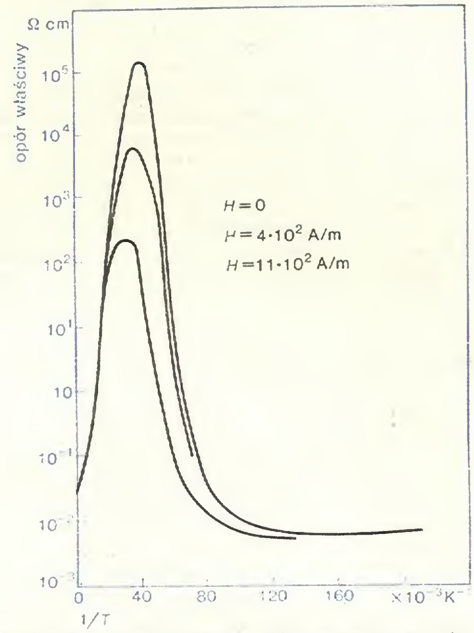
Obraz gęstości stanów w materiale wpływa zasadniczo na wszystkie własności kryształu. Wielkości fizyczne, charakteryzujące materiał, jak np. przewodność elektryczna, przewodność cieplna, siła termoelektryczna, a także własności optyczne — zależą od gęstości stanów. Można się również spodziewać, że szczególną rolę będzie odgrywała temperatura krytyczna: wszystkie własności półprzewodników magnetycznych w pobliżu tej temperatury powinny się zmieniać dość drastycznie.

Gigantyczny magnetoopór

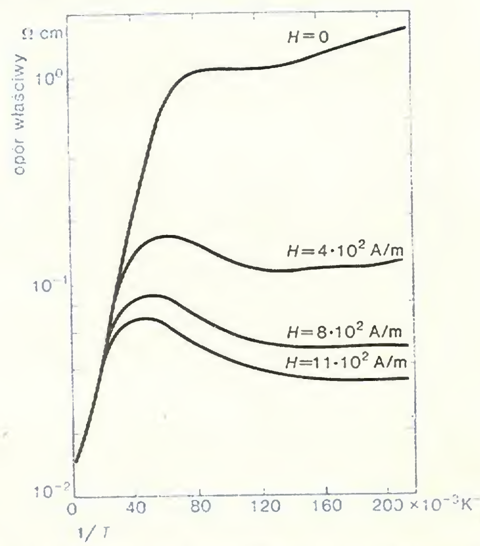
Z szerokiej klasy zjawisk transportu, a więc zjawisk związanych z przenoszeniem ładunku i ciepła omówmy tutaj tylko jedno zjawisko, a mianowicie zależność oporu właściwego półprzewodników magnetycznych od temperatury i pola magnetycznego (tzw. magnetoopór). Zachowanie się oporu w półprzewodnikach magnetycznych jest zupełnie inne niż w półprzewodnikach i metalach magnetycznych.

opór — zależność od temperatury

Na rys. 4 i 5 pokazany jest przebieg oporu właściwego dla ferromagnetycznego i antyferromagnetycznego półprzewodnika w funkcji temperatury dla kilku pól magnetycznych. W pobliżu temperatury krytycznej obserwuje się w obu wypadkach bardzo silny wzrost oporu. Jest to efekt wyjątkowo duży — opór materiału zmienia się o kilka rzędów wielkości. W miarę



Rys. 4. Zależność oporu właściwego ferromagnetycznego półprzewodnika ($Eu_{0,95}La_{0,005}S$) od temperatury T i pola magnetycznego H



Rys. 5. Zależność oporu właściwego antyferromagnetycznego półprzewodnika ($Eu_{0,80}Cd_{0,20}Te$) od temperatury T i pola magnetycznego H

wzrostu temperatury opór ferromagnetycznego półprzewodnika szybko maleje, nawet czasami poniżej wartości obserwowanej w niskich temperaturach. W antyferromagnetycznym półprzewodniku w miarę wzrostu temperatury opór próbki dalej rośnie, ale już znacznie wolniej.

Następnym niezwykłym zjawiskiem zachodzącym w pobliżu temperatury krytycznej jest zmniejszanie się oporu próbki (nawet setki razy) przy stosunkowo niewielkiej zmianie pola magnetycznego, czyli gigantyczny ujemny magnetoopór półprzewodników magnetycznych. Efekt działania pola magnetycznego

szybko słabnie w miarę oddalania się od temperatury krytycznej (szczególnie w stronę niższych temperatur).

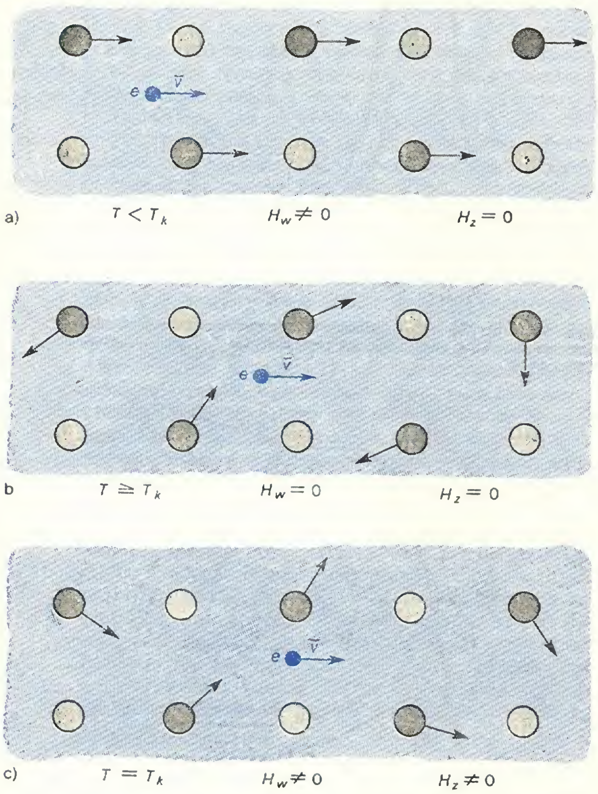
ujemny magnetoopór

Tak silnego efektu ujemnego magnetooporu nie obserwuje się w żadnych innych ciałach stałych, oprócz półprzewodników magnetycznych.

Wyjaśnienie powyższych zależności oporu elektrycznego półprzewodników magnetycznych od temperatury i pola magnetycznego jest dość złożone, postaramy się podać ogólny zarys rozumowania, prowadzący do jakościowego wyjaśnienia tych zależności.

W ferromagnetyku w niskich temperaturach istnieje silne, ukierunkowane, wewnętrzne pole magnetyczne. Elektron poruszając się wzdłuż kierunku pola magnetycznego porusza się tak, jakby pola nie było. Wynika to z podstawowych zasad ruchu elektronu w polu magnetycznym. Oczywiście, elektrony w krysztale poruszają się we wszystkich kierunkach i wskutek rozproszeń ustawicznie zmieniają swoje tory. Jednak w kierunku równoległym do pola magnetycznego rozpraszanie (związane z wewnętrznym polem magnetycznym) nie istnieje (rys. 6). W miarę wzrostu tempe-

ruch elektronu w ferromagnetyku



Rys. 6. Uproszczony schemat ferromagnetycznego kryształu; czarne kółka oznaczają atomy mające silny własny moment spinowy: a) w temperaturze niższej od krytycznej i zerowym zewnętrznym polu magnetycznym elektron poruszający się wzdłuż kierunku wewnętrznego pola magnetycznego H_w nie rozprasza się, b) w temperaturze wyższej lub równej krytycznej i w zerowym zewnętrznym polu magnetycznym elektron poruszający się w dowolnym kierunku ulega rozpraszaniu na atomach z momentem spinowym, c) w temperaturze równej krytycznej; zewnętrzne pole magnetyczne częściowo porządkuje układ spinów; elektron jest bardziej rozpraszany niż w przypadku a, ale mniej niż w przypadku b

ratury idealny porządek magnetyczny zaczyna się psuć. Pojawiają się obszary namagnesowane inaczej niż główna część kryształu. Rozpraszanie nośników prądu wzrasta, tym samym wzrasta również opór elektryczny.

Maksymalny nieporządek panuje w temperaturze krytycznej, gdy pękają sprzężenia spinowe najbliżej oddziałujących atomów. Lokalne, chaotycznie skierowane pola magnetyczne mogą być tak silne, że niemal zatrzymują poruszające się elektrony. Po zapadnięciu się ostatnich wiązań spinowych średnie wewnętrzne pole magnetyczne szybko spada do zera. Rozpraszanie maleje, opór również szybko maleje.

Przyłożenie zewnętrznego pola magnetycznego jest czynnikiem porządkującym sieć magnetyczną — przywraca zachwiany porządek magnetyczny, rozpraszanie maleje, opór również maleje — pojawia się ujemny magnetoopór. Wpływ pola magnetycznego jest więc w działaniu podobny do efektu obniżania temperatury.

Zastanówmy się teraz, dlaczego w obszarze wysokich temperatur opór ma mniejsze wartości niż obserwowane w niskich temperaturach. Otóż w ferromagnetyku, mimo że nie przykładamy zewnętrznego pola magnetycznego, wewnętrzne pole magnetyczne wywołuje samoistny magnetoopór. Pole magnetyczne w zwykłych półprzewodnikach powoduje zazwyczaj zwiększenie oporu, często bardzo znaczne. Powyżej temperatury krytycznej cieplne drgania sieci krystalicznej niszczą wewnętrzne pole magnetyczne i opór próbki może zmaleć nawet poniżej oporu obserwowanego w bardzo niskich temperaturach, ponieważ żaden samoistny magnetoopór już nie istnieje.

samoistny magnetoopór ferromagnetyków

W antyferromagnetycznym półprzewodniku w niskich temperaturach nie ma pola wewnętrznego — atomowe pola magnetyczne kompensują się nawzajem. Szybki wzrost oporu w miarę podwyższania temperatury jest spowodowany częściowym psuciem się kompensacji. Pojawiają się lokalne pola magnetyczne, a zatem wzrost rozpraszania nośników prądu i wzrost oporu. Po osiągnięciu temperatury krytycznej pole magnetyczne wewnętrzne dąży do zera wskutek nieustannych szybkich zmian kierunków momentów magnetycznych, a nie wskutek ponownej kompensacji. Z punktu widzenia pola magnetycznego wracamy do sytuacji, jaka była w temperaturach niskich, tzn. nieistnienia wewnętrznego pola magnetycznego. Opór próbki maleje jednak tylko nieznacznie, ponieważ drgania cieplne sieci krystalicznej są także silnym źródłem rozpraszania nośników prądu. Drgania te zwiększają się w miarę wzrostu temperatury i opór próbki znów zaczyna rosnąć. Warto podkreślić, że cały efekt został wytłumaczony zmianami rozpraszania nośników prądu. Nie zawsze sytuacja jest tak jasna. W miarę wzrostu temperatury może się jednocześnie zmieniać koncentracja nośników prądu, co znacznie komplikuje obraz zmian zachodzących w półprzewodnikach magnetycznych.

opór antyferromagnetyków

Ogólnie można stwierdzić, że wszystkie efekty transportu — efekty związane z ruchem nośników prądu, wykazują wiele osobliwości w pobliżu temperatury krytycznej. Anomalnym zmianom, oprócz oporu właściwego, ulegają: przewodność cieplna, siła termoelektryczna, współczynnik Halla itp. Charakterystyczna jest także zależność różnych parametrów od kierunku i rodzaju (a tym samym siły) oddziaływań magnetycznych. Ponieważ przez tworzenie stopów można zmieniać rodzaje oddziaływań spinowych, należy oczekiwać bardzo silnych zależności mierzonych parametrów od składu otrzymanych stopów.

anomalie w pobliżu temperatury krytycznej

Własności optyczne półprzewodników magnetycznych

Optyczne własności półprzewodników magnetycznych, a więc m.in. takie zjawiska, jak: absorpcja fali elektromagnetycznej w zależności od energii fali padającej, odbicie, efekt skręcania płaszczyzny polaryzacji (efekt Faradaya), również wykazują anomalie w porównaniu z podobnymi zjawiskami obserwowanymi w półprzewodnikach. Anomalie te widoczne są zarówno w kształcie obserwowanych krzywych, jak również w ich zależności od temperatury i pola magnetycznego. Interpretacja otrzymanych wyników jest na ogół bardzo trudna i wiele obserwowanych zależności nie znalazło dotychczas pełnego wyjaśnienia.

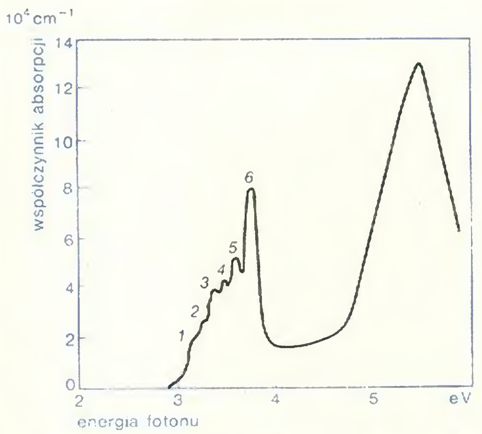
Własności optyczne kryształu stanowią odbicie mechanizmów oddziaływania materiału z falą elektro-

magnetyczną. W materiałach magnetycznych ze względu na obecność spinów, wewnętrznych pól magnetycznych, fal spinowych itp. powstają nowe formy oddziaływań fali elektromagnetycznej z kryształem, ponadto komplikują się znane efekty optyczne obserwowane w zwykłych półprzewodnikach.

absorpcja fali elektromagnetycznej

Rozpatrzmy np. jedno z najbardziej typowych zjawisk optycznych — absorpcję fali elektromagnetycznej w półprzewodniku magnetycznym. W zwykłym półprzewodniku, jeżeli energia fali padającej jest mniejsza od energii wzbronionej (rys. 1) kryształ nie absorbuje fali — jest przezroczysty dla padającego promieniowania. W półprzewodnikach magnetycznych, w przerwie energetycznej znajdują się pasma o dużej gęstości stanów związane z elektronami *d* lub *f*. Może to być jedno, jak na rys. 2, lub więcej pasm. Jeżeli energia fali elektromagnetycznej będzie równa różnicy energii między pasmem walencyjnym i pasmem *d* lub *f*, lub różnicy między *d* czy *f* a pasmem przewodnictwa (rys. 2), wówczas taka fala może być silnie absorbowana, mimo że jej energia jest znacznie mniejsza od energetycznej szerokości pasma wzbronionego. Na rys. 7 przedstawiono absorpcję we fluorku europu w zależności od energii fali padającej. Piki oznaczone numerami od *1* do *6* odpowiadają różni-

przykład: fluorek europu



Rys. 7. Widmo absorpcji dla cienkiej monokrystalicznej warstwy fluorku europu (EuF_2) w temperaturze 20 K

com energii pomiędzy pasmem walencyjnym a kolejnymi stanami elektronów powłoki *f* europu. Odpowiedni rysunek struktury gęstości stanów zawierałby więc sześć pików gęstości stanów w przerwie energetycznej, a nie jeden jak na rys. 2. Struktura widma absorpcji fluorku europu jest stosunkowo prosta i łatwa do interpretacji. Należy jednak pamiętać, że nie wszystkie możliwe do wyobrażenia przejścia optyczne są dozwolone; różnią się one ponadto natężeniem linii absorpcyjnych, regułami polaryzacji itp. Warto zauważyć, że takie czynniki, jak: natężenie linii, ich przesuwanie się w funkcji temperatury czy ciśnienia, reguły polaryzacji, są bezpośrednią wskazówką ułatwiającą identyfikację przejścia optycznego. Na rys. 7 warto zwrócić uwagę na wartość współczynnika absorpcji dla przejść *6* i *7*. Są to bardzo duże wartości obserwowane na ogół w zwykłych półprzewodnikach dla przejść międzypasmowych, co dodatkowo komplikuje eksperyment i może prowadzić do mylnej interpretacji.

Związki europu wykazują również charakterystyczną anomalię przesuwania się krawędzi absorpcji w stronę niższych energii, gdy temperatura maleje i kryształ zaczyna się porządkować magnetycznie. Podobny efekt powoduje przyłożenie zewnętrznego pola magnetycznego.

Półprzewodniki półmagnetyczne

Materiały, o których będzie mowa, mieszczą się w grupie półprzewodników magnetycznych i obie definicje podane na wstępie doskonale się do nich stosują. Nazwa półprzewodniki półmagnetyczne została użyta niejako eksperymentalnie, ponieważ będzie mowa o grupie materiałów jeszcze słabo zbadanych, ale o bardzo ciekawych własnościach. Będziemy pod tą nazwą rozumieli stopy półprzewodników z magnetykami, ale o raczej niewielkiej, najwyżej kilkunastoprocentowej zawartości materiału magnetycznego. W tego typu materiałach powinny dominować własności półprzewodnikowe, a własności magnetyczne będą istnieć niejako na drugim planie.

Dotychczas półprzewodnikami magnetycznymi nazywaliśmy magnetyki o pewnych własnościach charakterystycznych dla półprzewodników. Natomiast półprzewodniki półmagnetyczne są to półprzewodniki o pewnych własnościach charakterystycznych dla magnetyków. Materiałów o takich cechach może być bardzo wiele, ponieważ prawie każdy półprzewodnik może tworzyć stopy z różnymi magnetykami. Można sobie wyobrazić ogromną liczbę materiałów zarówno o szerokiej przerwie energetycznej i dużym oporze, jak np. stopy ZnS-MnS, CdSe-MnSe, CdTe-MnTe itd., a także materiały o małej, a nawet zerowej przerwie energetycznej i małym oporze, jak np. InSb-MnSb, HgTe-MnTe, HgTe-FeTe, HgSe-MnSe itp. W stopach mamy możliwość ciągłej i płynnej zmiany koncentracji atomów odpowiedzialnych za magnetyczne własności materiału. Daje to zarówno możliwość stopniowej zmiany własności półprzewodnika, jak również szanse regulacji tych własności. Dotychczas opublikowano niewiele prac na temat własności półprzewodników półmagnetycznych, można jednak oczekiwać jakościowo i ilościowo nowych efektów fizycznych. Przede wszystkim zewnętrzne pole magnetyczne powinno silnie wpływać na własności tych materiałów, zarówno na własności obserwowane w efektach transportu prądu, jak i efektach optycznych. Można się spodziewać nowych oddziaływań elektronów, plazmonów czy fononów z falami spinowymi, czy magnonami (→ Wzbudzenia elementarne w ciele stałym). Powinny również wystąpić pewne modyfikacje w strukturze energetycznej. Patrząc od strony magnetyków, mamy możliwość zmiany typów oddziaływań spinowych, co oczywiście jest związane bezpośrednio z koncentracją magnetyka w stopie. Otwiera się nowa dziedzina w fizyce ciała stałego, która może przynieść nie tylko głębszą wiedzę o mechanizmach istniejących oddziaływań i strukturach krystalicznych, ale również – nowe, a nawet rewelacyjne, praktyczne zastosowania tego typu materiałów.

stopy półprzewodników z magnetykami

S. Methfessel, D. C. Mattis *Magnetic Semiconductors*, Berlin 1968 (ros. Moskwa 1972).

Struktura elektronowa ciał stałych

Waldemar Gorzkowski

Badanie struktury elektronowej za pomocą metod mechaniki kwantowej stanowi jedno z najważniejszych zagadnień fizyki ciała stałego, a w szczególności fizyki metali, półprzewodników i izolatorów. Wykorzystuje się przy tym charakterystyczną budowę wewnętrzną ciał krystalicznych. W rezultacie otrzymuje się wartości dozwolonych poziomów energetycznych zgrupowanych w pasma. Mimo wyjaśnienia podstawo-

wych problemów związanych ze strukturą elektronową i mimo licznych sukcesów tej teorii, nadal istnieje wiele zagadnień nie rozwiązanych, głównie z powodu trudności rachunkowych. Ścisły opis układu cząstek stanowiących makroskopowe ciało stałe, tj. elektronów i jąder atomowych, wymagałby ułożenia i rozwiązania równania (w opisie kwantowym) lub układu równań (w opisie klasycznym) zawierającego liczbę zmiennych porównywalną z liczbą Avogadra (ok. $6\cdot 10^{23}$). Jest to niewykonalne nawet przy użyciu najpotężniejszych maszyn liczących i dlatego stosuje się uproszczenia umożliwiające otrzymanie rozwiązań przybliżonych.

Wiele ze stosowanych przybliżeń ma tę własność, że ich jakość jest tym lepsza, im więcej cząstek zawiera rozpatrywany układ fizyczny. Nie jest to cecha wyjątkowa. Na przykład rozpatrując gaz w naczyniu operujemy takimi pojęciami, jak ciśnienie, temperatura, entropia itd., które tym lepiej opisują układ, im więcej cząsteczek gazu on zawiera — gdyby w naczyniu było tylko kilka cząsteczek, posługiwanie się wymienionymi pojęciami byłoby mało sensowne, należałoby wtedy każdą cząsteczkę opisywać oddzielnie, uwzględniając jej oddziaływanie z innymi cząsteczkami i ze ściankami naczynia.

przybliżenie jednoelektronowe

Wspomnianą własność ma również stosowane najczęściej przybliżenie zwane przybliżeniem jednoelektronowym. Można je odnosić zarówno do bezpostaciowych ciał stałych, jak i do ciał krystalicznych. W przybliżeniu tym zakłada się, że ruch każdego elektronu walencyjnego można opisać niezależnie przyjmując, że porusza się on w pewnym polu potencjalnym $V(\vec{r})$ ($\vec{r}$ — wektor położenia), niezależnym od stanu rozpatrywanego elektronu. Potencjał ten jest wytwarzany przez pozostałe elektrony walencyjne oraz rdzenie atomowe, czyli jądra, wraz z elektronami niewalencyjnymi. Strukturę elektronową bezpostaciowych ciał stałych bada się od niedawna i nie wiadomo jeszcze, jakie metody postępowania okażą się najowocniejsze. W odniesieniu do ciał krystalicznych metody, oparte na przybliżeniu jednoelektronowym, są podstawowymi metodami badawczymi.

Oprócz przybliżenia jednoelektronowego do badania struktury elektronowej coraz częściej stosuje się metody teorii wielu ciał, czyli metody uwzględniające wzajemne oddziaływanie wszystkich elektronów walencyjnych. Tego rodzaju opis wielociałowy jest znacznie trudniejszy od opisu jednoelektronowego i nie stosuje się go tak powszechnie. Istnieją jednak zjawiska, których nie można wyjaśnić za pomocą opisu jednoelektronowego (np. zjawisko nadprzewodnictwa, którego istotę udało się wyjaśnić dopiero za pomocą ujęcia wielociałowego, → Nadprzewodnictwo).

Dla niektórych ciał krystalicznych przybliżoną postać potencjału $V(\vec{r})$ można wyznaczyć korzystając z mniej lub bardziej uzasadnionych założeń modelowych. Czasami jest to wystarczające, jednak dla żadnego kryształu nie znamy takiej postaci $V(\vec{r})$, która umożliwiałaby pełny opis podstawowych jego własności. Mimo tej trudności wiele zagadnień daje się rozwiązać dzięki symetrii kryształów i związanej z nią symetrii potencjału $V(\vec{r})$.

Rozmiary rzeczywistych kryształów (rzędu milimetrów) są wielokrotnie większe od stałej sieci (kilka angstremów). Z tego względu na ogół pomija się zjawiska powierzchniowe związane z istnieniem powierzchni kryształu (→ Stany powierzchniowe w ciałach stałych) i przyjmuje, że kryształ jest nieskończony.

warunki Borna–Kármána

Czasami na rozpatrywany kryształ nieskończony nakłada się dodatkowo warunki Borna–Kármána. Jest to równoważne założeniu, że cały nieskończony kryształ jest zbudowany z okresowo powtarzających się w przestrzeni makroskopowych segmentów, z których każdy jest równoważny badanemu kryształowi rzeczywistemu. Jeżeli rzeczywisty kryształ jest kostką o wymiarach $N_1a_1 \times N_2a_2 \times N_3a_3$, gdzie a_1, a_2 i a_3 są stałymi sieci, a N_1, N_2 i N_3 — bardzo dużymi liczbami naturalnymi, to warunki Borna–Kármána oznaczają po prostu, że dla dowolnej wielkości fizycznej $\mathcal{P}$ zachodzi związek:

$$\mathcal{P}(x, y, z) = \mathcal{P}(x+n_1N_1a_1, y+n_2N_2a_2, z+n_3N_3a_3),$$

gdzie n_1, n_2 i n_3 są dowolnymi liczbami całkowitymi.

potencjał krystaliczny

Przy badaniu struktury elektronowej kryształów najważniejsza okazuje się ich symetria translacyjna, tj. fakt, że kryształ ma budowę mikroskopową okresowo powtarzającą się w przestrzeni, co można wyrazić następującym związkiem spełnianym przez potencjał krystaliczny $V(\vec{r})$:

$$V(\vec{r}+n_1\vec{a}_1+n_2\vec{a}_2+n_3\vec{a}_3) = V(\vec{r}),$$

gdzie $\vec{a}_1$, $\vec{a}_2$ i $\vec{a}_3$ są wektorami charakterystycznymi dla danej sieci krystalicznej, a n_1, n_2 i n_3 — dowolnymi liczbami całkowitymi. Zauważmy, że o ile w wypadku warunków Borna–Kármána chodziło o makroskopową symetrię dowolnej wielkości fizycznej $\mathcal{P}$, o tyle w wypadku symetrii translacyjnej istotna jest mikroskopowa symetria konkretnej wielkości fizycznej — potencjału krystalicznego $V(\vec{r})$.

twierdzenie Blocha

Z symetrii translacyjnej wynika, że stacjonarne funkcje falowe elektronu w krysztale są modulowanymi falami płaskimi postaci (tzw. twierdzenie Blocha):

$$\psi_{\vec{k}}(\vec{r}) = u_{\vec{k}}(\vec{r})e^{i\vec{k}\vec{r}}, \tag{1}$$

gdzie $\vec{k}$ jest wektorem falowym, a $u_{\vec{k}}(\vec{r})$ — funkcją modulującą mającą taką samą symetrię przestrzenną jak potencjał $V(\vec{r})$. Czynnik $e^{i\vec{k}\vec{r}} = \cos\vec{k}\vec{r}+i\sin\vec{k}\vec{r}$ reprezentuje falę płaską zapisaną za pomocą liczb zespolonych (czynnik zawierający czas w zagadnieniach stacjonarnych nie występuje i został pominięty).

dozwolone wartości energii

Zgodnie z mechaniką kwantową wszystkie informacje o układzie fizycznym, w danym wypadku o elektronie w krysztale, są zawarte w równaniu Schrödingera. Przyjmując, że funkcje falowe będące rozwiązaniami równania Schrödingera mają postać (1), z równania tego uzyskuje się równanie wyznaczające postać funkcji $u_{\vec{k}}(\vec{r})$ oraz dozwolone wartości energii. Ogólnym wnioskiem z tego rodzaju rozważań jest to, że dozwolone energie E elektronów w krysztale są funkcjami wektora falowego $\vec{k}$. Wyznaczenie konkretnej zależności $E(\vec{k})$ wymagałoby rozwiązania otrzymanego równania.

W granicznym przypadku tzw. pustej sieci przyjmuje się, że potencjał krystaliczny V zależy od wektora położenia $\vec{r}$, ale bardzo słabo. Można wtedy przyjąć, że $V(\vec{r})$ jest stałe (lub równe zeru przy odpowiednim wyborze punktu zerowego dla potencjału), a mimo to ma symetrię taką jak kryształ, a nie taką, jak rzeczywiście zupełnie pusta przestrzeń. Zależność $E(\vec{k})$ ma wtedy postać:

$$E(\vec{k}) = \hbar^2k^2/2m,$$

gdzie $\hbar = h/2\pi$ (h — stała Plancka), m — masa elektronu.

Jak wiadomo z mechaniki, energia E i pęd p cząstki swobodnej związane są zależnością

$$E = p^2/2m.$$

kwazipęd elektronu

Zatem wielkość $\hbar\vec{k}$ odgrywa rolę analogiczną do pędu w mechanice. Podobieństw między wektorem $\hbar\vec{k}$ a pędem jest więcej i z tego względu $\hbar\vec{k}$ nazwano kwazipędem.

Zależność energii E od wektora falowego $\vec{k}$ traktowana jako jednoznaczna funkcja $\vec{k}$ (mogącego przyjmować dowolne wartości) na ogół nie jest funkcją ciągłą. Bliższe badania tej zależności wskazują, że gdy

$V(\vec{r}) \neq$ const, to wartość E dla pewnych wartości $\vec{k}$ może zmieniać się skokowo.

Nieciągłości funkcji $E(\vec{k})$ mogą występować tylko dla tych $\vec{k}$, które spełniają związek:

$$\vec{k}\frac{\vec{K}}{2} = \frac{K^2}{4},$$

gdzie

$$\vec{K} = l_1\vec{b}_1 + l_2\vec{b}_2 + l_3\vec{b}_3,$$

$$\vec{b}_1 = 2\pi \frac{\vec{a}_2 \times \vec{a}_3}{\vec{a}_1(\vec{a}_2 \times \vec{a}_3)},$$

$$\vec{b}_2 = 2\pi \frac{\vec{a}_3 \times \vec{a}_1}{\vec{a}_1(\vec{a}_2 \times \vec{a}_3)},$$

$$\vec{b}_3 = 2\pi \frac{\vec{a}_1 \times \vec{a}_2}{\vec{a}_1(\vec{a}_2 \times \vec{a}_3)};$$

l_1, l_2 i l_3 oznaczają liczby całkowite.

strefy Brillouina

Powierzchnie, na których mogą występować nieciągłości funkcji $E(\vec{k})$ rozbijają przestrzeń wektora $\vec{k}$ na rozłączne obszary zwane strefami Brillouina. (Przez przestrzeń wektora $\vec{k}$ rozumie się trójwymiarową przestrzeń euklidesową, w której na trzech kartezjańskich osiach współrzędnych odłożono k_x, k_y i k_z). Strefa zawierająca punkt $\vec{k} = (0, 0, 0)$ nazywa się pierwszą strefą Brillouina, strefa z nią sąsiadująca nazywa się drugą strefą Brillouina, itd. Wektory $\vec{b}_1$, $\vec{b}_2$ i $\vec{b}_3$ nazywają się wektorami podstawowymi sieci odwrotnej, zaś wektor $\vec{K}$ — wektorem sieci odwrotnej. Wszystkie strefy Brillouina mają jednakową objętość i każdą z nich można uważać za pewną komórkę elementarną sieci odwrotnej. Innymi słowy, całą sieć odwrotną można otrzymać z pierwszej strefy Brillouina przez jej równoległe przesuwanie. To samo odnosi się do drugiej, trzeciej i dalszych stref Brillouina. Strefy Brillouina o różnych numerach mają różne kształty. Spośród wszystkich stref wyróżnia się pierwsza. W odróżnieniu od innych strefa ta jest jednospójna, tzn. składa się tylko z jednej części. Każda inna strefa Brillouina składa się z kilku oddzielnych części.

warunki na wektory falowe

W przypadku kryształu nieskończonego bez nałożonych warunków Borna–Kármána wektory falowe $\vec{k}$ mogą przyjmować dowolne wartości. Innymi słowy $\vec{k}$ może się zmieniać w sposób ciągły. Nałożenie warunków Borna–Kármána, czyli uwzględnienie skończonych rozmiarów kryształu, połączone z pominięciem zjawisk powierzchniowych, jest równoważne nałożeniu na $\vec{k}$ pewnych warunków. Warunki te mogą być spełnione tylko dla niektórych wektorów $\vec{k}$. W rezultacie dla kryształu z nałożonymi warunkami Borna–Kármána wektory $\vec{k}$ mogą przyjmować tylko dyskretne (nieciągłe) wartości, np. dla kryształu skończonego mającego kształt równoległościanu o krawędziach $L_1 = N_1a_1$, $L_2 = N_2a_2$ i $L_3 = N_3a_3$ wektory $\vec{k}$ mogą przyjmować wartości dane wzorem:

$$\vec{k} = \left(\frac{2\pi}{N_1a_1}n_1, \frac{2\pi}{N_2a_2}n_2, \frac{2\pi}{N_3a_3}n_3\right), \tag{2}$$

gdzie n_1, n_2 i n_3 są dowolnymi liczbami całkowitymi. Wektorom $\vec{k}$ należącym do pierwszej strefy Brillouina odpowiadają wtedy wartości n_1, n_2 i n_3 spełniające nierówności:

$$-N_i/2 \leqslant n_i < N_i/2.$$

Liczba różnych wektorów falowych w każdej ze stref Brillouina jest jednakowa i wynosi $N_1N_2N_3$, czyli jest równa liczbie komórek elementarnych w krysztale.

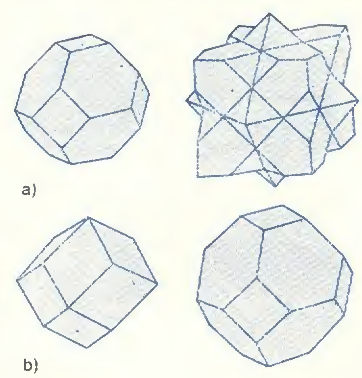
zredukowany wektor falowy

Dla każdego wektora $\vec{k}_2$ z drugiej (trzeciej, czwartej itd.) strefy Brillouina można znaleźć taki wektor $\vec{k}_1$ z pierwszej strefy Brillouina, że różnica $\vec{k}_2 - \vec{k}_1$ będzie jednym z wektorów sieci odwrotnej $\vec{K}$. Funkcje falowe $\psi_{\vec{k}}(\vec{r})$ odpowiadające wektorom $\vec{k}$ różniącym się o wektor $\vec{K}$ różnią się jedynie postacią funkcji $u_{\vec{k}}(\vec{r})$. W związku z tym stany stacjonarne elektronów w krysztale można opisywać wartością $\vec{k}$ z pierwszej strefy Brillouina (zredukowany wektor falowy) uważając $u_{\vec{k}}(\vec{r})$ za wieloznaczną funkcję $\vec{r}$, a $E(\vec{k})$ za wieloznaczną funkcję $\vec{k}$. Innymi słowy, przy ograniczeniu obszaru zmienności wektora $\vec{k}$ do pierwszej strefy Brillouina, każdej wartości $\vec{r}$ odpowiada więcej niż jedna wartość $u_{\vec{k}}(\vec{r})$, a każdej wartości $\vec{k}$ — więcej niż jedna wartość $E(\vec{k})$. Dokładniejsza analiza wskazuje, że wieloznaczną funkcję $E(\vec{k})$ określoną w całej przestrzeni wektora $\vec{k}$ można traktować jako zespół jednoznacznych funkcji ciągłych $E_n(\vec{k})$ określonych dla $\vec{k}$ z pierwszej strefy Brillouina, numerowanych parametrem n o wartościach naturalnych. Mówimy, że parametr n numeruje pasma energetyczne, czyli poszczególne gałęzie funkcji $E(\vec{k})$.

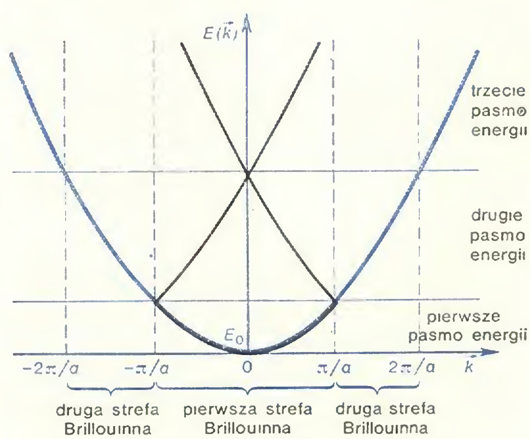
pasma energetyczne

kształt stref Brillouina

Kształt stref Brillouina zależy od symetrii translacyjnej sieci krystalicznej. Na rys. 1 pokazano kształt pierwszej i drugiej strefy Brillouina dla sieci regularnej płasko i przestrzennie centrowanej. Na rys. 2 pokazano zależność energii od wektora falowego dla pustej

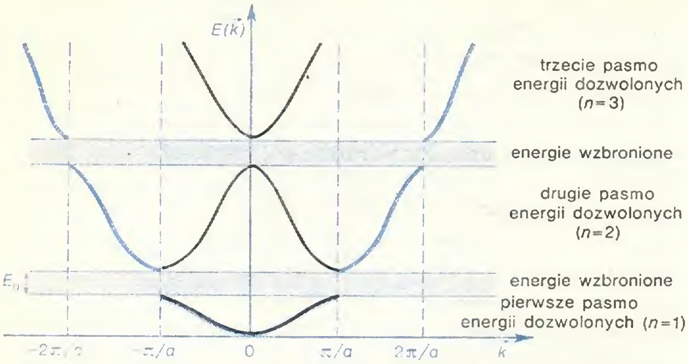


Rys. 1. Pierwsza i druga strefa Brillouina dla sieci regularnej: a) płasko centrowanej, b) przestrzennie centrowanej

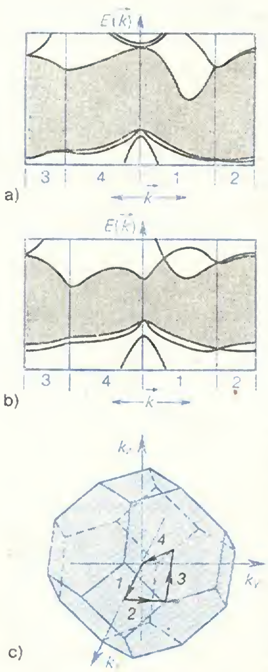


Rys. 2. Zależność $E(\vec{k})$ dla pustej sieci. Niebieska parabola odpowiada niezredukowanemu wektorowi $\vec{k}$. Po zredukowaniu wektora $\vec{k}$ do pierwszej strefy Brillouina funkcja $E(\vec{k})$ określona dla $\vec{k}$ spełniających warunek $-\pi/a \leqslant k < \pi/a$ staje się funkcją wieloznaczną (czarne linie). W przypadku pustej sieci dozwolone są wszystkie wartości $E \geqslant E_0$

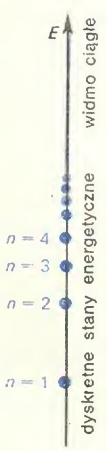
sieci, a na rys. 3 dla bardziej realistycznego, chociaż również fikcyjnego przypadku, gdy sieć nie jest pusta. Przykład rzeczywistej zależności $E_n(\vec{k})$ przedstawiono na rys. 4.



Rys. 3. Przykład zależności $E(\vec{k})$ dla $V(\vec{r}) \neq$ const. Widzimy, że na granicach stref Brillouina pojawiły się nieciągłości funkcji $E(\vec{k})$. Niebieskim kolorem zaznaczono zależność $E(\vec{k})$ dla niezredukowanego wektora falowego (wtedy E jest jednoznaczną funkcją $\vec{k}$). Po zredukowaniu $\vec{k}$ do pierwszej strefy Brillouina E staje się wieloznaczną funkcją $\vec{k}$ rozpadającą się na kilka jednoznacznych gałęzi ciągłych numerowanych parametrem n. Jak widać z rysunku, dozwolone energie elektronów mogą leżeć tylko wewnątrz pewnych przedziałów oddzielonych obszarami energii wzbronionych, tj. energii, których nie może mieć elektron w stanie stacjonarnym. Szerokość pasma energii wzbronionych między pasmem walencyjnym a pasmem przewodnictwa nazywa się przerwą energetyczną (E_g)



Rys. 4. Zależność energii E od zredukowanego wektora falowego $\vec{k}$ dla: a) krzemu i b) germanu. Ze względu na niemożliwość pokazania zależności E od $\vec{k}$ przebiegającego całą trójwymiarową strefę Brillouina pokazano jedynie zależność $E(\vec{k})$ dla $\vec{k}$ leżących na kilku liniach zaznaczonych na rys. c) przedstawiającym pierwszą strefę Brillouina dla germanu i krzemu



Rys. 5. Stany energetyczne atomu

W przytoczonym opisie kwantowym nie uwzględniono własnego momentu pędu elektronu, tj. spinu. Wprowadzenie tej wielkości podwaja liczbę pasm, lecz nie zmienia zasadniczego kształtu teorii. Nadal ważne jest stwierdzenie, że dozwolone energie elektronów mogą leżeć tylko wewnątrz pewnych obszarów oddzielonych przerwami energetycznymi i że stany elektronów wewnątrz tych pasm można „numerować" za pomocą wektora falowego $\vec{k}$.

Sytuacja w kryształach pod pewnymi względami przypomina sytuację w atomach i cząsteczkach. Na przykład w atomie (rys. 5) główna liczba kwantowa n numeruje poszczególne poziomy energetyczne, zaś stany elektronów odpowiadające danemu n różnią się wartością liczb kwantowych orbitalnej liczby kwantowej $l = 0, 1, 2, ..., n$, magnetycznej liczby kwantowej $m = 0, \pm 1, \pm 2, ..., \pm l$ i spinowej liczby kwantowej $s = \pm 1/2$. W krysztale liczba n numeruje pasma, zaś wektor $\vec{k}$ charakteryzuje stany wewnątrz pasma.

liczba stanów elektronowych w pasmie

Dla kryształu skończonego liczba różnych stanów elektronowych w pasmie, czyli liczba różnych dozwolonych wartości $\vec{k}$ w pierwszej strefie Brillouina równa jest $N_1N_2N_3$, tj. liczbie komórek elementarnych w krysztale. Z doświadczenia wiadomo, że szerokość pasm energetycznych, tj. różnica między najwyższą i najniższą energią elektronów w danym pasmie, jest rzędu 1 eV. Wynika stąd, że różnice energii między kolejnymi poziomami energetycznymi w pasmie są rzędu 10^{-23} eV (dla kryształu skończonego). Szerokość strefy Brillouina w kierunku równoległym do osi x_i wynosi $2\pi/a_i$. Odległości między kolejnymi dozwolonymi wektorami $\vec{k}$ w tym kierunku wynoszą $2\pi/N_ia_i$, czyli stanowią — z grubsza biorąc — jedną dziesięciomilionową część szerokości strefy ($1/N_i \approx 1/(\text{liczba Avogadra})^{1/3} \approx 10^{-7}$). Widać, że stany elektronowe w krysztale skończonym w przestrzeni wektora $\vec{k}$ są położone bardzo gęsto. Dzięki temu nawet w przypadku kryształu skończonego można postępować tak, jak postępuje się w przypadku kryształu nieskończonego bez nałożonych warunków Borna-Kármána, tzn. można przyjąć, że wewnątrz pasma każda energia jest dozwolona, a wewnątrz pierwszej strefy Brillouina dozwolone są wszystkie wartości $\vec{k}$. Innymi słowy, można przyjąć, że dopuszczalne wartości E i $\vec{k}$ zmieniają się w sposób ciągły, a nie dyskretny.

rola warunków Borna-Kármána

Warto w tym miejscu wyjaśnić rolę warunków Borna-Kármána. Fizyczne znaczenie tych warunków zostało już omówione poprzednio. Natomiast z matematycznego punktu widzenia nałożenie warunków Borna-Kármána prowadzi do tego, że zbiory dopuszczalnych wartości E i $\vec{k}$ stają się zbiorami dyskretnymi. Bez tych warunków zarówno E jak i $\vec{k}$ zmieniałyby się w sposób ciągły. Ponieważ nawet w przypadku dyskretnym w praktyce stosuje się opis ciągły ze względu na fakt, że sąsiednie wartości dyskretne są bardzo bliskie siebie, mogłoby się wydawać, że nakładanie warunków Borna-Kármána niczego nie daje. Tak jednak nie jest. Często bowiem zdarza się, że ważna jest liczba stanów w pasmie. Wiedząc np. jaka jest liczba elektronów walencyjnych przypadających na jedną komórkę elementarną, znając liczbę komórek elementarnych w krysztale i wiedząc, ile stanów „mieści się" w poszczególnych pasmach, bez trudu można obliczyć (uwzględniając zakaz Pauliego), ile najniższych pasm jest zapełnionych przez elektrony. Obliczenia takie przy ciągłej zmienności E i $\vec{k}$ są znacznie mniej przejrzyste.

zależność $\varrho_{\vec{k}}(\vec{r})$

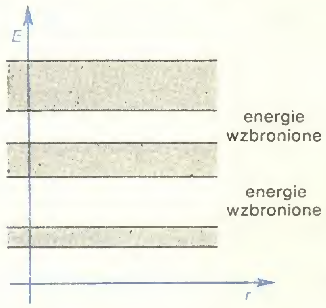
Gęstość prawdopodobieństwa $\varrho_{\vec{k}}(\vec{r})$ znalezienia elektronu w punkcie danym wektorem $\vec{r}$ jest równa kwadratowi wartości bezwzględnej funkcji $\psi_{\vec{k}}(\vec{r})$ lub funkcji $u_{\vec{k}}(\vec{r})$ (→ Chemia kwantowa):

$$\varrho_{\vec{k}}(\vec{r}) = |\psi_{\vec{k}}(\vec{r})|^2 = |u_{\vec{k}}(\vec{r})|^2.$$

Dla elektronów walencyjnych funkcja $\varrho_{\vec{k}}(\vec{r})$ zazwyczaj ma wartości istotnie różne od zera w całej komórce elementarnej, a więc i w całym krysztale (funkcja $\varrho_{\vec{k}}(\vec{r})$ ma taką samą symetrię translacyjną jak potencjał krystaliczny $V(\vec{r})$). Elektrony walencyjne nie są więc zlokalizowane w określonych punktach przestrzeni, np. przy atomach, lecz są rozmyte po całym krysztale. Sytuację taką mamy w większości substancji o praktycznym znaczeniu, chociaż nie musi to zawsze być prawdą. Wydaje się, że substancją, dla której funkcja $\varrho_{\vec{k}}(\vec{r})$ ma wartość istotnie różną od zera tylko w części komórki elementarnej jest bor.

Dopuszczalne wartości energii całkowitej elektronów $E_n(\vec{k})$ są „numerowane" parametrem dyskretnym n, wskazującym pasmo, oraz parametrem ciągłym lub „prawie ciągłym" $\vec{k}$, nie zależą natomiast od położenia elektronów w krysztale opisywanego wektorem $\vec{r}$. Gdybyśmy więc chcieli zrobić wykres dopuszczalnych energii E od $\vec{r}$, to dla każdej dozwolonej wartości E otrzymalibyśmy poziomą kreskę. Kreski odpowiadające wszystkim wartościom E, jakie mogą przyjmować całkowite energie elektronów, tworzą pasma pokazane na rys. 6. Rysunki tego rodzaju są uboższe niż wy-

zależność $E(\vec{r})$



Rys. 6. Pasma energetyczne w ciele stałym

kresy przedstawiające E w zależności od n i $\vec{k}$. Uwidocznione są na nich jedynie zakresy energii przyjmowane przez elektrony w poszczególnych pasmach i nic poza tym. W szczególności nie należy sądzić, że jeśli w pasmie mamy tylko jeden elektron, to może on „przemieszczać" się w nim swobodnie, bo „ma dużo miejsca". Argumentacja taka, często spotykana przy omawianiu urządzeń półprzewodnikowych takich, jak tranzystory, diody itp. (zwłaszcza w opracowaniach o charakterze popularnym), jest bardzo sugestywna, ale nieprawdziwa. Dla stanów stacjonarnych gęstość prawdopodobieństwa znalezienia elektronu w punkcie $\vec{r}$, czyli wielkość $\varrho_{\vec{k}}(\vec{r})$ nie zależy od czasu, a więc o żadnym ruchu w zwykłym znaczeniu nie może być mowy. W wykorzystywanych w praktyce materiałach często zdarza się, że istotnie w pasmie pustym elektron walencyjny może się przemieszczać stosunkowo swobodnie. Uzasadnienie tego jest jednak dość złożone i odwoływanie się tylko do intuicyjnych wyobrażeń na ogół zawodzi. Konieczny jest tu ścisły opis kwantowy i traktowanie elektronu jako paczki falowej, czyli pewnej kombinacji stanów stacjonarnych. Okazuje się, że o możliwości poruszania się elektronu w pasmie decyduje nie tylko liczba elektronów w pasmie lecz i inne wielkości, np. stopień nakładania się funkcji falowych o różnych wartościach wektora falowego $\vec{k}$ itp. Mogą zdarzać się przypadki, gdy mimo, że pasmo jest słabo obsadzone, swoboda ruchu elektronów jest mała. Trudno jest ustalić tu bezwzględnie obowiązującą regułę. Z grubsza biorąc można jednak powiedzieć, że im funkcje $\varrho_{\vec{k}}(\vec{r})$ są bardziej rozmyte, tym swoboda ruchu elektronów jest większa (stany zdelokalizowane). Odwrotnie — funkcjom wyraźnie zlokalizowanym w jakimś obszarze na ogół odpowieda mała swoboda ruchu elektronów (stany zlokalizowane).

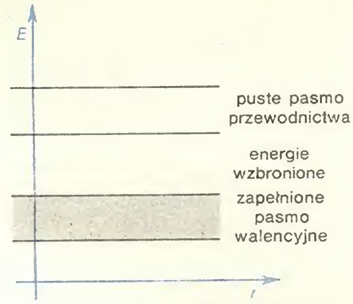
interpretacja wykresów energii

W temperaturze 0 K elektrony zajmują najniższe stany energetyczne. Liczba stanów w pasmie jest równa liczbie komórek elementarnych w krysztale. Tego samego rzędu jest też liczba elektronów walencyjnych kryształu. Najwyższe pasmo zawierające elektrony nazywamy pasmem walencyjnym, zaś pasmo leżące nad nim — pasmem przewodnictwa lub pasmem przewodzenia. W zależności od sposobu obsadzenia i wzajemnego położenia pasm oraz od swobody ruchu elektronów kryształy dzieli się na izolatory, półprzewodniki i metale.

pasmo walencyjne i pasmo przewodnictwa

W izolatorach i czystych półprzewodnikach w temperaturze 0 K niższe pasmo (walencyjne) obsadzone całkowicie elektronami jest oddzielone od zupełnie pustego pasma wyższego (przewodnictwa) przerwą energetyczną (rys. 7). W pasmie pustym nie ma co się

izolatory



Rys. 7. Struktura pasmowa izolatorów i półprzewodników samoistnych

poruszać, zaś w pasmie całkowicie obsadzonym elektrony zajmują wszystkie możliwe stany (o wszystkich wartościach $\vec{k}$) i też nie mogą się poruszać (zmieniać stanu). Tak więc w temperaturze 0 K ani izolatory ani półprzewodniki nie mogą przewodzić prądu elektrycznego. Między izolatorami i czystymi półprzewodnikami, czyli półprzewodnikami samoistnymi, nie ma różnicy jakościowej. Jest jedynie różnica ilościowa. Gdy przerwa energetyczna jest dostatecznie mała, to w zwykłych temperaturach wskutek wzbudzeń termicznych część elektronów z pasma walencyjnego przechodzi do pasma przewodnictwa.

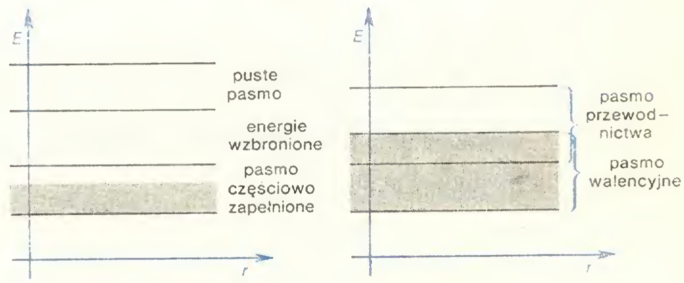
półprzewodniki samoistne

Jeżeli stany są zdelokalizowane (ruch elektronów jest wtedy w dużym stopniu swobodny), to przez przyłożenie zewnętrznego pola elektrycznego łatwo wywołać przepływ prądu i mamy półprzewodnik.

W izolatorach przerwa energetyczna jest na tyle duża, że termiczne przejścia elektronów z pasma walencyjnego do pasma przewodnictwa są prawie niemożliwe, bądź stany elektronowe są silnie zlokalizowane. W pierwszym wypadku brak jest nośników ładunku, natomiast w drugim ruch nośników jest bardzo utrudniony (przyjęło się dość arbitralnie uważać za izolatory materiały o przerwie energetycznej większej od 2,5–3 eV).

W metalach najwyższe pasmo zawierające niezlokalizowane eletrony jest zapełnione częściowo lub też zachodzi nakładanie pasma walencyjnego na puste pasmo przewodnictwa. W obu wypadkach możliwy jest ruch elektronów i przepływ prądu (rys. 8).

metale

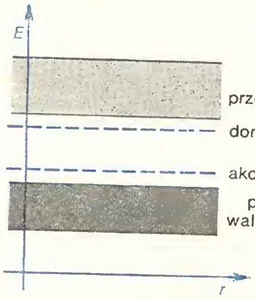


Rys. 8. Struktura pasmowa metali

Omówiona wyżej struktura elektronowa odnosi się do idealnych kryształów nie zawierających ani domieszek, ani defektów budowy sieci krystalicznej. W rzeczywistości sytuacja taka nigdy się nie zdarza. Idealny kryształ musiałby mieć entropię równą zeru, a to dla żadnej temperatury różnej od 0 K nie jest możliwe. Uwzględnienie domieszek i defektów w krysztale prowadzi do wniosku, że w strukturze elektronowej kryształu powinny pojawić się dodatkowe poziomy energetyczne. Szczególnie ważne okazują się dwa rodzaje poziomów domieszkowych leżące wewnątrz przerwy energetycznej: poziomy donorowe w pobliżu dna pasma przewodnictwa i poziomy akceptorowe w pobliżu wierzchołka pasma walencyjnego.

wpływ domieszek i defektów

Dla poziomów odpowiadających domieszkom lub defektom budowy gęstość prawdopodobieństwa znalezienia elektronu $\varrho(\vec{r})$ jest istotnie różna od zera tylko w pewnym obszarze wokół defektu lub domieszki. Tego rodzaju stany zlokalizowane zaznaczamy zwykle za pomocą linii przerywanej (rys. 9). Przy dużej koncentracji domieszek lub defektów obszary odpowiadające różnym od zera wartościom różnych funkcji



Rys. 9. Poziomy donorowe i akceptorowe

pasma domieszkowe

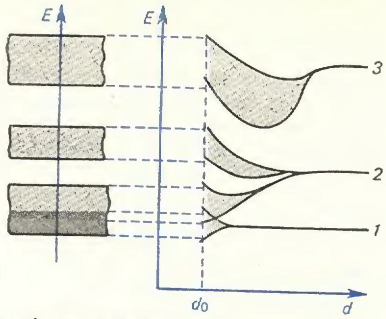
$\varrho(\vec{r})$ zaczynają się nakładać. Stany domieszkowe przestają wtedy być stanami zlokalizowanymi i powstaje wąskie pasmo domieszkowe.

Zazwyczaj poziomy donorowe lub akceptorowe odpowiadają domieszkom, mogą one jednak powstawać również wskutek występowania defektów budowy sieci krystalicznej, np. dyslokacji.

Poziomy donorowe i akceptorowe z reguły bardzo silnie wpływają na własności elektryczne kryształów. Przez odpowiednie domieszkowanie niektórym izolatorom można nadać własności półprzewodzące.

Powstawanie pasm energetycznych można wyjaśnić w inny sposób. Sposób ten jest bardziej poglądowy, nie stanowi jednak wygodnej podstawy do wykonywania niezbędnych obliczeń ilościowych. W sieci utworzonej z N komórek elemetarnych znajduje się N jednakowych, jednakowo rozmieszczonych atomów. Gdyby poszczególne atomy nie oddziaływały ze sobą, to każdy prosty poziom elektronowy atomu byłby N-krotnie zwyrodniały. W rzeczywistości atomy oddziałują ze sobą — w przeciwnym wypadku nie utworzyłyby kryształu. Oddziaływanie to znosi zwyrodnienie: każdy N-krotnie zwyrodniały poziom rozszczepia się na N pojedynczych poziomów leżących bardzo blisko siebie i tworzących prawie ciągłe pasmo — rys. 10. Gdy atomy są od siebie bardzo oddalone (duża odległość międzyatomowa d), rozszczepienie poziomów jest bardzo małe, a funkcje falowe są silnie zlokalizowane. W miarę zmniejszania d następują

rozszczepienie poziomów



Rys. 10. Powstawanie pasm energetycznych z atomowych poziomów energetycznych; d stała sieci, d_0 wartość stałej sieci w kryształach rzeczywistych

coraz większe rozszczepienia poziomów, a jednocześnie funkcje falowe elektronów stają się coraz bardziej zdelokalizowane.

parametry struktury pasmowej

Parametry struktury pasmowej, czyli wielkości pozwalające określić położenie i kształt funkcji $E_n(\vec{k})$, można w zasadzie obliczyć znając potencjał $V(\vec{r})$. Jest to jednak skomplikowane i częściowo udało się tylko w kilku przypadkach. Zazwyczaj strukturę pasmową konkretnych ciał krystalicznych wyznacza się metodami półempirycznymi: na podstawie rozważań teoretycznych definiuje się pewne parametry, a konkretne wartości tych parametrów przyjmuje się tak, aby przewidywane wartości mierzalne były zgodne z doświadczeniem dla możliwie szerokiej klasy zjawisk. Szczególnie cenne są tu dane z doświadczeń optycznych. Badanie krawędzi absorpcji oraz natężeń przejść optycznych między rozmaitymi stanami pozwala wyznaczyć wielkości przerw energetycznych oraz kształt pasm w pobliżu ekstremów funkcji $E_n(\vec{k})$.

A.I. Anselm *Wstep do teorii półprzewodników*, Warszawa 1967; F.J. Blatt *Fizyka zjawisk elektronowych w metalach i półprzewodnikach*, Warszawa 1973, C. Kittel *Wstęp do fizyki ciała stałego*, Warszawa 1970; Ch.A. Wert, R.M. Thomson *Fizyka ciała stałego*, Warszawa 1974.

Dynamika elektronu w ciałach stałych (kryształach)

Waldemar Gorzkowski

Ruch elektronu w przestrzennie okresowym polu krystalicznym, związanym z okresową budową kryształu, podlega nieco innym prawom niż ruch elektronu swobodnego. Prawa dynamiki elektronu w krysztale zazwyczaj staramy się formułować w sposób nawiązujący do praw mechaniki klasycznej, używając języka podobnego do języka zwykłej mechaniki. Prawa te wyprowadza się z praw mechaniki kwantowej. Są one słuszne z pewnymi ograniczeniami i dotyczą zwykle wielkości uśrednionych po obszarze komórki elementarnej, np. mówiąc o prędkości elektronu mamy na myśli średnią prędkość elektronu w komórce elementarnej. Prędkość tę wyraża wzór:

prędkość elektronu

$$\vec{v} = \frac{1}{\hbar}\operatorname{grad}_{\vec{k}} E_n(\vec{k}) = \frac{1}{\hbar}\left(\frac{\partial}{\partial k_x}E_n(\vec{k}), \frac{\partial}{\partial k_y}E_n(\vec{k}), \frac{\partial}{\partial k_z}E_n(\vec{k})\right). \quad (1)$$

Symbole $\frac{\partial}{\partial k_x}$, $\frac{\partial}{\partial k_y}$, $\frac{\partial}{\partial k_z}$ oznaczają pochodne funkcji, która po nich następuje, po odpowiednich zmiennych, przy założeniu, że pozostałe zmienne mają ustalone wartości (pochodne cząstkowe); $\hbar = h/2\pi$ (h — stała Plancka); $E_n(\vec{k})$ oznacza energię elektronu z n-tego pasma (w stanie stacjonarnym) w punkcie $\vec{k}$ strefy Brillouina (→ Struktura elektronowa ciał stałych), gdy na kryształ nie oddziałują pola zewnętrzne. Określona wyżej średnia prędkość elektronu jest prędkością grupową fal Blocha w n-tym pasmie.

Gdy sieć jest pusta zależność $E(\vec{k})$ opisuje wzór:

$$E = \hbar^2k^2/2m.$$

Prędkość $\vec{v}$ jest wtedy dana wzorem:

$$\vec{v} = \frac{1}{m}\hbar\vec{k}.$$

W rzeczywistych kryształach zależność $E(\vec{k})$ jest bardziej złożona i wyrażenia na $\vec{v}$ się komplikują.

W odniesieniu do wielkości średnich podstawowe prawa dynamiki elektronów opisujące zmiany kwazipędu $\hbar\vec{k}$, energii E i prędkości $\vec{v}$ elektronu pod wpływem makroskopowego pola elektrycznego $\vec{\mathcal{E}}$ i magnetycznego $\vec{H}$ można sformułować w sposób przypominający klasyczne prawa ruchu:

$$\underbrace{\frac{d}{dt}(\hbar\vec{k})}_{\substack{\text{pochodna}\\ \text{kwazipędu}}} = \underbrace{-e\vec{\mathcal{E}}}_{\substack{\text{siła pocho-}\\ \text{dząca od pola}\\ \text{elektrycznego}}} \underbrace{-\frac{e}{c}(\vec{v}\times\vec{H})}_{\substack{\text{siła pochodząca}\\ \text{od pola magne-}\\ \text{tycznego}}}, \tag{2}$$

$$\underbrace{\frac{d}{dt}E}_{\substack{\text{pochodna}\\ \text{energii}}} = \underbrace{-e\vec{\mathcal{E}}\,\vec{v}-\frac{e}{c}(\vec{v}\times\vec{H})\vec{v}}_{\text{siła razy prędkość}}, \tag{3}$$

$$\underbrace{\frac{d}{dt}\vec{v}}_{\substack{\text{przyśpie-}\\ \text{szenie}}} = \underbrace{-\frac{e}{m^*}\vec{\mathcal{E}}-\frac{e}{cm^*}(\vec{v}\times\vec{H})}_{\text{stosunek siły do masy}}; \tag{4}$$

e oznacza tu bezwzględną wartość ładunku elektronu, a m^* nosi nazwę krzywiznowej masy efektywnej. Podane wyżej równania odnoszą się do przypadku, gdy energia E zależy tylko od wartości wektora $\vec{k}$, a nie zależy od jego kierunku. Pasma takie nazywamy pasmami sferycznymi. Krzywiznowa masa efektywna, a ściślej biorąc jej odwrotność, dana jest wzorem:

masa efektywna

pasma sferyczne

$$\frac{1}{m^*}=\frac{1}{\hbar^2}\frac{\partial^2E(\vec{k})}{\partial k^2}. \tag{5}$$

Analizując wzory (2)–(5) należy zwrócić uwagę na następujące sprawy: 1) W równaniach mamy masę efektywną a nie masę spoczynkową elektronu m_0. Jest to bardzo istotna różnica, gdyż wielkości te mogą się od siebie różnić nawet o 2–3 rzędy. Dla wielu metali masa efektywna m^* jest prawie równa masie elektronu swobodnego m_0. Natomiast dla innych substancji na ogół zdarza się to dość rzadko. (Wartość masy efektywnej m^* można uzyskać z częstości cyklotronowej badając ruch elektronu w zewnętrznym polu magnetycznym; → Metale). 2) Jeżeli funkcja $E_n(\vec{k})$ jest liniową funkcją k^2, to masa efektywna m^* jest wielkością stałą, niezależną od $\vec{k}$ (a więc i od energii E). Pasma o takiej zależności $E(\vec{k})$ nazywają się pasmami parabolicznymi. W innych wypadkach masa efektywna zależy od energii elektronu.

pasma paraboliczne

Gdy pasma są niesferyczne, tj. gdy E zależy nie tylko od wartości wektora $\vec{k}$, lecz i od jego kierunku, układ równań (2)–(4) należy nieco zmodyfikować. Pierwsze dwa pozostają nadal ważne. Natomiast równanie (4) przyjmuje postać:

$$\frac{dv_j}{dt}=-e\sum_i\left(\frac{1}{m^*}\right)_{ji}\left[\mathcal{E}_i+\frac{1}{c}(\vec{v}\times\vec{H})_i\right];$$

$$i,j=x,y,z. \tag{4'}$$

Wielkość $(1/m^*)_{ji}$ nazywa się tensorem odwrotności masy efektywnej. Wielkość ta jest określona wzorem:

$$\left(\frac{1}{m^*}\right)_{ji}=\frac{1}{\hbar^2}\frac{\partial^2E(\vec{k})}{\partial k_i\partial k_j};\ i,j=x,y,z. \tag{5'}$$

Dyskusja wzoru (5′) w ogólnej postaci nie jest wygodna. Zazwyczaj dobierając odpowiedni układ współrzędnych tensor odwrotności masy efektywnej sprowadza się na podstawie pewnych reguł przekształcania tensorów do szczególnie prostej, tzw. diagonalnej postaci, w której wielkości $(1/m^*)_{ji}$ dla $i \neq j$ są równe zeru; różne od zera są wtedy tylko trzy wielkości: $(1/m^*)_{11}$, $(1/m^*)_{22}$, $(1/m^*)_{33}$.

anizotropowość masy efektywnej

We wspomnianym wyżej układzie współrzędnych, zwanym układem osi głównych tensora odwrotności masy efektywnej, ruch elektronu w każdym z trzech kierunków równoległych do osi współrzędnych odbywa się niezależnie, chociaż dla każdego z tych kierunków elektron ma na ogół inną masę efektywną. W praktyce często zdarza się, że dwie spośród trzech mas efektywnych są równe. Na ich oznaczenie stosuje się symbol $m_\perp$. Trzecią masę oznacza się wtedy symbolem $m_\parallel$. Anizotropowy charakter masy efektywnej jest przyczyną występowania rozmaitego rodzaju efektów, których nie obserwuje się w wypadku elektronów swobodnych. Na przykład przyłożenie pola elektrycznego do kryształu o pasmie niesferycznym może powodować przepływ elektronów w kierunku skośnym do kierunku pola (pojawia się wtedy napięcie poprzeczne w krysztale).

Osobliwością masy efektywnej jest to, że może ona być ujemna. Ujemnej masie efektywnej odpowiada, mówiąc językiem mechaniki klasycznej, przyspieszenie w kierunku przeciwnym do kierunku działania siły. Z różnych względów wygodniej jest wówczas mówić o dziurach zamiast o elektronach, czyli o cząstkach mających dodatni ładunek i dodatnią masę efektywną. Wielkości te co do wartości bezwzględnej są równe odpowiednim wartościom dla elektronu.

ujemna masa efektywna

Zdarza się, że spośród trzech mas efektywnych charakteryzujących pasmo niesferyczne np. tylko jedna jest dodatnia a dwie ujemne. Wtedy tylko w jednym kierunku elektron zachowuje się jak cząstka o dodatniej masie.

Jak widać z dotychczasowych rozważań, przyspieszenie elektronu określone jako dv/dt nie musi być równoległe do siły zewnętrznej. Nieizotropowość masy efektywnej i nierównoległość przyspieszenia do siły zewnętrznej biorą się stąd, że bariery potencjału, na które napotyka elektron w swym ruchu, w różnych kierunkach krystalograficznych są różne. Warto przy tym zwrócić uwagę, że w równaniach (2)–(4) nie występują w postaci jawnej siły wewnętrzne związane z istnieniem pola krystalicznego. Siły te jednak zostały w tych równaniach uwzględnione w postaci niejawnej właśnie w masie efektywnej. Gdyby pole krystaliczne nie zostało uwzględnione, to w równaniach występowałaby masa elektronu swobodnego m_0 a nie m^*.

Zależność $E_n(\vec{k})$ w rzeczywistych kryształach może być bardzo zawiła. Zwykle najbardziej interesują nas jedynie ekstrema funkcji $E_n(\vec{k})$, a konkretnie maksima pasma walencyjnego i minima pasma przewodnictwa, bowiem obsadzenie pasm wokół tych punktów jest szczególnie podatne na zmiany temperatury i innych czynników. Ekstrema $E_n(\vec{k})$ nie muszą występować w środku strefy. W germanie i krzemie maksimum energii pasma walencyjnego odpowiada $\vec{k}=0$ (środek strefy), natomiast minima pasma przewodnictwa w krzemie znajdują się wewnątrz, w germanie — na brzegu strefy Brillouina.

ekstrema pasm

W minimach wartość drugiej pochodnej jest dodatnia, a w maksimach — ujemna. W odniesieniu do funkcji $E_n(\vec{k})$ oznacza to, że minimom pasma przewodnictwa odpowiada dodatnia masa efektywna elektronów. Podobnie maksimom pasma walencyjnego odpowiada ujemna masa efektywna elektronów. W częściowo obsadzonych (np. wskutek wzbudzeń termicznych) pasmach przewodnictwa elektrony zajmują przede wszystkim stany najniższe, tj. w pobliżu minimum (ewentualną nadwyżkę energii elektrony tracą wskutek różnego rodzaju efektów rozproszeniowych na domieszkach, na drganiach sieci itp.). W związku z tym elektrony w pasmie przewodnictwa mają zwykle dodatnią masę efektywną. Inaczej sprawa przedstawia się z pasmem walencyj-

nym. Pasmo to jest prawie całkowicie wypełnione. Niewielka część elektronów wskutek wzbudzeń termicznych opuściła to pasmo przechodząc do pasma przewodnictwa i w pobliżu wierzchołka pasma mamy stany nieobsadzone. W obszarach strefy Brillouina, w których stany elektronowe są obsadzone, zmiany stanu elektronów są utrudnione ze względu na zasadę Pauliego. Stan mogą zmieniać na ogół tylko te elektrony, które znajdują się w pobliżu stanów nieobsadzonych, a więc te, które znajdują się w pobliżu wierzchołka pasma. W tym wypadku masa efektywna elektronów jest ujemna. W zewnętrznych polach elektrycznych i magnetycznych nieobsadzone stany elektronowe zachowują się jak pewne cząstki o dodatnim ładunku elektrycznym. Cząstki te przyjęto nazywać dziurami. Dzięki wprowadzeniu pojęcia dziury, zamiast mówić o pasmie prawie całkowicie zapełnionym przez elektrony, wygodnie jest mówić o pasmie prawie pustym, zawierającym niewielką liczbę dziur. Bliższa analiza wskazuje, że dziurze po elektronie o ujemnej masie efektywnej należy przypisać dodatnią masę efektywną i odwrotnie. Należy jednak podkreślić, że pole elektryczne bądź magnetyczne nie oddziałuje na puste miejsca po elektronach, ale na same elektrony. Wprawdzie przy posługiwaniu się opisem za pomocą dziur zapominamy o elektronach zapełniających pasmo, jednakże należy sobie zdawać sprawę, że własności tych elektronów określają własności dziur, ich ładunek, masę efektywną itd. Często spotykany pogląd, że dziura to miejsce po zwykłym elektronie, jest niezupełnie poprawny. Istotnie dziura o dodatniej masie efektywnej jest miejscem po elektronie, ale po elektronie o ujemnej masie efektywnej.

W pasmach sferycznych, dla $\vec{k}$ bliskich wartości ekstremalnej, zależność energii od $\vec{k}$ jest paraboliczna, ale przy $\vec{k}$ bardziej odległych — pasmo przestaje być paraboliczne. Nieparaboliczność pasm zaobserwowano np. w InSb, InAs, HgTe itp. W wielu interesujących substancjach zależność E od $\vec{k}$ dla pasma przewodnictwa daje się opisać wzorem (przykład pasma nieparabolicznego):

$$E\left(1+\frac{E}{E_g}\right) = \frac{\hbar^2 k^2}{2m_c^*}; \qquad (6)$$

E_g oznacza tu szerokość przerwy energetycznej, a m_c^* — masę efektywną na dnie pasma przewodnictwa.

Warto tu zwrócić uwagę, że zależność (6) jest podobna do zależności między pędem i energią elektronów swobodnych w mechanice relatywistycznej. Pełne podobieństwo odpowiednich wzorów otrzymuje się przez następujące przyporządkowanie:

$$m_c^* \to m_0,\ E_g \to 2m_0c^2$$

(c — prędkość światła). Analogia ta, mimo że jest czysto formalna i nie niesie głębszych treści, często może ułatwić zrozumienie zjawisk dotyczących wspomnianych substancji osobom, które są obyte ze szczególną teorią względności.

pędowa masa efektywna

Przy badaniu dynamiki elektronów w pasmach sferycznych oprócz krzywiznowej masy efektywnej m^* często wprowadza się pędową masę efektywną m^{**} określoną wzorem:

$$\underbrace{\hbar k}_{\text{kwazipęd}} = \underbrace{m^{**}v}_{\text{analog pędu}}.$$

Masa pędowa m^{**} nie musi być równa i często nie jest równa masie krzywiznowej m^*, ponieważ związek $\hbar k = m^*v$ na ogół nie zachodzi. Masę m^{**} można wyrazić następująco:

$$\frac{1}{m^{**}} = \frac{1}{\hbar^2}\frac{1}{k}\frac{dE_n\vec{k}}{dk}.$$

Dla pasm parabolicznych $m^* = m^{**}$. Dla pasm opisanych zależnością (6) zachodzi związek:

$$m^{**} = m_c^*\left(1+\frac{2E}{E_g}\right).$$

Masa pędowa, podobnie jak masa krzywiznowa, może zależeć od energii elektronu.

Masa m^{**} jest wielkością, która odgrywa rolę we wszystkich zjawiskach transportu nośników oraz w zjawiskach optycznych. Oprócz zasygnalizowanych tu kilku rodzajów zależności $E_n(\vec{k})$ istnieje wiele innych ciekawych typów pasm, jednak nie wszystkie z nich są dostatecznie zbadane i nie wszystkie mają praktyczne znaczenie.

Bibliografia → Struktura elektronowa ciał stałych.

Dynamika sieci krystalicznej

Wacław Nazarewicz

periodyczna struktura przestrzenna

Podstawową własnością odróżniającą stan krystaliczny od innych stanów (gazowego, ciekłego, amorficznego) jest periodyczna struktura przestrzenna. Przy opisie tej struktury korzystamy z pojęcia sieci krystalicznej. Rozróżniamy czternaście typów przestrzennych sieci krystalicznych, zwanych sieciami Bravais'go. Węzły sieci reprezentują położenie identycznych atomów lub grupy atomów (jonów). Możliwość przyporządkowania węzłom sieci różnych grup atomów prowadzi do dużej liczby odrębnych struktur krystalicznych (→ Budowa kryształów).

Dzięki własności periodyczności przestrzennej (symetrii translacyjnej) można w krysztale rozpatrywać tylko pewien ograniczony obszar, zwany komórką elementarną. Przez odpowiednie powtarzanie komórki elementarnej w trzech kierunkach otrzymuje się cały kryształ, który stanowi zbiór olbrzymiej liczby takich komórek. Określenie kształtu komórki elementarnej oraz rodzaju i położenia wchodzących w jej skład atomów daje pełny opis własności strukturalnych kryształu. Dane te można uzyskać metodą dyfrakcji promieniowania rentgenowskiego.

Rodzaj struktury krystalicznej, jaki ma dana substancja, zależy od natury oddziaływań między atomami. Ze względu na te oddziaływania wyodrębniamy cztery zasadnicze grupy kryształów: kryształy kowalencyjne, jonowe, metaliczne i molekularne (tutaj nie rozpatrywane). Wielkość oddziaływań charakteryzuje energia wiązania, która przybiera wartości w przedziale od dziesiątych części eV do kilku eV na atom.

oddziaływania między atomami

Oddziaływanie między atomami w krysztale można opisać używając funkcji energii potencjalnej. Funkcja ta osiąga minimum, gdy atomy znajdują się w ściśle określonych położeniach względem siebie. Położenia te są położeniami równowagi.

Przy opisie niektórych własności kryształów można z powodzeniem przyjąć model statyczny, który zakłada, że atomy są nieruchome. Założenie to jest jednak idealizacją. W rzeczywistości atomy zawsze wykonują małe oscylacje wokół swoich położeń równowagi, których amplituda zależy od temperatury. Oscylacje te nazywamy drganiami sieci krystalicznej.

drgania sieci krystalicznej

Drgania sieci krystalicznej odgrywają ważną rolę w wielu procesach zachodzących w ciałach stałych. Uwzględnienie drgań sieci jest niezbędne przy rozpatrywaniu własności cieplnych kryształów, jak ciepło właściwe, przewodnictwo i rozszerzalność cieplna.

Spośród wielu innych zjawisk, które zależą od drgań sieci, wymienimy tylko rozpraszanie elektronów w metalach i półprzewodnikach oraz absorpcję promieniowania podczerwonego w dielektrykach i półprzewodnikach.

Podstawowe przybliżenia

Kryształ można rozpatrywać jako układ złożony z dwóch rodzajów cząstek: jąder atomowych i elektronów. Między tymi cząstkami zachodzą oddziaływania elektrostatyczne typu kulombowskiego. Według mechaniki kwantowej stan kryształu opisuje funkcja falowa, która jest rozwiązaniem równania Schrödingera dla układu omawianych cząstek. Funkcja ta zależy zarówno od współrzędnych jąder atomowych, jak i współrzędnych elektronów (→ Struktura elektronowa ciał stałych).

przybliżenie adiabatyczne

U podstaw teorii ciała stałego, a tym samym i teorii dynamiki sieci krystalicznej, leży przybliżenie adiabatyczne, które pozwala rozpatrywać ruch jąder atomowych i ruch elektronów osobno. Inaczej mówiąc, stosując przybliżenie adiabatyczne, można opisać stan kryształu za pomocą iloczynu dwóch funkcji falowych, z których jedna zależy tylko od współrzędnych jąder atomowych, druga — od współrzędnych elektronów.

Przybliżenie adiabatyczne stanie się zrozumiałe, gdy zwrócimy uwagę na znaczną różnicę mas jąder i elektronów, a co za tym idzie — również różnicę prędkości ruchu tych cząstek. Jądra atomowe są znacznie cięższe od elektronów i w związku z tym poruszają się znacznie wolniej. Przy zmianie położenia jąder atomowych w krysztale prawie natychmiast ustala się rozkład przestrzenny elektronów, odpowiadający nowym położeniom jąder. Z jednej strony można więc rozpatrywać ruch elektronów w polu potencjalnym nieruchomych jąder (w ten sposób postępuje się m.in. przy badaniu struktury pasmowej kryształów), z drugiej zaś strony przy rozpatrywaniu ruchu jąder atomowych ważne są nie chwilowe położenia elektronów, ale pole powstające na skutek ich średniego przestrzennego rozkładu. Pole to jest w pełni określone przez współrzędne jąder atomowych.

W teorii kwantowej drgań sieci krystalicznej przybliżenie adiabatyczne pozwala rozpatrywać równanie Schrödingera zależne tylko od współrzędnych jądrowych. Przy badaniu drgań sieci w ramach mechaniki klasycznej przybliżenie adiabatyczne umożliwia przedstawienie oddziaływań międzyatomowych przy użyciu funkcji energii potencjalnej $U(\vec{R})$, która zależy tylko od współrzędnych jąder atomowych.

Pomimo wyeliminowania współrzędnych elektronowych otrzymanie funkcji energii potencjalnej $U(\vec{R})$ w konkretnej postaci nie jest rzeczą łatwą. Zakładamy jednak, że funkcję tę można rozwinąć w szereg potęgowy względem odchyleń atomów od położeń równowagi $\vec{u} = \vec{R} - \vec{R}_0$ (gdzie $\vec{R}_0$ oznacza położenie równowagi):

$$U = U_0 + Au + Bu^2 + Cu^3 + \dots,$$

czyli

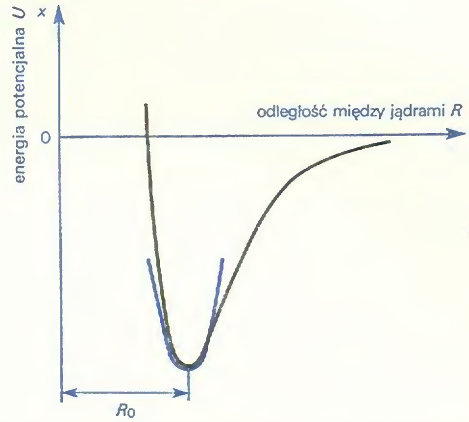
$$U = U_0 + U_1 + U_2 + U_3 + \dots \quad (1)$$

Współczynniki A, B, C, ... zawierają odpowiednio pierwszą, drugą itd. pochodną energii potencjalnej obliczoną w położeniu równowagi.

Pierwszy wyraz w rozwinięciu jest energią potencjalną układu nieruchomych atomów znajdujących się w położeniu równowagi. Wyraz ten nie odgrywa roli w zagadnieniach dynamiki sieci krystalicznej.

Wyraz U_1 przedstawia liniową zależność energii potencjalnej od odchylenia atomów od położeń równowagi. Jest on równy zeru, gdyż pierwsza pochodna w minimum funkcji jest równa zeru. Wyraz U_2 reprezentuje zależność kwadratową energii potencjalnej od odchyleń atomów od ich położeń równowagi. Ograniczenie rozwinięcia funkcji energii potencjalnej do wyrazu U_2 stanowi istotę tzw. przybliżenia harmonicznego. Sens tego przybliżenia w wypadku dwóch atomów ilustruje rys. 1, który przedstawia rzeczywisty i przybliżony (paraboliczny) kształt funkcji energii potencjalnej. W krysztale energia potencjalna pochodzi od oddziaływań wielu atomów. Funkcja energii potencjalnej przedstawia złożoną powierzchnię w wielowymiarowej przestrzeni.

przybliżenie harmoniczne



Rys. 1. Zależność energii potencjalnej od odchylenia atomu od położenia równowagi R_0. Linia niebieska odpowiada przybliżeniu harmonicznemu

Funkcja U w przybliżeniu harmonicznym zadowalająco opisuje oddziaływania międzyatomowe wtedy, gdy odchylenia atomów od położeń równowagi są małe w porównaniu z odległościami między atomami. Z wyjątkiem niektórych kryształów warunek ten jest stosunkowo dobrze spełniony w niezbyt wysokich temperaturach. W wielu kryształach w temperaturze pokojowej odchylenia atomów nie przekraczają 0,1 Å, co stanowi tylko kilka procent odległości międzyatomowych.

Równania ruchu

Wiele problemów dynamiki sieci krystalicznej można rozwiązać posługując się metodami mechaniki klasycznej. Punktem wyjścia są newtonowskie równania ruchu ($m\vec{a} = \vec{F}$), w których występują masy, przyspieszenia i siły. Równania tego typu należy napisać dla każdego atomu w sieci krystalicznej. Jest oczywiste, że ze względu na dużą liczbę atomów w krysztale musimy wprowadzić odpowiednie oznaczenia, które pozwolą te atomy rozróżniać. Najodpowiedniejszym sposobem określenia atomu jest podanie numeru komórki elementarnej, w której on się znajduje, oraz numeru atomu w komórce elementarnej. Po tych wyjaśnieniach wprowadzimy następujące oznaczenia niezbędne do dalszych rozważań: m_s — masa atomu s, $\vec{R}_0(ls)$ — położenie równowagi atomu s znajdującego się w komórce elementarnej l, $\vec{u}(ls)$ — jego odchylenie od położenia równowagi, t — czas, α, β — zbiór trzech współrzędnych kartezjańskich x, y, z. Klasyczne równania ruchu w przybliżeniu harmonicznym mają postać:

klasyczne równania ruchu

$$m_s \frac{d^2 u_\alpha(ls)}{dt^2} = \sum_{\substack{l's' \\ \alpha\beta}} B_{\alpha\beta}(ls, l's')\, u_\beta(l's'). \quad (2)$$

Lewa strona równania jest iloczynem masy i przyspieszenia atomu (ls), prawa — przedstawia siłę działającą na ten atom pochodzącą od wszystkich pozostałych atomów $(l's')$. Siła ta zależy liniowo od

stałe siłowe

odchyleń poszczególnych atomów od położeń równowagi. Współczynniki $B_{\alpha\beta}(ls, l's')$ noszą nazwę stałych siłowych. Wyrażają się one przez drugą pochodną energii potencjalnej względem odchyleń atomów od położeń równowagi, obliczoną w położeniu równowagi. W doskonałym krysztale symetria translacyjna powoduje, że stałe siłowe zależą tylko od różnicy wskaźników l i l', a nie zależą od ich wartości wziętych osobno.

Liczba równań w układzie (2) wynosi $3rN$, gdzie N oznacza liczbę komórek elementarnych, r — liczbę atomów w jednej komórce. Wartość ta pokrywa się z liczbą stopni swobody wszystkich atomów w krysztale. Opisanie drgań sieci w makroskopowej próbce krystalicznej objętości 1 cm^3 wymaga olbrzymiej liczby równań ruchu, rzędu 10^{23}.

Warto zauważyć, że prawa strona każdego równania w układzie (2) składa się z dużej liczby wyrazów opisujących oddziaływania międzyatomowe. Zwykle oddziaływania te silnie maleją ze wzrostem odległości między atomami. Można to wykorzystać do znacznego zmniejszenia liczby rozpatrywanych wyrazów.

Istotna trudność, jaką przedstawia zapisany układ równań, wynika z faktu, że każde równanie zawiera odchylenia wszystkich atomów. Inaczej mówiąc, równania w układzie (2) są wzajemnie powiązane. Z fizycznego punktu widzenia oznacza to, że odchyleniu dowolnego atomu w krysztale od położenia równowagi towarzyszy odchylenie innych atomów. Ruch atomów w krysztale ma zatem charakter kolektywny.

współrzędne normalne

W teorii drgań sieci krystalicznej wykazuje się, że zamiast $\vec{u}(ls)$ można wprowadzić nowe zmienne Q_i, które usuwają niedogodną własność układu (2), polegającą na wzajemnym powiązaniu równań. Nowy układ równań składa się z $3rN$ niezależnych równań postaci:

$$\frac{d^2Q_i}{dt^2}+\omega_i^2 Q_i = 0 \quad (i = 1,2,\dots,3rN). \tag{3}$$

Równania tego typu są dobrze znane. Opisują one ruch niezależnych oscylatorów harmonicznych. Rozwiązania ich zawierają harmoniczną zależność od czasu:

$$Q_i = Q_i^0 e^{-i\omega_i t}; \tag{4}$$

ω_i — częstość własna oscylatora. W każdym równaniu w układzie (3) występuje teraz tylko jedna zmienna Q_i, którą nazywamy współrzędną normalną.

Można wykazać, że w doskonałym krysztale, posiadającym symetrię translacyjną, każdy oscylator wysyła falę płaską postaci:

$$\psi_j(\vec{q}) = \vec{e}_s(\vec{q}j)e^{-i\omega j(\vec{q})+i\vec{q}\vec{R}_0(ls)}. \tag{5}$$

Fala ta jest scharakteryzowana przez wektor falowy $\vec{q}$, częstość kołową $\omega_j(\vec{q})$ i wektor polaryzacji $\vec{e}_s(\vec{q}j)$. Wektor falowy $\vec{q}$ określa kierunek rozchodzenia się fali, a jego wartość liczbowa jest równa $2\pi/\lambda$ (λ — długość fali). Wektor polaryzacji $\vec{e}_s(\vec{q}j)$ wskazuje kierunek drgań atomów i zależy od rodzaju atomu (wskaźnik s). Wskaźniki $\vec{q}$ i j, występujące przy częstościach i wektorach polaryzacji, klasyfikują rozwiązania równań ruchu. Do zagadnienia tego jeszcze wrócimy.

Gdy fale rozchodzą się w kierunkach wysokiej symetrii w krysztale (np. w kierunku [111] w kryształach o strukturze regularnej), orientacja wektora polaryzacji w stosunku do wektora falowego jest łatwa do określenia. Drgania można wówczas klasyfikować jako ściśle podłużne (gdy $\vec{e} \parallel \vec{q}$) lub ściśle poprzeczne (gdy $\vec{e} \perp \vec{q}$). Gdy fale rozchodzą się w innych kierunkach, zazwyczaj wektor polaryzacji tworzy z wektorem falowym pewien kąt różny od kąta zerowego lub prostego.

W ramach przybliżenia harmonicznego odchylenia atomów $\vec{u}(ls)$ od położenia równowagi można przedstawić w postaci superpozycji $3rN$ fal płaskich:

$$\vec{u}(ls) = \sum_{j}^{3r}\sum_{\vec{q}}^{N}\psi_j(\vec{q}) = \sum_{j}^{3r}\sum_{\vec{q}}^{N}\vec{e}_s(\vec{q}j)e^{-i\omega_j(\vec{q})t+i\vec{q}\vec{R}_s(ls)}. \tag{6}$$

W równaniu (6) rolę współrzędnych normalnych $Q_j(\vec{q})$ odgrywa wyrażenie:

$$e_s(\vec{q}j)e^{-i\omega_j(\vec{q})t} = Q_j(\vec{q}). \tag{7}$$

W tym wypadku współrzędne normalne interpretujemy jako wielkości określające amplitudy fal płaskich. Ruch opisany przez współrzędne normalne nosi nazwę drgań normalnych. W doskonałej sieci krystalicznej drgania normalne mają postać fal płaskich. W każdym drganiu normalnym biorą udział wszystkie atomy, które drgają z tą samą częstością. Między drganiami poszczególnych atomów istnieje stała różnica faz. Ogólna liczba drgań normalnych jest równa liczbie stopni swobody dla atomów w krysztale ($3rN$).

drgania normalne

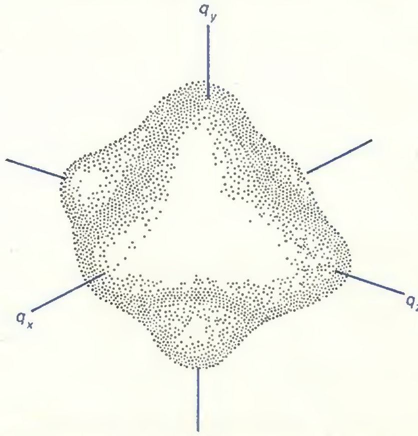
Aby opis drgań sieci krystalicznej w języku drgań normalnych był pełny, pozostaje jeszcze wyznaczyć częstość $\omega_j(\vec{q})$ i wektory polaryzacji $\vec{e}_s(\vec{q}j)$. Można to osiągnąć przez rozwiązanie układu równań ruchu (2). Zgodnie z przedstawionymi wyżej rozważaniami zakłada się, że rozwiązania mają postać fali płaskiej.

Reasumując, współrzędne normalne pozwalają sprowadzić zagadnienie ruchu wzajemnie oddziaływających atomów do zagadnienia ruchu niezależnych oscylatorów harmonicznych. Oscylator harmoniczny jest jednym z niewielu układów, dla którego znane są ściśle rozwiązania, zarówno w ramach mechaniki klasycznej, jak i kwantowej. Nie można jednak zapominać, że wprowadzenie formalizmu oscylatora i drgań normalnych jest związane z kwadratową zależnością energii potencjalnej od odchyleń atomów od położeń równowagi, która stanowi wynik przybliżenia harmonicznego.

Relacje dyspersji

Zależność $\omega_j(\vec{q})$ charakteryzuje podstawowe własności dynamiczne sieci krystalicznej i jest przedmiotem wielu badań, zarówno teoretycznych, jak i doświadczalnych. Funkcja $\omega(\vec{q})$ nosi nazwę relacji dyspersji lub prawa dyspersji.

Ogólnie relacje dyspersji przedstawiają zależność częstości drgań sieci od trzech zmiennych, którymi



Rys. 2. Powierzchnia stałej częstości w przestrzeni wektora falowego $\vec{q}$

powierzchnie stałej częstości

są składowe wektora falowego $\vec{q}$. W trójwymiarowej przestrzeni mogą one być przedstawione graficznie jako powierzchnie stałej częstości (rys. 2). Analiza własności tych powierzchni umożliwia otrzymanie ważnych informacji o strukturze widma częstości drgań. W praktyce jednak ograniczamy się zwykle do rozpatrywania relacji dyspersji tylko dla niektórych, wybranych kierunków w przestrzeni wektora falowego $\vec{q}$ (a tym samym w krysztale). Z reguły są to kierunki odznaczające się wysoką symetrią, jak np. kierunki [111], [110] i [100] w kryształach o strukturze regularnej. Obrazem graficznym relacji dyspersji są

krzywe dyspersji

wówczas krzywe płaskie, które nazywamy krzywymi dyspersji.

Jest zrozumiałe, że kształt krzywych dyspersji winien zależeć od natury kryształu, a w szczególności od charakteru oddziaływań międzyatomowych w krysztale. Interesują nas jednak przede wszystkim te własności relacji dyspersji, które są wspólne dla wszystkich kryształów, niezależnie od typu oddziaływań międzyatomowych. Do takich ogólnych wła-

okresowość i parzystość

sności relacji dyspersji należy okresowość i parzystość. Własności te oznaczają, że spełnione są następujące zależności:

$$\omega_j(\vec{q}) = \omega_j(\vec{q}+\vec{G}) \qquad (8)$$

$$\omega_j(\vec{q}) = \omega_j(-\vec{q}); \qquad (9)$$

$\vec{G}$ jest wektorem sieci odwrotnej. Okresowość funkcji $\omega_j(\vec{q})$ jest następstwem istnienia symetrii translacyjnej w krysztale, która powoduje, że wszystkie nierównoważne fizycznie drgania normalne odpowiadają wartościom $\vec{q}$ położonym w pewnym ograniczonym obszarze. Podobnie jak w zagadnieniach struktury elektronowej kryształów, obszar ten wygodnie jest wybrać w postaci tzw. strefy Brillouina. W trójwymiarowej przestrzeni $\vec{q}$ strefa Brillouina jest zamkniętym wielościanem, którego kształt odzwierciedla symetrię kryształu.

wieloznaczność

Następną charakterystyczną cechą relacji dyspersji jest wieloznaczność. Oznacza to, że każdej wartości wektora falowego $\vec{q}$ odpowiada więcej niż jedna wartość częstości ω. W związku z tym krzywe dyspersji składają się z gałęzi, które rozróżniamy przy pomocy wskaźnika j.

Ogólna liczba gałęzi drgań w trójwymiarowym krysztale jest równa $3r$, gdzie r jest liczbą atomów w komórce elementarnej. Analiza ruchu drgającego prowadzi do klasyfikacji tych gałęzi na akustyczne i optyczne. Liczba gałęzi akustycznych jest równa 3, a liczba gałęzi optycznych $3r-3$. Tak np. w krysztale trójwymiarowym, którego komórka elementarna zawiera jeden atom, występują tylko trzy gałęzie akustyczne (nie ma gałęzi optycznych). W krysztale trójwymiarowym z dwoma atomami w komórce elementarnej jest ogółem sześć gałęzi, z których trzy są gałęziami akustycznymi i trzy — gałęziami optycznymi. Własności symetrii kryształu mogą powodować zwyrodnienie niektórych z tych gałęzi, co redukuje ich liczbę.

Pozostaje teraz sprecyzować, czym różnią się między sobą gałęzie akustyczne i optyczne. Różnica ta najwyraźniej występuje w środku strefy Brillouina, dlatego omówimy własności drgań długofalowych ($\vec{q} \approx 0$).

drgania akustyczne

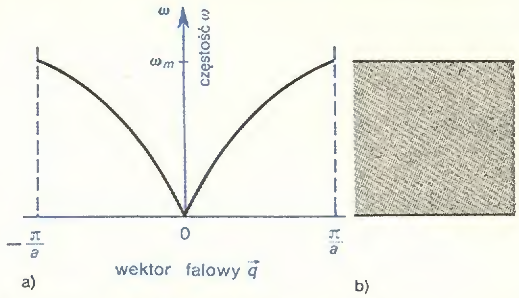
Drgania akustyczne odpowiadają sytuacji, gdy wszystkie atomy w komórce elementarnej poruszają się w zgodnej fazie. Gdy $\vec{q} = 0$, częstość drgań akustycznych jest równa zeru, a przy małych wartościach $\vec{q}$ zmienia się liniowo ze wzrostem $\vec{q}$. W pobliżu środka strefy Brillouina drgania akustyczne nie różnią się od drgań towarzyszących rozchodzeniu się fali akustycznej w krysztale. Fakt ten znajduje odbicie w nazwie, która została przyjęta dla omawianych drgań.

drgania optyczne

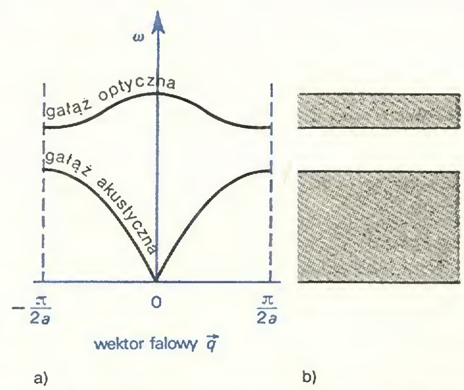
W przypadku drgań optycznych sąsiednie atomy w komórce elementarnej poruszają się w fazach przeciwnych, przy tym środek masy komórki elementarnej pozostaje w spoczynku. Oznacza to, że amplituda drgań atomów jest odwrotnie proporcjonalna do ich masy. Gdy $\vec{q} = 0$, częstości drgań optycznych dążą zawsze do skończonej wartości, różnej od zera. W kryształach jonowych drgania optyczne wytwarzają oscylujący moment elektryczny. Ponieważ z tym momentem związane są silne oddziaływania optyczne (pochłanianie), drgania tego typu otrzymały nazwę drgań optycznych.

krzywe dyspersji — kryształy jednowymiarowe

Relacje dyspersji w postaci nie skomplikowanego wyrażenia matematycznego można otrzymać tylko dla kryształów jednowymiarowych. Wykresy typowych zależności dla tych kryształów są przedstawione na rys. 3a i 4a. Zgodnie z poprzednio omawianą wła-



Rys. 3. Krzywe dyspersji (a) i pasmo dozwolonych częstości drgań (b) sieci liniowej z jednym atomem w komórce elementarnej



Rys. 4. Krzywe dyspersji (a) i pasma dozwolonych częstości drgań (b) sieci liniowej z dwoma atomami w komórce elementarnej

snością okresowości funkcji $\omega_j(\vec{q})$ zakres zmienności wektora falowego $\vec{q}$ jest ograniczony do przedziału $(-\pi/a, +\pi/a)$ lub $(-\pi/2a, +\pi/2a)$, który w tym wypadku określa strefę Brillouina (a — odległość między atomami). W krysztale mającym 1 atom w komórce elementarnej występuje tylko gałąź akustyczna, a w krysztale z dwoma atomami w komórce elementarnej — gałąź akustyczna i optyczna. Liczba gałęzi drgań jest mniejsza niż w krysztale trójwymiarowym; jest to związane z mniejszą liczbą stopni swobody przypadających na jeden atom.

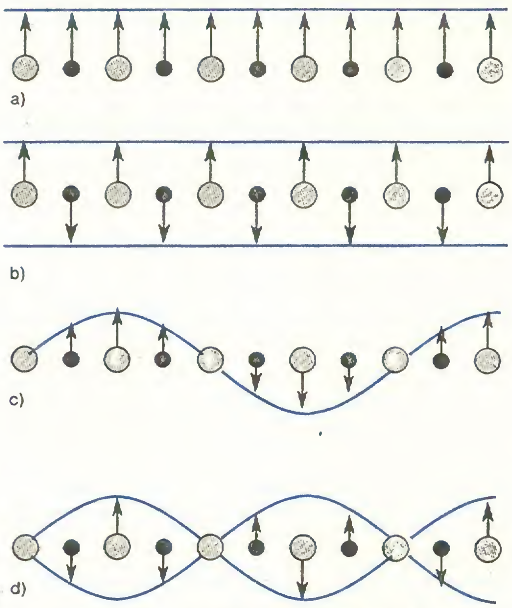
częstość drgań normalnych

Warto zauważyć, że częstości drgań normalnych sieci przybierają tylko wartości leżące w ograniczonym przedziale, zawartym między 0 i pewną maksymalną częstością ω_m. W najprostszym wypadku cały obszar dozwolonych częstości tworzy jedno pasmo (rys. 3b). W kryształach zawierających dwa atomy w komórce elementarnej mogą występować dwa pasma rozdzielone przerwą, z których jedno odpowiada drganiom akustycznym, drugie — optycznym (rys. 4b). Struktura pasmowa i istnienie maksymalnej częstości drgań jest cechą ogólną, niezależną od wymiarowości kryształu. Maksymalna częstość drgań

optycznych w kryształach rzeczywistych jest zwykle rzędu $10^{13}\ s^{-1}$.

analiza drgań — kryształy jednowymiarowe

Analizę charakteru drgań można stosunkowo łatwo przeprowadzić dla kryształów jednowymiarowych. Wyniki dotyczące drgań poprzecznych w krysztale zawierającym dwa atomy w komórce elementarnej są przedstawione na rys. 5. Wzięto pod uwagę dwie



Rys. 5. Ruch falowy w sieci liniowej zawierającej dwa atomy w komórce elementarnej: a) drganie poprzeczne akustyczne w środku strefy Brillouina ($\vec{q} = 0$); b) drganie poprzeczne optyczne w środku strefy Brillouina ($\vec{q} = 0$); c) drganie poprzeczne akustyczne wewnątrz strefy Brillouina; d) drganie poprzeczne optyczne wewnątrz strefy Brillouina

wartości wektora falowego $\vec{q}$; jedna odpowiada środkowi strefy Brillouina ($\vec{q} = 0$), druga znajduje się wewnątrz strefy, w pewnej odległości od jej środka.

krzywe dyspersji — kryształy trójwymiarowe

Otrzymanie krzywych dyspersji dla rzeczywistych trójwymiarowych kryształów (rys. 6) wymaga skomplikowanych obliczeń, do których są niezbędne maszyny matematyczne. Przy obliczeniach zakłada się odpowiedni model oddziaływań międzyatomowych. Stosunkowo prosta sytuacja jest wtedy, gdy oddziaływania międzyatomowe maleją szybko ze wzrostem odległości. Pozwala to rozpatrywać tylko niewielką liczbę stałych siłowych, które można wyrazić przez współczynniki sprężystości lub inne dane doświadczalne. W wielu kryształach jednak odgrywają istotną rolę oddziaływania dalekiego zasięgu, które muszą być uwzględnione przy obliczeniach, jeżeli chcemy uzyskać krzywe dyspersji zgodne z wynikami pomiarów. W zależności od natury badanych kryształów stosuje się różne modele oddziaływań międzyatomowych.

metody eksperymentalne

Podstawową metodą eksperymentalną, dzięki której można otrzymać krzywe dyspersji, jest badanie niesprężystego rozproszenia neutronów powolnych. Porównanie obliczonych krzywych dyspersji z wynikami pomiarów traktuje się jako metodę kontroli poprawności przyjętego modelu oddziaływań międzyatomowych. Zagadnienie niesprężystego rozpraszania neutronów powolnych na drganiach sieci omówimy w dalszej części tego artykułu, w rozdziale „Doświadczalne metody badania drgań sieci".

Warto jeszcze zwrócić uwagę na różnice, jakie istnieją między falą sprężystą w sieci krystalicznej i falą sprężystą w ośrodku ciągłym, nie mającym dyskretnej struktury atomowej. Jak to powyżej przedstawiliśmy, drgania sieci krystalicznej mają ograniczony zakres zmienności częstości i wektora falowego, a zależność między tymi wielkościami wykazuje dyspersję. W wypadku fali sprężystej rozchodzącej się w ośrodku ciągłym częstość i wektor falowy mogą w zasadzie przybierać dowolne wartości, a zależność $\omega(\vec{q})$ jest liniowa.

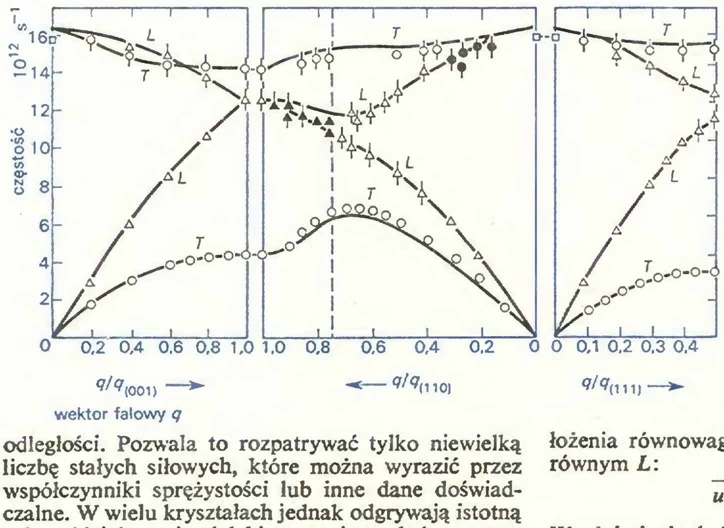
Widmo wartości wektora falowego

wpływ powierzchni

W dotychczasowych rozważaniach wychodziliśmy z założenia, że kryształ jest nieskończony. W rzeczywistości wszystkie kryształy są ograniczone powierzchniami i tym samym mają skończone rozmiary. Jest oczywiste, że atomy znajdujące się na powierzchni lub w jej pobliżu drgają inaczej niż atomy znajdujące się wewnątrz kryształu, daleko od powierzchni (inne stałe siłowe). Drgania powierzchniowe stanowią przedmiot specjalnych badań (→ Stany powierzchniowe w ciałach stałych).

W tej chwili interesuje nas jednak, jak obecność powierzchni wpływa na drgania sieci krystalicznej wewnątrz kryształu. Można przypuszczać, że jeśli oddziaływania międzyatomowe szybko maleją ze wzrostem odległości, to atomy położone daleko od powierzchni „nie czują" jej obecności. Dlatego w rzeczywistych kryształach warunki brzegowe, jakimi się posługujemy przy rozwiązywaniu równań ruchu, nie mają wielkiego znaczenia. W tej sytuacji możemy je tak sformułować, aby maksymalnie ułatwić rozwiązanie równań ruchu. Najwygodniejszą postać pod tym względem mają warunki cykliczności Borna–Kármána. Zgodnie z nimi odchylenia atomów od po-

warunki Borna–Kármána



Rys. 6. Krzywe dyspersji drgań sieci krystalicznej krzemu dla trzech kierunków krystalograficznych. Linie ciągłe przedstawiają wyniki obliczeń, a punkty — wyniki pomiarów metodą dyfrakcji neutronów powolnych. Pionowa linia przerywana na środkowym rysunku określa granicę strefy Brillouina. Litery L i T oznaczają odpowiednio gałąź drgań podłużnych i poprzecznych. W rozpatrywanych kierunkach krystalograficznych gałęzie poprzeczne są zwyrodniałe

łożenia równowagi muszą się powtarzać z okresem równym L:

$$\vec{u}(ls) = \vec{u}(l+L, s). \qquad (10)$$

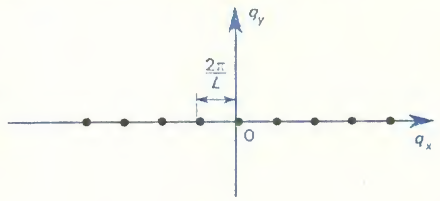
W odniesieniu do kryształu mającego postać sześcianu o krawędzi L oznacza to, że odchylenia atomów po-

łożonych w punktach symetrycznych na powierzchniach sześcianu są identyczne (jednakowe amplitudy i fazy). Warunki cykliczności Borna–Kármána pozwalają przedstawić ruch atomów w skończonym krysztale w postaci fal płaskich bieżących, podobnie jak w krysztale nieograniczonym.

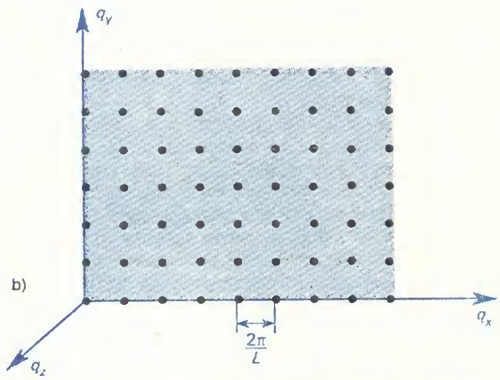
dyskretne wartości wektora falowego

Można łatwo wykazać, że z warunków cykliczności wynika dyskretny zbiór wartości, które może przybierać wektor falowy $\vec{q}$ wewnątrz strefy Brillouina. Wartości te są rozmieszczone w sposób jednorodny: w każdym z trzech kierunków są one oddalone od siebie o wartość równą $2\pi/L$ (rys. 7).

a)



b)



Rys. 7. Rozmieszczenie dozwolonych stanów w przestrzeni $\vec{q}$ wynikające z cyklicznych warunków brzegowych: a) sieć liniowa, b) sieć trójwymiarowa (pokazano tylko stany położone na płaszczyźnie q_zq_y)

Całkowita liczba dyskretnych wartości $\vec{q}$ w strefie Brillouina jest równa liczbie komórek elementarnych N w krysztale. W makroskopowych kryształach gęstość dyskretnych wartości $\vec{q}$ jest tak duża, że można traktować $\vec{q}$ jako zmienną ciągłą. Pozwala to przy obliczeniach zastępować sumowanie po dyskretnych wartościach $\vec{q}$ bardziej wygodnym całkowaniem po zmiennej $\vec{q}$. Przykładem traktowania $\vec{q}$ jako zmiennej ciągłej są omawiane przed chwilą krzywe dyspersji (rys. 6).

Kwantowanie drgań sieci

Przedstawiony dotychczas obraz drgań sieci krystalicznej otrzymano przy zastosowaniu praw mechaniki klasycznej. Może powstać pytanie, czy wyniki te nie są przybliżone, a w związku z tym, czy nie ulegną zmianie, gdy zastosujemy metody mechaniki kwantowej. Okazuje się, że nie. Oscylator harmoniczny stanowi taki układ mechaniczny, którego częstość własna ω jest zawsze taka sama — niezależnie od tego, czy otrzymano ją metodami mechaniki klasycznej, czy kwantowej. Własność ta rozciąga się na drgania sieci krystalicznej, które są równoważne zbiorowi niezależnych oscylatorów harmonicznych.

kwantowanie oscylatora harmonicznego

W tej sytuacji zwykle najpierw rozwiązujemy klasyczne równania ruchu i otrzymujemy częstości drgań (oraz krzywe dyspersji i widmo częstości), a następnie stosujemy procedurę kwantowania oscylatora harmonicznego.

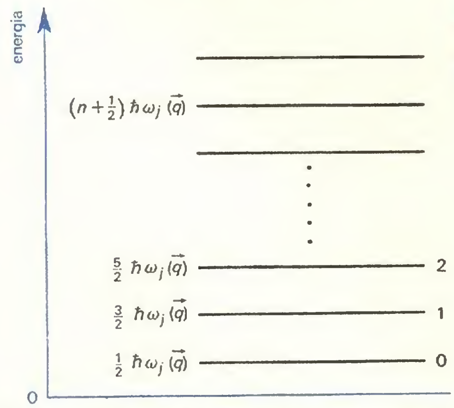
Mechanika kwantowa przewiduje, że oscylator harmoniczny może przybierać tylko dyskretne wartości energii, określone wzorem:

$$E_{n(\vec{q}j)} = [n(\vec{q}j)+\tfrac{1}{2}]\hbar\omega_j(\vec{q}), \tag{11}$$

gdzie $n(\vec{q}j) = 0, 1, 2 \ldots$, $\hbar$ — stała Plancka podzielona przez 2π. Schemat dozwolonych poziomów energetycznych wynikających z tego wzoru jest przedstawiony na rys. 8.

poziomy oscylatora harmonicznego

Każdemu dozwolonemu poziomowi energetycznemu odpowiada pewna liczba kwantowa $n(\vec{q}j)$, która



Rys. 8. Diagram poziomów energetycznych oscylatora harmonicznego

może przybierać tylko wartości całkowite nieujemne. Odległość między sąsiednimi poziomami energii jest równa $\hbar\omega_j(\vec{q})$. Wielkość ta jest najmniejszą porcją (kwantem) energii, która może być pobrana lub oddana przy zmianie stanu energetycznego oscylatora.

fonon

Kwant energii drgań sieci krystalicznej nosi nazwę fononu. Oscylatorowi związanemu z drganiami sieci krystalicznej znajdującemu się w stanie kwantowym $n(\vec{q}j)$ odpowiada $n(\vec{q}j)$ fononów, z których każdy ma tę samą wartość energii, równą $\hbar\omega_j(\vec{q})$. Liczby kwantowe $n(\vec{q}j)$ często nazywamy liczbami obsadzenia.

drgania zerowe

Według mechaniki klasycznej każdy oscylator może mieć dowolną wartość energii, która jest proporcjonalna do kwadratu amplitudy drgań. W temperaturze zera bezwzględnego amplituda i energia drgań są równe zeru. Natomiast w opisie mechaniki kwantowej w temperaturze zera bezwzględnego pozostają tzw. drgania zerowe o energii $\frac{1}{2}\hbar\omega_j(\vec{q})$, które odpowiadają stanowi podstawowemu oscylatora określonemu liczbą kwantową $n(\vec{q}j) = 0$. Występowanie drgań zerowych można wyjaśnić na gruncie zasady nieokreśloności Heisenberga. Zgodnie z tą zasadą atomy nie mogą być nieruchome, gdyż oznaczałoby to, że ich położenie jest dokładnie określone, a pęd i energia może mieć nieskończenie dużą wartość.

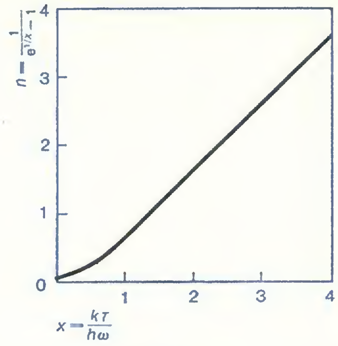
Fonony podlegają kwantowej statystyce Bosego–Einsteina. W odróżnieniu od statystyki Fermiego–Diraca, która ogranicza liczbę elektronów w tym samym stanie kwantowym do dwóch (o przeciwnie skierowanych spinach), statystyka Bosego–Einsteina nie zabrania, aby w tym samym stanie kwantowym $\vec{q}j$ znajdowała się dowolna liczba fononów. Zgodnie z mechaniką statystyczną średnia wartość liczby obsadzenia $\langle n(\vec{q}j)\rangle$ dla oscylatora sieciowego $\vec{q}j$ w warunkach równowagi termodynamicznej dana jest przez następujące wyrażenie:

$$\langle n(\vec{q}j)\rangle = \frac{1}{e^{\hbar\omega_j(\vec{q})/kT}-1}, \tag{12}$$

gdzie k — stała Boltzmanna, T — temperatura bezwzględna.

średnia liczba fononów

Ze wzoru (12) widać, że średnia liczba obsadzenia (równoważna średniej liczbie fononów) zależy od temperatury. Kształt tej zależności przedstawia rys. 9. Można zauważyć, że w dostatecznie wysokich tem-



Rys. 9. Wykres funkcji rozkładu Bosego-Einsteina

peraturach ($\hbar\omega_j(\vec{q}) \ll kT$) zależność między średnią liczbą fononów i temperaturą ma liniowy charakter.

Należy podkreślić, że wszystkie dotychczasowe rozważania odnoszą się do fononów wzbudzonych termicznie. Można jednak generować fonony za pomocą promieniowania elektromagnetycznego pochłanianego w krysztale lub inaczej. Częstości i wektory falowe wytwarzanych wówczas fononów przybierają tylko niektóre spośród możliwych wartości. Nie stosuje się do nich prawo rozkładu statystycznego (12). Liczba generowanych fononów jest funkcją natężenia padającego na kryształ promieniowania i nie zależy od temperatury kryształu. W odróżnieniu od drgań sieci wzbudzonych termicznie, które są niespójne, promieniowanie elektromagnetyczne (pochodzące z lasera) może wytwarzać drgania normalne posiadające własność spójności.

fonony generowane promieniowaniem elektromagnetycznym

gaz fononowy

Przy opisie różnych procesów, w których odgrywają rolę drgania sieci, posługujemy się często modelem gazu fononowego. Model ten można wprowadzić opierając się na zasadzie dualizmu korpuskularno-falowego, stanowiącej jedno z podstawowych założeń mechaniki kwantowej. W myśl tej zasady każdemu drganiu normalnemu o częstości ω_j i wektorze falowym $\vec{q}$ można przypisać określoną liczbę nierozróżnialnych kwazicząstek (fononów) o energii $\hbar\omega_j(\vec{q})$ i kwazipędzie $\hbar\vec{q}$. W ten sposób zbiór drgań normalnych jest zastąpiony przez zbiór kwazicząstek. W ramach przybliżenia harmonicznego kwazicząstki stanowią gaz idealny, w którym nie występują wzajemne oddziaływania.

oddziaływanie fononów z innymi kwazicząstkami

Oddziaływanie drgań sieci z innymi kwazicząstkami (np. elektronem w krysztale) można traktować jako zderzenie, dla którego obowiązuje zasada zachowania energii i pędu. W oddziaływaniach tych fonony mogą powstawać i znikać. Oddziaływanie drgań sieci między sobą odpowiada w tym obrazie zderzeniu fononów. Dla pojawienia się tego oddziaływania konieczne jest jednak wyjście poza ramy przybliżenia harmonicznego.

Można się oczywiście domyślać, że analogia między kwazicząstkami i zwykłymi cząstkami, takimi jak elektrony, protony, neutrony, jest tylko częściowa. Wskazuje już na to posługiwanie się terminami kwazicząstka i kwazipęd.

W odróżnieniu od zwykłych cząstek kwazicząstki nie mogą istnieć w próżni. Do ich powstania i istnienia niezbędny jest pewien ośrodek. Dla fononów takim ośrodkiem jest kryształ. Fonon jako kwazicząstka nie może być „wyjęty" z kryształu.

Następna istotna różnica wiąże się z pojęciem pędu. Jak wiadomo, każda rzeczywista cząstka ma określony pęd, a w oddziaływaniach rzeczywistych cząstek obowiązuje zasada zachowania pędu. Fononowi nie można przypisać pędu fizycznego, ale w oddziaływaniach zachowuje się on tak, jakby rolę pędu odgrywała wielkość $\hbar\vec{q}$. Niekiedy rola ta jest w pewnym stopniu ograniczona, gdyż zasada zachowania pędu dla oddziaływań z fononami zawiera wyrazy postaci $\hbar\vec{q}+\hbar\vec{G}$ ($\vec{G}$ — wektor sieci odwrotnej). W związku z tym mówimy czasem, że $\hbar\vec{q}$ określa pęd fononu z dokładnością do wielkości $\hbar\vec{G}$. Sens fizyczny wyrazu $\hbar\vec{G}$ można wyjaśnić przyjmując możliwość przekazania pędu do całej sieci krystalicznej. Przedstawione wyżej uwagi o znaczeniu wielkości $\hbar\vec{q}$ w dynamice sieci krystalicznej uzasadniają posługiwanie się terminem kwazipęd fononu.

kwazipęd fononu

cząstka a kwazicząstka

Warto jeszcze zwrócić uwagę na niektóre dalsze różnice między cząstkami i kwazicząstkami. Przejawiają się one w formie zależności energii od pędu. Jak wiadomo, dla zwykłych cząstek zależność ta ma charakter uniwersalny: dla fotonu jest ona liniowa, a dla cząstek nierelatywistycznych — kwadratowa. Jeśli chodzi o fonony, to nie mają one takiej uniwersalnej zależności energii od pędu. W każdym krysztale jest ona inna i zależy od natury oddziaływań międzyatomowych. Warto dodać, że zależność energii od kwazipędu dla fononów jest analogiczna do rozpatrywanej przez nas wcześniej zależności częstości drgań normalnych od wektora falowego, reprezentowanej przez krzywe dyspersji.

Mimo pewnych różnic występujących między rzeczywistymi cząstkami i kwazicząstkami (w szczególności fononami) posługiwanie się pojęciem kwazicząstki jest bardzo wygodne przy opisie różnych oddziaływań zachodzących w krysztale. Jako przykład można wymienić oddziaływanie między drganiami sieci, oddziaływanie drgań sieci z elektronami w krysztale oraz z polem elektromagnetycznym (z fotonami).

Widmo częstości drgań

Przy analizie niektórych własności kryształów związanych z drganiami sieci wygodnie jest wprowadzić pojęcie rozkładu częstości drgań lub widma częstości drgań sieci. Widmo częstości drgań można przedstawić w postaci pewnej funkcji, którą oznaczymy przez $g(\omega)$. Funkcja ta jest zdefiniowana w ten sposób, że wyrażenie $g(\omega)\,d\omega$ określa liczbę drgań normalnych przypadających na przedział częstości $(\omega, \omega+d\omega)$. Funkcja $g(\omega)$ zwykle spełnia warunek normalizacji, wynikający z żądania, aby liczba wszystkich drgań normalnych równała się liczbie oscylacyjnych stopni swobody w krysztale.

warunek normalizacji funkcji $g(\omega)$

Stosunkowo łatwo można obliczyć $g(\omega)$ dla sieci jednowymiarowych. Schematyczny kształt tej funkcji jest przedstawiony na rys. 10. Widzimy, że dla dwuatomowej sieci jednowymiarowej widmo częstości drgań składa się z dwóch części, z których jedna odpowiada drganiom akustycznym, druga — drga-



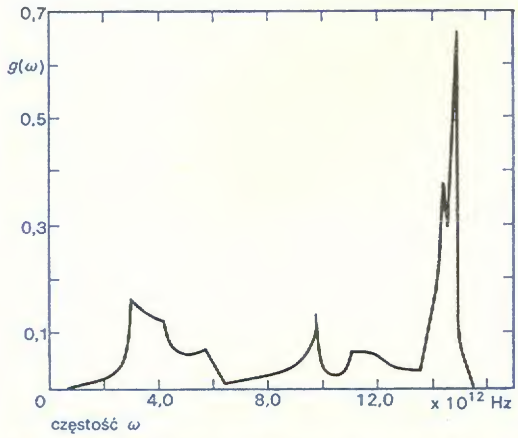
Rys. 10 Widmo drgań sieci liniowej: a) z jednym atomem w komórce elementarnej, b) z dwoma atomami w komórce

niom optycznym. Można również zauważyć, że przy pewnych wartościach ω funkcja ta dąży do nieskończoności, ale nie powinno to dziwić, gdyż rozpatrywany model jest nierealny.

$g(\omega)$ w kryształach rzeczywistych

W rzeczywistych trójwymiarowych kryształach obliczenia widma częstości drgań wykonuje się wyłącznie metodami numerycznymi za pomocą elektronowych maszyn matematycznych. W tym celu rozwiązuje się równania ruchu dla dostatecznie gęsto rozmieszczonych wartości wektora falowego $\vec{q}$ w strefie Brillouina i następnie zlicza się drgania z częstościami zawartymi w przedziałach $(\omega, \omega + d\omega)$.

Widmo częstości drgań ma zazwyczaj złożoną strukturę, która zależy od rodzaju kryształu. Rys. 11 przedstawia wyniki obliczeń dla krzemu. W wypadku



Rys. 11. Widmo częstości drgań sieci krzemu (obliczone)

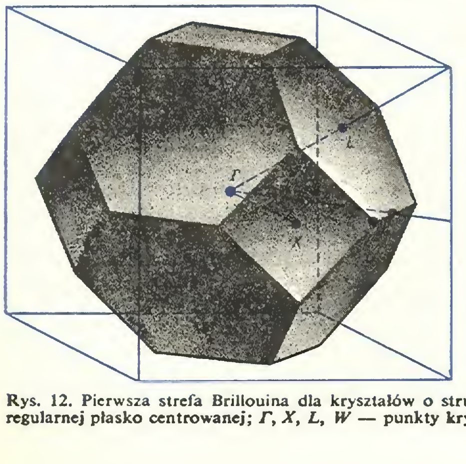
kryształów trójwymiarowych funkcja $g(\omega)$ nie ma nieciągłości typu nieskończoności, jak to miało miejsce dla sieci liniowych. Przy dokładnych obliczeniach otrzymuje się jednak na krzywych $g(\omega)$ charakterystyczne ostre załamania.

Teoria drgań sieci pozwala ustalić interesujące związki między funkcją $g(\omega)$ i relacjami dyspersji $\omega(\vec{q})$. Można mianowicie wykazać, że osobliwości funkcji $g(\omega)$ odpowiadają określonym punktom w strefie Brillouina. Noszą one nazwę punktów krytycznych.

punkty krytyczne

Położenie punktów krytycznych wyznacza się z warunków określających miejsca w strefie Brillouina, w których otoczeniu powierzchnie stałej częstości mają kształt trójwymiarowego maksimum, minimum lub siodła. Po odpowiednich obliczeniach otrzymujemy kształt funkcji $g(\omega)$ w otoczeniu punktów krytycznych związanych z różnymi kształtami powierzchni stałej częstości.

Z rozważań teoretycznych wynika, że liczba różnego typu punktów krytycznych i odpowiadających

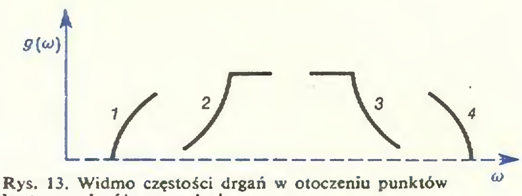


Rys. 12. Pierwsza strefa Brillouina dla kryształów o strukturze regularnej płasko centrowanej; Γ, X, L, W — punkty krytyczne

im osobliwości funkcji $g(\omega)$ jest określona wyłącznie przez symetrię kryształu, a nie zależy od charakteru oddziaływań międzyatomowych.

osobliwości funkcji $g(\omega)$

Punkty krytyczne tego rodzaju występują zwykle w strefie Brillouina w miejscach odznaczających się wysoką symetrią. Na rys. 12 wskazane są położenia punktów krytycznych w strefie Brillouina dla kryształów o strukturze soli kamiennej, blendy cynkowej i diamentu. Są one oznaczone symbolami Γ, X, L i W. Rysunek 13 przedsta-



Rys. 13. Widmo częstości drgań w otoczeniu punktów krytycznych różnego rodzaju

wia schematycznie kształt funkcji $g(\omega)$ w sąsiedztwie różnych punktów krytycznych. Krzywa *1* i *4* odpowiada sytuacji, gdy na powierzchni stałej częstości występuje odpowiednio minimum i maksimum; krzywa *2* i *3* — gdy powierzchnia ta ma kształt siodła.

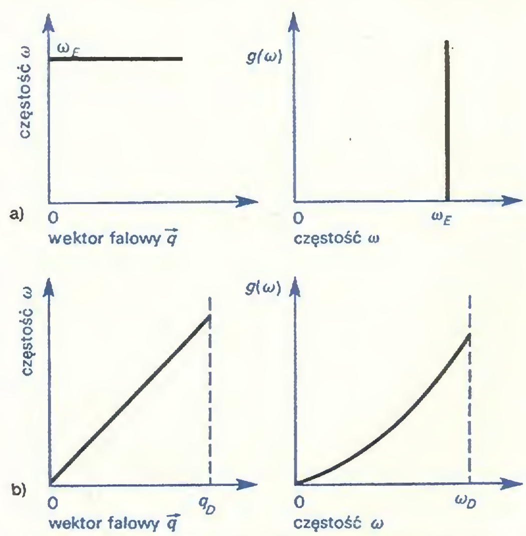
Opierając się na wynikach teorii punktów krytycznych, można przyjąć, że widmo częstości drgań każdego kryształu składa się z szeregu przyczynków pochodzących od punktów krytycznych różnego rodzaju. Po tych uwagach staje się jasne pochodzenie charakterystycznych załamań widocznych na krzywej $g(\omega)$ przedstawionej na rys. 11.

własności termodynamiczne kryształu

Określenie własności termodynamicznych kryształów (m.in. ciepła właściwego) nie wymaga tak dokładnej znajomości kształtu widma częstości drgań. Stosunkowo dobre wyniki otrzymujemy posługując się bardzo przybliżonymi funkcjami $g(\omega)$, otrzymanymi dla bardzo prostych modeli dynamicznych kryształu.

model Einsteina

Jednym z takich modeli jest model Einsteina, który się opiera na założeniu, że każdy atom w sieci jest niezależnym oscylatorem. Widmo częstości drgań składa się z jednej wartości ω_E, wspólnej dla wszystkich oscylatorów (rys. 14a). Częstość tę nazywamy częstością Einsteina. W modelu Einsteina drgania



Rys. 14. Krzywe dyspersji i widmo częstości drgań; a) model Einsteina (ω_E — częstość Einsteina); b) model Debye'a (ω_D — częstość Debye'a)

sieci nie mają dyspersji. W związku z tym model ten źle opisuje drgania akustyczne, odznaczające się dużą dyspersją. Można go jednak stosować do drgań optycznych, które z reguły wykazują słabą zależność częstości od wektora falowego.

model Debye'a

Bardziej odpowiednim do opisu drgań akustycznych jest model Debye'a, który traktuje kryształ jako ośrodek ciągły i izotropowy. Relacje dyspersji są w tym wypadku liniowe, a widmo częstości ma postać $g(\omega) = \text{const}\cdot\omega^2$ (rys. 14b). Maksymalna częstość ω_D (częstość Debye'a) jest określona z warunku normalizacji (całkowita liczba drgań w przedziale od zera do ω_D musi być równa $3rN$). Jest to pewnego rodzaju odstępstwo od własności ośrodka ciągłego, który nie ogranicza liczby drgań normalnych. W modelu Debye'a strefa Brillouina ma kształt kuli, której promień $\vec{q}_D$, zwany wektorem falowym Debye'a, odpowiada w przybliżeniu rozmiarom strefy Brillouina dla rzeczywistych kryształów.

Model Debye'a stosunkowo dobrze opisuje drgania akustyczne, którym odpowiada długość fali większa od odległości międzyatomowych. Kryształ może być wtedy traktowany jako ośrodek ciągły, a krzywe dyspersji mają charakter liniowy.

Drgania akustyczne o stosunkowo małych częstościach są znacznie silniej wzbudzane niż drgania optyczne. W związku z tym decydujący wkład do energii wewnętrznej kryształu jest związany z drganiami akustycznymi. Dzięki temu model Debye'a daje dobre wyniki przy obliczeniach wielkości termodynamicznych. Między innymi pozwala on uzyskać poprawną zależność ciepła właściwego kryształów od temperatury.

Oddziaływania anharmoniczne

Konsekwencją przybliżenia harmonicznego jest możliwość przedstawienia drgań sieci w postaci zbioru niezależnych drgań normalnych. W obrazie tym energia dostarczona selektywnie jednemu drganiu normalnemu nie może być przekazana innym drganiom normalnym. Tym samym w doskonałym krysztale nie istnieje mechanizm, który by gwarantował przywrócenie równowagi termodynamicznej.

W opisie kwantowym niezależnym drganiom normalnym odpowiadają fonony, między którymi nie zachodzą oddziaływania. Czas życia i droga swobodna fononów w doskonałym, nieograniczonym krysztale powinny więc osiągać nieskończenie dużą wartość.

W rzeczywistości jednak żaden kryształ nie jest ściśle harmoniczny. Dokładna teoria dynamiki sieci winna uwzględniać wyrazy trzeciego rzędu i wyższych w rozwinięciu energii potencjalnej względem odchyleń od położeń równowagi. Wyrazy te nazywamy anharmonicznymi.

wyrazy anharmoniczne

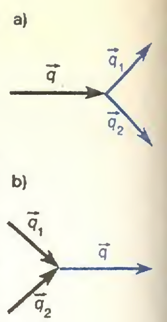
Z wyrazami anharmonicznymi wiąże się szereg makroskopowych własności kryształów, takich jak przewodnictwo cieplne, rozszerzalność cieplna, zależność od temperatury współczynników sprężystości oraz liniowa zależność od temperatury ciepła właściwego w zakresie wysokich temperatur. Nie będziemy się tutaj zajmować zastosowaniami teorii anharmonicznej drgań sieci do wyjaśnienia wymienionych własności. Ograniczymy się wyłącznie do przedstawienia ogólnej interpretacji fizycznej wyrazów anharmonicznych.

Dla szeregu kryształów współczynniki występujące przy wyrazach anharmonicznych w rozwinięciu energii potencjalnej są małe w porównaniu ze współczynnikami przy wyrazach drugiego rzędu. W tej sytuacji wyrazy anharmoniczne można traktować jako małe zaburzenie drgań normalnych doskonałego kryształu. Odpowiednie obliczenia wykazują, że zaburzenie to powoduje niewielką zmianę częstości drgań sieci (w porównaniu z wynikami teorii harmonicznej) oraz pojawienie się czasu relaksacji, czyli powrotu wzbudzonych drgań normalnych do stanu równowagi termodynamicznej.

czas życia fononów

W ujęciu kwantowym efektem wyrazów anharmonicznych jest pojawienie się oddziaływań między fononami, w wyniku czego fonony mogą powstawać i znikać. Prowadzi to do skończonej wartości czasu życia i drogi swobodnej fononów. Średni czas życia fononów jest równoważny czasowi relaksacji energii wzbudzonych drgań normalnych.

Rys. 15. Dwa rodzaje procesów trójfononowych: a) absorpcja jednego fononu i emisja dwóch innych fononów, b) absorpcja dwóch fononów i emisja jednego fononu



Zarówno częstość, jak i czas życia fononów zależą od wartości współczynników anharmonicznych i temperatury. Ze wzrostem temperatury czas życia fononów maleje, gdyż rośnie liczba fononów i w następstwie tego — prawdopodobieństwo oddziaływań między nimi. Dla wielu kryształów efekt ten jest stosunkowo mały. Kształt krzywych dyspersji i widma częstości drgań zaledwie w niewielkim stopniu zależy od oddziaływań anharmonicznych. Dlatego obliczenia tych wielkości prowadzi się z reguły w ramach przybliżenia harmonicznego.

Bardzo często można się ograniczyć do uwzględnienia pierwszego wyrazu anharmonicznego w rozwinięciu energii potencjalnej. Wyraz ten reprezentuje dwa typy procesów, które schematycznie są przedstawione na rys. 15. W procesie (a) jeden fonon znika, a dwa inne powstają, w procesie (b) dwa fonony znikają, a na ich miejsce powstaje jeden inny. W oddziaływaniach tych spełnione są zasady zachowania energii i kwazipędu, które dla procesu (b) mają postać:

$$\hbar\omega_1 + \hbar\omega_2 = \hbar\omega,$$
$$\vec{q}_1 + \vec{q}_2 = \vec{q}, \tag{13}$$
$$\vec{q}_1 + \vec{q}_2 = \vec{q} + \vec{G}.$$

Ze względu na zasadę zachowania pędu możliwe są dwa rodzaje oddziaływań. Jedno z nich nosi nazwę procesu normalnego, drugie — procesu przerzutu lub procesu *Umklapp*. Charakterystyczną cechą procesu przerzutu jest to, że część pędu określona wielkością $\hbar\vec{G}$ jest przekazana całej sieci krystalicznej. Procesy przerzutu odgrywają dużą rolę w zagadnieniach dotyczących mechanizmu ustalania się równowagi termodynamicznej w krysztale, mechanizmu, który leży u podstaw teorii sieciowego przewodnictwa cieplnego.

procesy przerzutu

Wpływ defektów na drgania sieci

Kryształy, z którymi mamy do czynienia w praktyce, nigdy nie są doskonałe. Zawierają one zawsze pewną koncentrację defektów. Defekty te powodują istotne modyfikacje w obrazie drgań sieci kryształów doskonałych.

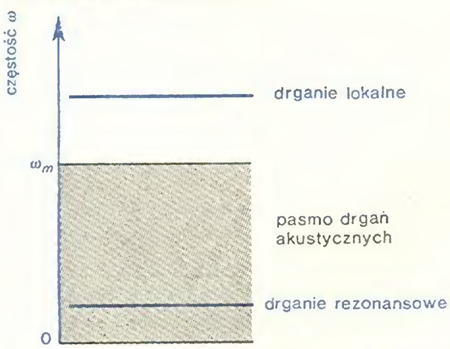
Wiele ogólnych wyników dotyczących drgań sieci z defektami można otrzymać rozpatrując stosunkowo proste modele defektów. Do takich należy m.in. defekt sieci wytworzony przez atom obcy (domieszkę chemiczną) umieszczony w węźle sieci zamiast atomu własnego kryształu. Defekt taki wprowadza lokalne zaburzenie, polegające na zmianie masy i stałych siłowych w stosunku do wartości w doskonałej sieci. Oddziaływania domieszki z atomami sieci zwykle mają krótki zasięg, co w praktyce pozwala rozpatrywać tylko stałe siłowe między atomem obcym i jego najbliższymi sąsiadami. Poza tym w najprostszych rozważaniach przyjmuje się, że defekty nie oddziałują między sobą. Jest to uzasadnione zwłaszcza wtedy, gdy koncentracja defektów jest mała.

domieszki

Własności drgań sieci w kryształach z defektami otrzymuje się przez rozwiązywanie odpowiednich równań ruchu. Ograniczymy się tutaj do przedstawienia wyników dla sieci jednowymiarowej zawierającej jeden atom w komórce elementarnej. Rozwiązanie równań ruchu dla takiej sieci z atomem obcym umieszczonym w węźle wykazuje, że widmo drgań sieci składa się z pasma częstości (podobnie jak w sieci doskonałej) i jednej dyskretnej częstości położonej po-

drganie lokalne

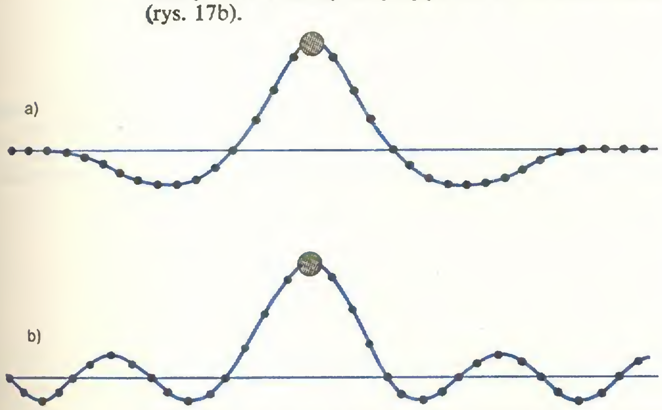
wyżej tego pasma (rys. 16). Dyskretna częstość odpowiada tzw. drganiu lokalnemu. Charakterystyczną cechą tego drgania jest wykładnicze malenie amplitudy ze wzrostem odległości od atomu obcego (rys. 17a). Innymi słowy, w drganiu lokalnym bierze udział tylko atom obcy i mała liczba atomów znajdujących się w jego sąsiedztwie.



Rys. 16. Położenie drgań lokalnych i rezonansowych w stosunku do pasma częstości drgań dozwolonych sieci doskonałej

Wpływ defektów na częstości drgań normalnych w obszarze pasmowym jest pomijalnie mały. Pasmo dozwolonych częstości drgań w krysztale zawierającym małą koncentrację defektów pokrywa się z analogicznym pasmem w krysztale doskonałym. Modyfikacji ulegają amplitudy drgań normalnych, zwłaszcza w pobliżu defektów. Czasem obserwuje się rezonansowy wzrost amplitudy dla określonej częstości drgań znajdującej się wewnątrz pasma. Drgania mające tę własność noszą nazwę drgań rezonansowych lub drgań pseudolokalnych. Biorą w nich udział tylko atomy obce i atomy znajdujące się w ich sąsiedztwie (rys. 17b).

drgania rezonansowe

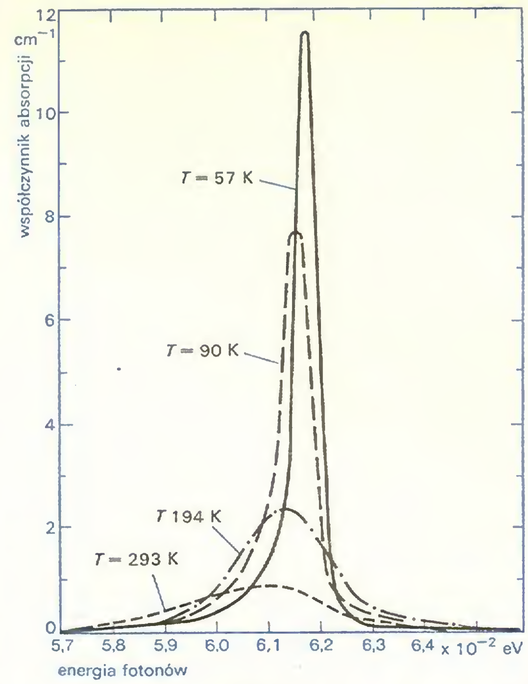


Rys. 17. Ruch atomów w sieci liniowej zawierającej atom obcy: a) drganie lokalne, b) drganie rezonansowe

Drgania rezonansowe występują wtedy, gdy przekazanie energii drgań od defektu do sieci jest z jakichś powodów utrudnione i zachodzi powoli. Przyczyną utrudniającą ten proces jest zwykle słabe sprzężenie atomu obcego z siecią (małe stałe siłowe) lub mała wartość $g(\omega)$ dla ω odpowiadającej częstości drgań atomu obcego.

Drgania lokalne pojawiają się wówczas, gdy atom obcy ma masę znacznie mniejszą niż atom macierzysty sieci. Do tego samego efektu prowadzi „wzmocnienie" stałych siłowych sprzęgających atom obcy z siecią. Natomiast, aby powstały drgania rezonansowe, muszą być wprowadzone do sieci atomy obce o masie większej niż masa atomów macierzystych sieci lub musi nastąpić znaczne „osłabienie" stałych siłowych.

Doświadczalne badanie drgań lokalnych, jak również rezonansowych przeprowadza się w kryształach jonowych i kowalencyjnych metodami optycznymi.



Rys. 18. Absorpcja podczerwieni przez drgania lokalne jonów H^- (jon H^- zastępuje w sieci KCl jon Cl^-) w krysztale KCl (chlorku potasu) w kilku temperaturach. Energia fotonu w maksimum krzywych odpowiada długości fali około 21 μm

W widmach absorpcji podczerwieni obserwuje się pasma takie, jak np. na rys. 18; położenie maksimum pasma pozwala określić częstość drgania lokalnego lub rezonansowego, a jego rozmycie — czas życia.

badania optyczne

Doświadczalne metody badania drgań sieci

Niesprężyste rozproszenie neutronów powolnych

Wśród doświadczalnych metod dostarczających informacji o drganiach sieci szczególnie ważną rolę odgrywa niesprężyste rozproszenie neutronów powolnych. Metodą tą można otrzymać krzywe dyspersji, a w wypadku kryształów o strukturze regularnej — również widmo częstości drgań.

W eksperymentach rozproszeniowych stosuje się neutrony o energii rzędu 0,1 eV. Neutrony o tej energii mają długość fali de Broglie'a porównywalną z odległościami międzyatomowymi w krysztale. Dla takich fal kryształ stanowi swego rodzaju przestrzenną siatkę dyfrakcyjną. Monochromatyczna fala neutronowa padająca na kryształ w wyniku oddziaływania z jądrami atomowymi ulega rozproszeniu. Gdy jądra atomowe w każdym węźle sieci mają identyczne własności rozpraszające, fale rozproszone nakładają się wzajemnie, dając zjawisko interferencji. Tego rodzaju rozproszenie nazywa się spójnym. Występuje ono wówczas, gdy kryształ jest doskonały (brak defektów), jednorodny pod względem izotopowym, a spiny jąder atomowych są równe zeru. Jeżeli natomiast kryształ zawiera przypadkowo rozmieszczone izotopy lub jego jądra atomowe mają spiny różne od zera, to mamy do czynienia z rozproszeniem niespójnym.

kryształ jako siatka dyfrakcyjna

Spójne rozproszenie neutronów w krysztale może może być zarówno sprężyste, jak i niesprężyste. W procesach sprężystych energia neutronu w czasie rozproszenia nie ulega zmianie. Efekty interferencyjne

odbicie braggowskie

prowadzą do tzw. odbicia braggowskiego od różnych płaszczyzn w krysztale, co opisuje wzór:

$$2d\sin\Theta = n\lambda,$$

gdzie d — odległość między płaszczyznami sieciowymi w krysztale, Θ — kąt odbicia braggowskiego, n — rząd odbicia, λ — długość fali de Broglie'a neutronu. Z wzoru tego można wyznaczyć długość fali λ, jeśli się zna kąt Θ z pomiaru. Posługując się związkami:

$$E = \hbar^2/2M_n\lambda^2 \quad (M_n \text{ — masa neutronu}),$$

$$|\vec{k}| = 2\pi/\lambda,$$

można następnie obliczyć energię i wartość wektora falowego neutronu.

Z punktu widzenia dynamiki sieci krystalicznej bardziej interesujące są procesy niesprężystego rozproszenia, w których rozproszonemu neutronowi towarzyszy pojawienie się lub zniknięcie jednego lub więcej fononów. Spójne, niesprężyste rozproszenie neutronów zachodzące z udziałem jednego fononu wykorzystuje się do otrzymania relacji dyspersji fononów. Jeżeli neutron ulega rozproszeniu i jednocześnie zachodzi absorpcja jednego fononu, to zasada zachowania energii i pędu wymaga spełnienia następujących zależności:

niesprężyste rozpraszanie z udziałem 1 fononu

$$E' - E = \frac{\hbar^2}{M_n}(k'^2 - k^2) = \hbar\omega_j(\vec{q}), \tag{14}$$

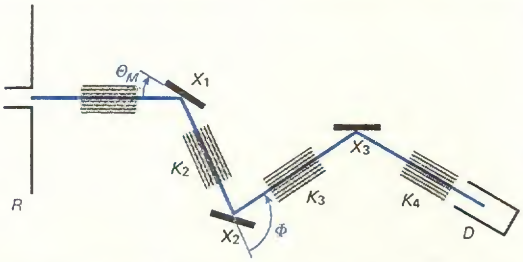
$$\vec{k}' - \vec{k} = \vec{q} + \vec{G}, \tag{15}$$

gdzie M_n — masa, E', E i $\vec{k}'$, $\vec{k}$ — odpowiednio energia i pęd neutronu padającego i rozproszonego. Gdy w doświadczeniu zmierzy się zmiany energii i pędu neutronu, to posługując się równaniami (14) i (15), można będzie wyznaczyć częstość i wektor falowy fononu uczestniczącego w procesie rozproszenia.

badanie krzywych dyspersji

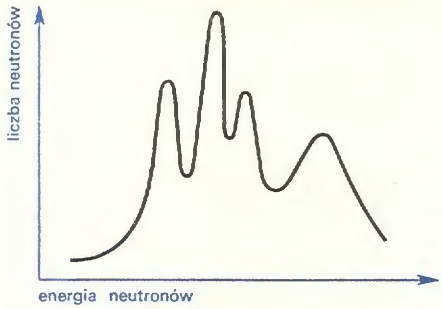
W badaniach niesprężystego rozproszenia neutronów źródłem neutronów są reaktory jądrowe. Prędkie neutrony wychodzące z reaktora przepuszcza się przez moderator (grafit, woda), w którym w wyniku szeregu kolejnych zderzeń z lekkimi atomami wytracają one energię lub inaczej mówiąc — ulegają spowolnieniu. Gdy temperatura moderatora jest bliska pokojowej, średnia energia neutronu osiąga wartość ok. 0,03 eV. Neutrony takie nazywamy termicznymi.

Rys. 19 przedstawia typowy spektrometr neutronowy przeznaczony do badania krzywych dyspersji neutronów. Wiązka termicznych neutronów jest formo-



Rys. 19. Schemat spektrometru neutronowego do badania krzywych dyspersji: R reaktor, K_1, K_2, K_3 i K_4 kolimatory, X_1 kryształ monochromatora, X_2 kryształ badany, X_3 kryształ analizatora, D detektor neutronów, Φ kąt rozproszenia

wana w kolimatorze K_1 (kanały w bloku substancji pochłaniającej neutrony). Monochromatyzację wiązki uzyskuje się przez braggowskie odbicie od kryształu X_1, który odgrywa rolę monochromatora. Energię E i wektor falowy $\vec{K}$ neutronu wyznacza się na podstawie pomiaru kąta Θ_M. Następnie monochromatyczna wiązka pada na badany kryształ X_2, który jest zorientowany w określony sposób w stosunku do wiązki neutronów. Energia neutronów rozproszonych pod kątem Φ jest analizowana przez kryształ X_3. Detektor D rejestruje natężenie rozproszonych neutronów w zależności od ich energii dla danego kąta rozproszenia Φ. W widmie rozproszenia obserwuje się maksima interferencyjne odpowiadające rozproszeniu z udziałem różnych fononów (rys. 20). Każde z tych maksimów pozwala wyznaczyć jedną wartość



Rys. 20. Typowe widmo neutronów rozproszonych na fononach

funkcji $\omega_j(\vec{q})$. Przeprowadzając badanie dla odpowiednio dobranych różnych kątów rozproszenia Φ i różnych orientacji kryształu X_2, można wykreślić całą funkcję $\omega_j(\vec{q})$. Specjalne metody, których nie będziemy tutaj omawiać, pozwalają określić wektor polaryzacji $\vec{e}_s(\vec{q}j)$ drgań i przyporządkować je odpowiedniej gałęzi krzywej dyspersji. Rozmycie maksimów interferencyjnych przy dobrej zdolności rozdzielczej spektrometru neutronowego charakteryzuje czas życia fononów — związany głównie z procesami anharmonicznymi w krysztale.

W niektórych kryształach dominującą rolę odgrywają procesy niespójnego rozproszenia niesprężystego. Nie można wówczas otrzymać krzywych dyspersji. W kryształach o strukturze regularnej zależność natężenia rozproszonych neutronów od energii (przy stałym kącie rozproszenia) odzwierciedla kształt widma częstości drgań $g(\omega)$.

zalety stosowania neutronów termicznych

Niewątpliwą zaletą neutronów termicznych w badaniach dynamiki sieci jest to, że ich energia i wartość wektora falowego są porównywalne z energią i wartością wektora falowego fononów w całym przedziale strefy Brillouina. Dzięki temu w procesie oddziaływania zachodzą znaczne zmiany tych wielkości, łatwe do zmierzenia. Duże znaczenie ma także fakt, że w rozproszeniu neutronów mogą brać udział wszystkie fonony, jeżeli tylko spełniona jest zasada zachowania energii i pędu. Pod tym względem zachodzi istotna różnica w stosunku do oddziaływania promieniowania elektromagnetycznego z fononami, w którym występują dodatkowe ograniczenia, nie wynikające z zasady zachowania energii i pędu.

Absorpcja w podczerwieni

Jednym z efektów oddziaływania promieniowania elektromagnetycznego z drganiami sieci jest jego absorpcja w krysztale, która występuje w szerokim zakresie spektralnym, obejmującym tzw. średnią i daleką podczerwień (od kilku do 1000 μm). Analiza widma absorpcji dostarcza wielu cennych danych o drganiach sieci krystalicznej.

Absorpcja promieniowania na drganiach sieci krystalicznej może być badana tylko w kryształach zawierających niezbyt dużą koncentrację swobodnych nośników ładunku. W kryształach metali i silnie domieszkowanych półprzewodników absorpcja na swobodnych nośnikach nie pozwala obserwować znacznie słabszych pasm absorpcji związanych z drganiami sieci.

warunek absorpcji

Warunkiem wystąpienia absorpcji promieniowania elektromagnetycznego na drganiach sieci jest istnienie oscylującego dipolowego momentu elektrycznego.

W kryształach jonowych moment ten powstaje w wyniku przeciwfazowego ruchu jonów przeciwnego znaku, odpowiadającego drganiom optycznym. W komórce elementarnej zawierającej dwa jony moment elektryczny $\vec{M}$ wyraża się wzorem:

$$\vec{M} = e(\vec{u}_+ - \vec{u}_-), \tag{16}$$

gdzie e jest ładunkiem jonów, $\vec{u}_+$ i $\vec{u}_-$ — odchyleniami dodatniego i ujemnego jonu od położenia równowagi. W bardziej ścisłych rozważaniach należy uwzględnić wkład do dipolowego momentu elektrycznego pochodzący od deformacji powłok elektronowych atomów, która zachodzi w czasie drgań sieci.

absorpcja — proces kwantowy

Absorpcja promieniowania jest procesem kwantowym, w którym mogą brać udział jeden, dwa lub więcej fononów. W procesach tych obowiązuje zasada zachowania energii i kwazipędu. W oddziaływaniu jednofononowym zasady te wymagają spełnienia następujących warunków:

$$\omega_{\text{foton}} = \omega_{\text{fonon}}, \tag{17}$$

$$\vec{k} = \vec{q}. \tag{18}$$

Z wzorów (17 i (18) wynika, że absorpcja promieniowania elektromagnetycznego z towarzyszącą jej emisją jednego fononu zachodzi wtedy, gdy częstość fotonu jest równa częstości fononu, a wektory falowe fotonu i fononu są zgodnie skierowane i równe co do bezwzględnej wartości.

Częstość fononów optycznych jest zwykle rzędu 10^{13} s^{-1}. Częstości tej odpowiada wartość wektora falowego $2{,}1 \cdot 10^3$ cm^{-1}, która jest o kilka rzędów mniejsza od rozmiarów strefy Brillouina. Można stąd wyciągnąć wniosek, że tylko fonony znajdujące się w pobliżu środka strefy Brillouina mogą oddziaływać z promieniowaniem elektromagnetycznym. Ze względu na bardzo małą wartość wektora falowego tych fononów w skali strefy Brillouina mówimy, że mają one prawie zerową wartość $\vec{q}$.

absorpcja jednofononowa

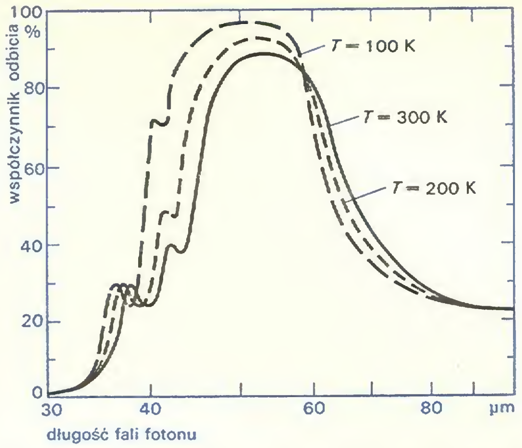
Oddziaływania optyczne podlegają określonym regułom wyboru (→ Spektroskopia atomowa). W wypadku absorpcji jednofononowej reguły te można sformułować żądając, aby dipolowy moment elektryczny wytwarzany przez drgania sieci był różny od zera i miał składową równoległą do wektora elektrycznego $\vec{E}$ promieniowania elektromagnetycznego, czyli:

$$\vec{M}\vec{E} \neq 0. \tag{19}$$

Jeżeli przyjmiemy, że moment elektryczny $\vec{M}$ jest równoległy do wektora polaryzacji $\vec{e}$ drgań sieci, to proste rozważania doprowadzą do wniosku, że tylko fonony optyczne poprzeczne (TO) oddziałują z promieniowaniem elektromagnetycznym. Fonony optyczne podłużne (LO) nie dają absorpcji promieniowania, chociaż wytwarzają w krysztale dipolowy moment elektryczny.

W kryształach jonowych procesy jednofononowe charakteryzuje bardzo duży współczynnik absorpcji (rzędu 10^5–10^6 cm^{-1}). Do badania tej absorpcji konieczne są bardzo cienkie próbki, które z różnych względów trudno jest otrzymać. W związku z tym znacznie wygodniej jest badać współczynnik odbicia. Stosując odpowiednie wzory matematyczne (tzw. relacje Kramersa–Kroniga), z widma odbicia można otrzymać stałe optyczne, a w szczególności współczynnik absorpcji.

Pasmo odbicia związane z oddziaływaniem fotonu z fononem optycznym nosi nazwę *reststrahlen band* (pasmo promieni resztkowych). Występuje ono w podczerwieni przy długościach fali rzędu kilkudziesięciu µm. Rysunek 21 przedstawia omawiane pasmo odbicia dla kryształu NaCl. Analiza pasma odbicia pozwala otrzymać częstość fononu optycznego poprzecznego (TO) o wektorze falowym $\vec{q} \approx 0$, jego czas życia oraz siłę oscylatora, która określa wielkość



Rys. 21. Zależność współczynnika odbicia od długości fali dla kryształu NaCl w różnych temperaturach

sprzężenia promieniowania elektromagnetycznego z danym fononem.

W kryształach kowalencyjnych o strukturze diamentu (krzem, german) atomy są neutralne i elektryczny moment dipolowy może pochodzić tylko od deformacji powłok elektronowych atomów wywoływanej przez drgania sieci. Na dwóch identycznych atomach w komórce elementarnej drgania sieci wytwarzają dipolowe momenty elektryczne równe co do wielkości, lecz przeciwnie skierowane. W rezultacie całkowity moment elektryczny w komórce elementarnej, a tym samym w całym krysztale znika. A zatem w kryształach tych nie zachodzą procesy jednofononowe. W widmie odbicia nie pojawia się pasmo odbicia resztkowego, charakterystyczne dla kryształów jonowych lub kryształów mających wiązania częściowo jonowe.

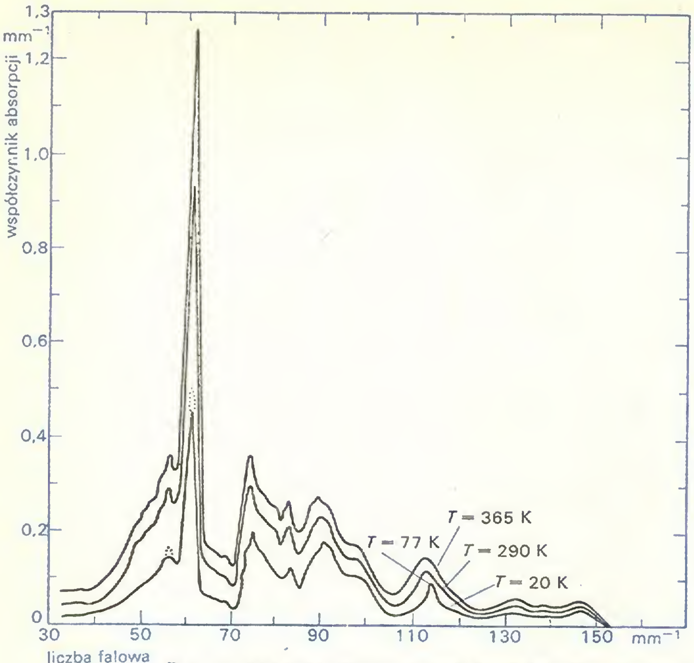
absorpcja dwufononowa

Do absorpcji w kryształach kowalencyjnych o strukturze diamentu przyczyniają się głównie procesy dwufononowe. Oddziaływania dwufononowe w tych kryształach związane są z dipolowym momentem elektrycznym drugiego rzędu, który powstaje, gdy deformacja powłok elektronowych atomów jest rezultatem współdziałania dwóch fononów. Procesy dwufononowe występują, oczywiście, nie tylko w kryształach kowalencyjnych. W kryształach jonowych i częściowo jonowych są jednak możliwe inne mechanizmy oddziaływania niż te, które przed chwilą przedstawiliśmy. Absorpcja związana z procesami dwufononowymi jest na ogół znacznie słabsza niż absorpcja odpowiadająca procesom jednofononowym.

W procesach dwufononowych w wyniku absorpcji fotonu powstają dwa fonony (proces sumacyjny) lub jeden fonon powstaje, a drugi znika (proces różnicowy). Oddziałujące fonony mają równe co do wielkości, ale przeciwnie skierowane wektory falowe. W procesach tych mogą brać udział fonony z całej strefy Brillouina, jeżeli reguły wyboru związane z symetrią komórki elementarnej nie zabraniają oddziaływań. Rysunek 22 przedstawia widmo dwufononowej absorpcji krzemu. Składa się ono z szeregu sumacyjnych pasm absorpcji, które można przyporządkować określonym kombinacjom dwóch fononów. Analiza dwufononowego widma absorpcji pozwala wyznaczyć częstość fononów w tych miejscach, w których funkcja rozkładu częstości drgań ma maksima.

analiza widma absorpcji

Przy dużej zdolności rozdzielczej aparatury pomiarowej w widmie absorpcji można zauważyć szczegóły związane z punktami krytycznymi omawianymi wyżej. Analiza widma absorpcji pozwala w takich razach wyznaczyć częstości drgań sieci w ściśle określonych punktach strefy Brillouina. Do interpretacji widma absorpcji konieczna jest znajomość reguł wyboru, które określają dozwolone kombinacje fononów w danym punkcie krytycznym.



Rys. 22. Absorpcja dwufononowa w krzemie w kilku temperaturach

absorpcja jednofononowa na drganiach defektów

Jak już wcześniej zaznaczono, badanie procesów jednofononowych daje tylko informacje o fononach ze środka strefy Brillouina. Ograniczenie to jest związane z zasadą zachowania kwazipędu w oddziaływaniach foton–fonon. Okazuje się jednak, że można „aktywować" zabronione oddziaływania jednofononowe przez wprowadzenie do sieci krystalicznej defektów (np. domieszek chemicznych). Drgania specjalnie dobranych defektów zaledwie nieznacznie różnią się od drgań atomów własnych kryształu. W czasie drgań defektów powstaje oscylujący dipolowy moment elektryczny, który oddziałuje z promieniowaniem elektromagnetycznym. Jednofononowa absorpcja promieniowania elektromagnetycznego na drganiach defektów odzwierciedla widmo częstości drgań w prawie doskonałym krysztale.

Rozproszenie ramanowskie

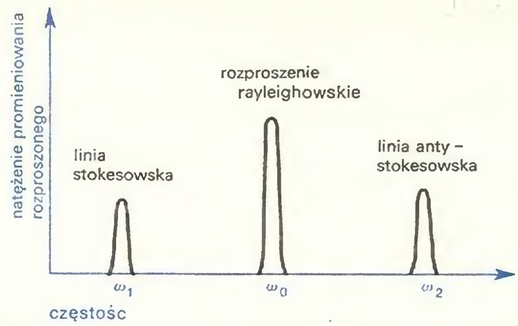
Dotychczas rozpatrywaliśmy procesy, w których zachodzi absorpcja fotonu i powstanie (emisja) fononów. Istnieje jeszcze inny rodzaj oddziaływań między promieniowaniem elektromagnetycznym i kryształem. Oddziaływanie to prowadzi do rozproszenia fotonu z jednoczesnym powstaniem (lub zniknięciem) fononów. Nosi ono nazwę rozproszenia ramanowskiego.

analiza promieniowania rozproszonego

W doświadczalnych badaniach rozproszenia ramanowskiego na kryształ pada monochromatyczna wiązka światła, która ulega rozproszeniu we wszystkich kierunkach. Analizuje się skład widmowy promieniowania rozproszonego. W widmie tym znajdujemy nie tylko częstość ω_0 promieniowania padającego, ale także inne częstości, mniejsze i większe niż ω_0. Analiza widma promieniowania rozproszonego dostarcza informacji o fononach, które biorą udział w procesie rozproszenia.

W rozproszeniu ramanowskim pierwszego rzędu powstaje lub znika fonon optyczny ze środka strefy Brillouina ($\vec{q} \approx 0$). Rysunek 23 przedstawia schematycznie widmo ramanowskie odpowiadające tego rodzaju procesom. Linia centralna o częstości ω_0 odpowiada rozproszeniu Rayleigha, które zachodzi bez udziału fononów. Linia o częstości ω_1 związana jest z procesem emisji jednego fononu i nosi nazwę linii stokesowskiej — światło rozproszone ma mniejszą częstość niż światło padające. Linia o częstości ω_2, większej niż częstość światła padającego, powstaje w wyniku absorpcji fononu i nazywa się linią antystokesowską.

linia stokesowska i antystokesowska

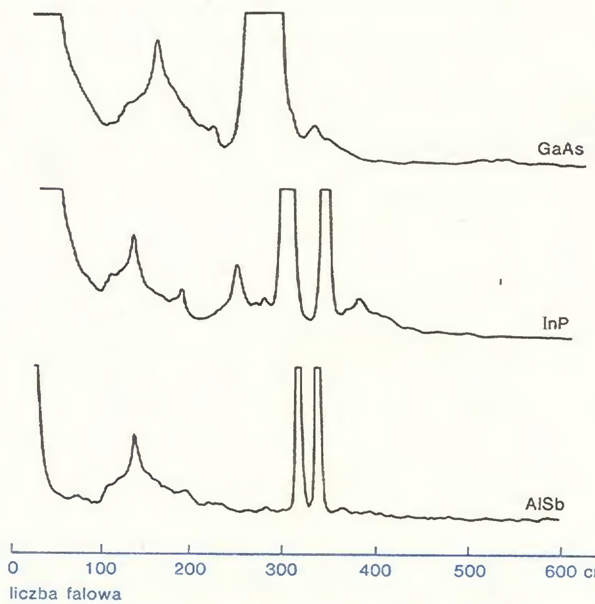


Rys. 23. Schematyczne widmo ramanowskie pierwszego rzędu odpowiadające oddziaływaniu z jednym fononem

Różnica $(\omega_0-\omega_1)$ lub $(\omega_2-\omega_0)$ określa częstość fononu biorącego udział w procesie rozproszenia światła. Gdy w rozproszeniu jest czynna większa liczba fononów, widmo ramanowskie składa się z odpowiednio większej liczby linii. Rozmycie linii jest zwykle związane ze skończonym czasem życia fononów.

W rozproszeniu ramanowskim obserwuje się także procesy drugiego rzędu, w których biorą udział dwa fonony. Analiza tych procesów dostarcza danych o fononach analogicznych do danych z analizy dwufononowych procesów absorpcyjnych.

Rysunek 24 przedstawia widmo ramanowskie trzech kryształów związków półprzewodnikowych grupy $A^{III}B^{V}$. W widmie tym widoczne są dwa silne maksima związane z fononami LO i TO oraz słabsze maksima odpowiadające procesom drugiego rzędu.



Rys. 24. Widmo rozpraszania ramanowskiego w kryształach GaAS, InP i AlSb w temperaturze pokojowej (część antystokesowska)

prawdopodobieństwo rozproszenia Ramana

Prawdopodobieństwo rozproszenia ramanowskiego zależy od wielkości zmian polaryzowalności elektronowej kryształu wywołanej przez drgania sieci. Natężenie promieniowania rozproszonego w procesach jednofononowych jest zwykle 10^6–10^7 razy mniejsze od natężenia promieniowania padającego.

Rozproszenie światła, przy którym powstaje lub znika fonon akustyczny, nazywa się rozproszeniem Brillouina. Rozproszenie to charakteryzuje bardzo

rozproszenie Brillouina mała zmiana częstości (rzędu kilku cm^{-1}), która zależy od częstości światła padającego. Dla procesów rozproszenia ramanowskiego zmiana częstości jest mniej więcej o dwa rzędy wielkości większa i ma wartość stałą, niezależną od częstości światła padającego. Badanie rozproszenia Brillouina jest jedną z metod wyznaczenia współczynników sprężystości kryształów.

reguły wyboru Rozproszenie ramanowskie podlega określonym regułom wyboru. Reguły te nie pokrywają się z regułami dla procesów absorpcyjnych. Dlatego w widmach ramanowskich można często obserwować fonony, które nie pojawiają się w widmach absorpcyjnych, np. w kryształach o strukturze diamentu. W rozproszeniu ramanowskim pierwszego rzędu biorą udział fonony optyczne ze środka strefy Brillouina, które nie mogą uczestniczyć w absorpcji jednofononowej.

W badaniach rozproszenia Ramana i Brillouina stosuje się obecnie lasery jako źródła promieniowania pobudzającego. Odznaczają się one dużą mocą i wysoką monochromatycznością. Dzięki tym cechom stało się możliwe wykonanie bardzo precyzyjnych badań zarówno procesów pierwszego, jak i drugiego rzędu. Eksperymenty z określoną geometrią i polaryzacją wiązek światła padającego i rozproszonego dostarczają dokładnych informacji nie tylko o częstościach fononów, ale również o ich przynależności do określonej gałęzi dyspersyjnej.

Aktualne kierunki badań

Dynamika sieci krystalicznej jest jednym z najstarszych działów fizyki ciała stałego. Jej podstawy teoretyczne zostały sformułowane przez P. Debye'a, M. Borna i T. Kármána już w 1912 r. Przez długi jednak okres nie było odpowiednich metod doświadczalnych pozwalających badać drgania sieci. Wyniki teorii były kontrolowane tylko pośrednio, poprzez obliczanie wielkości termodynamicznych i porównywanie tych obliczeń z wynikami pomiarów. Do wymienionych celów zwykle wykorzystywano zależność ciepła właściwego od temperatury.

Właściwe metody umożliwiające badanie własnoścj poszczególnych drgań normalnych sieci krystalicznej pojawiły się dopiero w okresie ostatnich 20 lat. W 1955 r. wykonano pierwsze badania niesprężystego rozproszenia neutronów powolnych na drganiach sieci. Mniej więcej od początku lat sześćdziesiątych zaczęły się pojawiać coraz liczniej prace stosujące w badaniach drgań sieci technikę spektroskopii w podczerwieni, a w kilka lat później — laserowej spektroskopii ramanowskiej. Intensywny rozwój badań eksperymentalnych stał się czynnikiem stymulującym prace teoretyczne.

Wśród aktualnie prowadzonych badań dynamiki sieci krystalicznej można wyróżnić kilka kierunków. Jednym z nich jest niewątpliwie badanie dynamiki sieci kryształów doskonałych, nie zawierających defektów. Celem tych badań jest otrzymanie krzywych dyspersji i widma częstości drgań dla rzeczywistych kryształów. Tematyka ta wiąże się ściśle z zagadnieniem oddziaływań międzyatomowych w kryształach. **dynamika sieci bez defektów**

Początkowo badania dynamiki sieci były z reguły ograniczone do kryształów zawierających niewiele (jeden lub dwa) atomów w komórce elementarnej i mających jednocześnie stosunkowo prostą strukturę krystaliczną o wysokiej symetrii. Obecnie coraz częściej spotyka się prace dotyczące kryształów o złożonej budowie. Specjalne miejsce w tych badaniach zajmują kryształy molekularne, w których oprócz zwykłych drgań, nazywanych czasem drganiami translacyjnymi, występują drgania torsyjne (skrętne) oraz drgania atomów wewnątrz molekuł.

Oprócz dalszego zainteresowania dynamiką sieci kryształów doskonałych, od początku lat sześćdziesiątych zaczął się silny rozwój nowego kierunku badań, który można określić jako dynamikę sieci kryształów zawierających defekty i niedoskonałości. Znaczna grupa prac w tej dziedzinie koncentruje się wokół zagadnień wpływu na widmo drgań sieci małej koncentracji defektów punktowych. Wiele uwagi poświęca się także badaniu dynamiki nieuporządkowanych kryształów mieszanych, chociaż napotyka się tutaj znacznie większe trudności teoretyczne niż przy defektach punktowych. Interesujące zagadnienia, których badania znajdują się dopiero w początkowym stadium, są związane z dynamiką substancji amorficznych (bezpostaciowych). **dynamika sieci z defektami**

Dokonując przeglądu aktualnych kierunków badań nie sposób pominąć obszernej dziedziny, jaką stanowi badanie oddziaływań fononów z innymi kwazicząstkami w krysztale, takimi jak elektrony, plazmony, magnony i in. Poznanie tych oddziaływań jest niezbędne do pełniejszego zrozumienia szeregu własności fizycznych kryształów.

W. Cochran *The Dynamics of Atoms in Crystals*, New York 1974; B. Donovan, J. F. Angress *Lattice Vibrations*, London 1971; J. A. Reissland *The Physics of Phonons*, London 1973.

Wzbudzenia elementarne w ciałach stałych

Jerzy Czerwonko

Ogólne pojęcie wzbudzenia

Wyobraźmy sobie dowolny układ fizyczny o skwantowanych poziomach energetycznych E_k ($k = 0$, 1, 2 ...); załóżmy np., że E_k rośnie wraz z k. Układ może znajdować się w stanie podstawowym, tj. w stanie o energii najniższej E_0, lub też w którymś ze stanów wzbudzonych, o energii E_k ($k>0$). Energię E_k-E_0 ($k>0$) nazywamy energią wzbudzenia. Jeśli układ jest izolowany i znajduje się w którymś ze stanów o energii E_k ($k = 1, 2, 3, ...$), to pozostanie w tym stanie nieskończenie długo, ze względu na prawo zachowania energii; stany wzbudzone układu będą więc stanami stabilnymi. Żaden z układów fizycznych nie jest jednak izolowany dokładnie, chyba że myślimy ewentualnie o tak skomplikowanym układzie fizycznym, jakim jest Wszechświat. (Owo „ewentualnie" ma dodatkowy sens, bowiem dla współczesnej kosmologii kwestia izolacji Wszechświata nie jest bynajmniej sprawą trywialną). Jeśli rozpatrywany układ nie jest izolowany, a oddziaływanie układu z otoczeniem jest odstatecznie słabe, to stany wzbudzone nie są stabilne — stabilny jest tylko stan podstawowy. Ten ostatni fakt można łatwo zrozumieć, jeśli się uwzględni, że prawo zachowania energii obowiązuje obecnie w układzie rozszerzonym: układ + otoczenie. Wobec tego, nawet przy dowolnie słabym oddziaływaniu z otoczeniem, układ w stanie wzbudzonym może oddać energię otoczeniu i przejść do stanu podstawowego. Z drugiej strony, przejście ze stanu podstawowego do wzbudzonego jest możliwe tylko wtedy, gdy otoczenie może dostarczyć układowi w jednej chwili energię E_1-E_0 lub większą, co jest osiągalne jedynie przy dostatecznie silnym oddziaływa-

energia wzbudzenia

stabilność stanów wzbudzonych

niu. Jak widzimy, istnieje wyraźna analogia między stanem podstawowym a stanem równowagi trwałej i stanem wzbudzonym a stanem równowagi chwiejnej.

Zgodnie z podstawowymi zasadami mechaniki kwantowej przejście układu ze stanu k do jakiegokolwiek innego stanu nie odbywa się „na komendę", lecz w sposób statystyczny. Przejście to można scharakteryzować średnim czasem przebywania w danym stanie, zwanym inaczej średnim czasem życia danego stanu lub też czasem relaksacji τ_k. Przejściu ze stanu $k>0$ do stanu $k = 0$ odpowiada zmiana energii układu o $E_k - E_0$ (której, zgodnie z mechaniką kwantową, odpowiada częstość $\nu_k = (E_k - E_0)/h$, gdzie h jest stałą Plancka) oraz emisja fali o tej częstości (może to być np. fala elektromagnetyczna). Zwróćmy uwagę, że informacje o różnicy poziomów energii uzyskuje się wyłącznie na podstawie pomiaru częstości fal, jeśli włączyć tu również fale opisujące cząstki. Stąd, jeśli $\nu_k \gg 1/\tau_k$, to jest możliwe określenie częstości fali w czasie życia poziomu (oscylacja w czasie życia poziomu zachodzi wielokrotnie), jeśli zaś częstość ν_k jest porównywalna z $1/\tau_k$ lub mniejsza, to częstości precyzyjnie określić się nie da. Dlatego też nie powinno zaskakiwać nazwanie takich wzbudzeń, dla których $\nu_k\tau_k \gg 1$, tj. $(E_k - E_0)\tau_k \gg h$, wzbudzeniami dobrze określonymi. Zwróćmy uwagę na związek powyższych rozważań z zasadą nieokreśloności dla energii, którą zapisać można w postaci $\Delta E\ \Delta t \approx h/2\pi$.

wzbudzenia dobrze określone

analogia z czystością dźwięku

Aby łatwiej uzmysłowić sobie sens warunku dobrej określoności wzbudzeń, odwołajmy się do analogii akustycznej, podanej przez twórcę cybernetyki N. Wienera. Jak wiadomo wysokość dźwięku jest związana z najniższą częstością w nim występującą czyli częstością podstawową. Dźwięk będzie czysty, jeśli oprócz tej częstości wystąpią w nim wyłącznie częstości harmoniczne będące wielokrotnościami częstości podstawowej. Aby można było wyczuć periodyczność związaną z częstością podstawową, tj. aby była ona „dobrze określona", potrzeba, aby co najmniej $\tau\nu > 1$, przy czym τ to tym razem czas trwania dźwięku, ν zaś — jego częstość podstawowa. Stąd też, jeśli włączyć dźwięk na czas krótszy niż okres drgań dźwięku, to odbierzemy wrażenie dźwięku nieczystego. Dlatego właśnie jest niemożliwe czyste zagranie polki czy taranteli przy wykorzystaniu dźwięków najniższych, np. przy grze na organach — włączanych pedałami.

Układy wielu cząstek; cząstki Fermiego

Wyobraźmy sobie obecnie ciąg poziomów energii E_k ($k = 0, 1, 2, \ldots$) takich, że $E_{k+l} > E_k$ przy $l>0$. Poziomy te są ponadto niezwyrodniałe, tzn. każdemu poziomowi E_k odpowiada dokładnie jeden stan. Na poziomach tych mogą znajdować się cząstki podlegające statystyce Fermiego, oddziałujące ze sobą na tyle słabo, że energia E układu może być z dobrą dokładnością zapisana jako

$$E_0N_0 + E_1N_1 + E_2N_2 + \ldots + E_kN_k + \ldots,$$

gdzie $N_0, N_1, N_2, \ldots, N_k, \ldots$, to liczby przybierające wartości zero lub jeden, a określające ile cząstek znajduje się na poziomach $E_0, E_1, E_2, \ldots, E_k, \ldots$ (Ze względu na związek spinu ze statystyką, wynikający z relatywistycznej kwantowej teorii pola, cząstki podlegające statystyce Fermiego mają spin połówkowy, co oznacza, że w zerowym zewnętrznym polu magnetycznym zwyrodnienie każdego poziomu jest parzyste; podkreślmy jednak, że rozważamy sytuację modelową — brak zwyrodnienia). Jeśli liczba cząstek Fermiego jest ustalona, tzn. gdy $N_0 + N_1 + N_2 + \ldots + N_k = N = \text{const}$, to układ ma najniższą energię przy $N_k = 1$ dla $k = 0, 1, 2, \ldots, N-1$ i $N_k = 0$ dla $k \geqslant N$. Tak więc stany z $k \leqslant N-1$ są obsadzone, a stany z $k \geqslant N$ są puste. Można również łatwo określić wzbudzenia podanego układu. Jeśli cząstkę ze stanu $k \geqslant N-1$ przenieść do stanu $l \geqslant N$, to energia powstałego w ten sposób stanu będzie wyższa od energii stanu podstawowego $E_G = E_0 + E_1 + E_2 + \ldots + E_{N-1}$ o $E_l - E_k$, przy czym $E_l - E_k$ musi być dodatnie przy $l > k$. Powstały stan można opisać następująco: w stanie k powstała antycząstka, w stanie l zaś — cząstka.

antycząstki a dziury

Termin „antycząstka" użyty tu zgodnie z A. A. Abrikosowem, nie jest bynajmniej powszechny w literaturze naukowej; częściej używa się wówczas określenia „dziura". Jest ono jednak mylące ponieważ, jak widać z powyższego, ujemność masy efektywnej, istotna przy określeniu dziury w teorii pasmowej (→ Dynamika elektronu w ciałach stałych), nie była tu w ogóle brana pod uwagę. Można powiedzieć, że jeśli antycząstka znajduje się tak blisko maksimum pasma walencyjnego, że jej masa efektywna jest ujemna, to jest dziurą. Pojęcie antycząstki jest więc pojęciem ogólniejszym od pojęcia dziury, z tym że antycząstki w półprzewodnikach, przy dostatecznie niskich temperaturach, są dziurami.

Używając pojęcia cząstki i antycząstki można opisać wszystkie wzbudzenia rozpatrywanego układu o energii

$$E_{l_1} + E_{l_2} + \ldots + E_{l_s} - E_{k_1} - E_{k_2} - \ldots - E_{k_s}$$

przy $l_1, l_2, \ldots, l_s \geqslant N$, i $k_1, k_2, \ldots, k_s \leqslant N-1$ z cząstkami występującymi w stanach $l_1, l_2, \ldots, l$ i antycząstkami w stanach $k_1, k_2, k_3, \ldots, k_s$. Również wszystkie wzbudzenia układu nie oddziałujących ze sobą cząstek Fermiego można opisać nie zakładając, że stany dla $k = 0, 1, 2, \ldots$ są niezwyrodniałe, tj. podając zwyrodnienie stanów, określane liczbami naturalnymi g_k wyrażającymi, ile mamy stanów kwantowych o energii E_k.

związek ze strukturą elektronową

Związek między rozpatrywanym modelem a strukturą elektronową ciał stałych jest dość oczywisty. Stanom zajętym w przedstawianym tu modelu, gdy energia całkowita ma najmniejszą wartość, odpowiadają stany spod powierzchni Fermiego w metalach i stany pasma walencyjnego w półprzewodnikach; stanom pustym — stany znad powierzchni Fermiego i stany pasma przewodnictwa.

Oddziałujące cząstki Fermiego; wzbudzenia elementarne; kwazicząstki

Dotychczas zakładaliśmy, że cząstki z sobą nie oddziałują, co może być jedynie przybliżeniem — gorszym lub lepszym. Oddziaływanie cząstek prowadzi do wymiany energii między nimi. Wydzielmy myślowo jedną z cząstek układu; pozostałe potraktujmy jako jej otoczenie. Rozważania, podobne do przeprowadzonych poprzednio, wykazują, że jeśli oddziaływanie międzycząstkowe (a więc i oddziaływanie wydzielonej cząstki z otoczeniem) jest dostatecznie słabe, to spowoduje niewielkie przesunięcia poziomów energetycznych wydzielonej cząstki, a także ich niestabilność. Poziomom energetycznym cząstki można więc przypisać średni czas życia. Analogicznie, przy dostatecznie słabych oddziaływaniach poziomy jednocząstkowe, a więc wzbudzenia układu, będą dobrze określone.

średni czas życia poziomów energetycznych

Z drugiej strony, trudno jednak oprzeć się wrażeniu, że kryterium słabości oddziaływania cząstek między sobą rzadko jest spełnione w ciałach stałych. Aby to uzasadnić, rozpatrzmy np. elektrony przewodnictwa w metalu. Wiadomo z teorii pasmowej, że oddziaływanie z jonami zmienia w sposób zasadniczy własności elektronów przewodnictwa, np. ich masy efektywne. Oddziaływanie to ma charakter głównie elektrostatyczny, czyli jest określone przez średnią odległość pomiędzy elektronem i jonem. Elektrony oddziałują ze sobą również elektrostatycznie, a że jest ich mniej więcej tyle samo co i jonów, więc średnia odległość między elektronami będzie zbliżona do średniej odległości między elektronem i jonem. Dzięki temu, oddziaływanie elektronów między sobą powinno być wielkością tego samego rzędu co oddziały-

wanie elektronów z jonami, które modyfikuje w sposób zasadniczy własności układów elektronów. Oddziaływanie elektron–elektron powinno zatem istotnie modyfikować własności układów elektronów.

Powstaje pytanie, czy w warunkach występujących w ciałach stałych można w ogóle mówić o dobrze określonych wzbudzeniach. Okazuje się, że można, i że na przykład cząstka i antycząstka w metalach leżące dostatecznie blisko powierzchni Fermiego, tj. o energiach bliskich energii Fermiego E_F, tak że energia cząstki lub antycząstki E spełnia nierówność $|E-E_F| \ll E_F$, są typowymi przykładami wzbudzeń dobrze określonych. Podobnie przedstawia się z reguły sprawa z elektronami i dziurami w półprzewodnikach, jeśli mamy do czynienia ze wzbudzeniami położonymi odpowiednio blisko ekstremów pasm. Dlaczego tak jest? Wyjaśnijmy ten fakt na przykładzie elektronów w metalu. Jak wiadomo, pod powierzchnią Fermiego znajdują się stany zajęte przez elektrony, ponad nią — stany puste. Elektron można opisać podając liczby kwantowe kwazipędu $\vec{p}$, energii oraz pasma, do którego należy elektron. Oddziaływanie międzyelektronowe prowadzi do rozpraszania elektronów na elektronach, tj. do przekazywania sobie nawzajem przez elektrony energii i kwazipędu. W każdym akcie rozpraszania musi być zachowana całkowita energia oraz całkowity kwazipęd elektronów (dokładniej — kwazipęd musi być zachowany z dokładnością do wektora sieci odwrotnej pomnożonego przez stałą Plancka). Ponadto elektrony — ze względu na zasadę Pauliego — muszą trafić po rozproszeniu do stanów nie zajętych przez elektrony.

przykłady wzbudzeń dobrze określonych

statystyczny opis rozpraszania

Ów akt rozpraszania można, zgodnie z podstawami mechaniki kwantowej, ująć również w sposób statystyczny. Mianowicie, prawdopodobieństwo rozproszenia określonego stanu w jednostce czasu, odwrotnie proporcjonalne do średniego czasu życia danego stanu, będzie tym większe, im więcej jest stanów końcowych, w które może przejść dany stan. Aby uściślić te rozważania załóżmy, że na kwazipęd narzucono warunki Borna–Kármána (→ Struktura elektronowa ciał stałych), tj. że

$$p_x = hn_x/L,\ p_y = hn_y/L,\ p_z = hn_z/L,$$

przy czym h jest stałą Plancka, L — długością krawędzi sześciennej kostki kryształu, p_x, p_y, p_z — składowymi pędu elektronu odpowiednio w kierunku x, y, z, zaś n_x, n_y, n_z — liczbami całkowitymi leżącymi w przedziale $-N/2+1$, $N/2$, gdzie N^3 jest liczbą komórek w krysztale. Załóżmy, że zależność energii od pędu jest najprostsza, tj.

$$E(p_x, p_y, p_z) = (p_x^2+p_y^2+p_z^2)/2m^*,$$

przy czym m^* jest masą efektywną. Niech stan podstawowy układu będzie scharakteryzowany przez obsadzenie wszystkich stanów z

$$p_x^2+p_y^2+p_z^2 \leqslant p_F^2,$$

czyli

$$n_x^2+n_y^2+n_z^2 \leqslant L^2p_F^2/h^2$$

i nieobsadzenie wszystkich pozostałych stanów; załóżmy dodatkowo, że $L^2p_F^2/h^2 < N^2/4$ po to, aby wszystkie stany obsadzone mieściły się w strefie Brillouina. Cząstki jako wzbudzenia będą występowały przy $n_x^2+n_y^2+n_z^2 > L^2p_F^2/h^2$, antycząstki zaś — przy nierówności odwrotnej. Niech w układzie znajduje się cząstka z liczbami n_x, n_y, n_z oraz antycząstka z n_x', n_y', n_z'. Jeśli w wyniku rozproszenia mamy otrzymać również cząstkę z m_x, m_y, m_z oraz antycząstkę z m_x', m_y', m_z' to jako rezultat praw zachowania pędu i energii mamy: $n_x-n_x' = m_x-m_x'$ itd. dla składowych y i z, oraz $n_x^2+n_y^2+n_z^2-n_x'^2-n_y'^2-n_z'^2 = m_x^2+m_y^2+m_z^2-m_x'^2-m_y'^2-m_z'^2$. Nietrudno zauważyć, że im bardziej $n_x^2+n_y^2+n_z^2$ oraz $n_x'^2+n_y'^2+n_z'^2$ odległe są od $L^2p_F^2/h^2$, tj. od powierzchni Fermiego, tym większe są możliwości wyboru liczb m_x, m_y, m_z oraz m_x', m_y', m_z'. Dokładniejsza analiza wykazuje, że liczba stanów, w które może przejść cząstka czy też antycząstka o energii E, jest proporcjonalna do $E_F(1-E/E_F)^2$, gdzie $E_F = p_F^2/2m^*$ jest energią Fermiego.

Dotychczas nie określaliśmy energii wzbudzeń cząstki czy też antycząstki, zawsze określano energię wzbudzenia układu złożonego z cząstek i antycząstek. Właściwą definicję energii wzbudzenia pojedynczego otrzymamy, jeśli będziemy dążyli z energiami wszystkich wzbudzeń, z wyjątkiem jednego, do energii E_F, tzn. do niewzbudzania innych antycząstek — lub cząstek. W efekcie da to E_k-E_F dla cząstek i E_F-E_k dla antycząstek, co można zapisać jako $|E_k-E_F|$. Porównanie energii wzbudzenia z przytoczonym wyżej rzędem wielkości odwrotności czasu życia wskazuje, że przy $|E-E_F| \gg E_F$ wzbudzenia powinny być dobrze określone. Dla rozpatrywanego przez nas izotropowego modelu energie wzbudzeń układu cząstka–antycząstka będą dane w obszarze energii, w którym wzbudzenia te są dobrze określone (przy $p \approx p_F$), przez

energia wzbudzenia pojedynczego

$$E_p = |p^2-p_F^2|/2m^* = |p-p_F|\,(p+p_F)/2m^* \approx$$
$$\approx |p-p_F|V,$$

gdzie $V = p_F/m^*$ nazywa się prędkością cząstek na powierzchni Fermiego. Co więcej, nawet dla układów silnie oddziałujących, dowolne wzbudzenie o odpowiednio małej energii — w skali E_F — da się zestawić z określonych wyżej cząstek i antycząstek, co znaczy, że są to wzbudzenia elementarne. W takim sensie będziemy używali tego pojęcia również dla innych układów fizycznych. Nadmieńmy, że przedstawione rozważania dotyczą nie tylko rozpatrywanego tutaj jako przykład układu izotropowego. Dodajmy również, że sprawa wzbudzeń elementarnych staje się bardziej skomplikowana w wypadku elektronów w nadprzewodnikach, mimo iż podział wzbudzeń elementarnych na cząstki i antycząstki pozostaje aktualny.

pojęcie wzbudzenia elementarnego

Przedstawiony wyżej mechanizm rozpraszania cząstek Fermiego na sobie, a więc m.in. elektronów na elektronach w ciałach stałych, nie jest jedynym mechanizmem prowadzącym do powstania niestabilności wzbudzeń elementarnych. Innym mechanizmem jest np. rozpraszanie na kwantach drgań sieci (a więc fononach) oraz domieszkach i wszelkiego rodzaju niedoskonałościach struktury kryształu. Niemniej jednak, działanie dodatkowych mechanizmów rozpraszania nie sprawia, przynajmniej dla dostatecznie czystych i pozbawionych defektów kryształów, że wzbudzenia przestają być dobrze określone dla wzbudzeń o dostatecznie niskiej energii. Rozpatrywano tu również wzbudzenia ze stanu podstawowego, w którym układ znajduje się jedynie w temperaturze zera bezwzględnego, jak wiadomo — nieosiągalnej. Okazuje się jednak, że sytuacja pozostaje jakościowo niezmieniona, dopóki temperatura bezwzględna w skali energetycznej kT (k — stała Boltzmanna) jest dostatecznie niska. Warto przy tym zwrócić uwagę, że stwierdzenie, czy kT jest małe, nie jest sprawą tak prostą jak dla wielkości bezwymiarowej. W tym wypadku określenie niska czy też wysoka temperatura zależy od rozpatrywanego układu jak i rozpatrywanego zjawiska. Jeżeli oddziaływanie elektronów z fononami można pominąć, to w metalach jedyną wielkością o wymiarze energii, określającą układ elektronów jest E_F, przy czym $E_F/k \approx 5 \cdot 10^4$ K (→ Metale), a więc dla metali wystarczy niekiedy spełnić nierówność $T \ll 5 \cdot 10^4$ K.

wpływ temperatury na określoność wzbudzenia

Jeśli rozpatrywać bardziej skomplikowane zależności energii elektronów od wektora kwazipędu i uwzględniać, że może być więcej pasm częściowo zapełnionych, to model gazu nieoddziałujących elektronów opisuje dobrze jakościowo, a niekiedy i ilościowo, własności elektronowego układu metali, mimo że oddziaływanie elektronów między sobą jest

mniej więcej tak samo silne jak oddziaływanie elektronów z siecią krystaliczną. Jest w ogóle rzeczą zadziwiającą, jak bardzo przydaje się model idealnego gazu kwantowego (Fermiego lub Bosego) przy opisie niskotemperaturowych własności wielu układów fizycznych z silnym oddziaływaniem. Fakt ten wyjaśniono dopiero na gruncie kwantowej teorii wielu ciał w latach pięćdziesiątych; wyjaśnienie to było właściwie oparte na rozważaniach analogicznych do przeprowadzonych rozważań czasów życia poziomów antycząstkowych i cząstkowych w pobliżu powierzchni Fermiego. Niezależnie od tego nie należy sobie wyobrażać, że wzbudzenie w układzie silnie oddziałujących cząstek Fermiego polega na przesunięciu realnej cząstki spod powierzchni Fermiego nad tę powierzchnię należy to rozumieć w ten sposób, że istnieje wzajemnie jednoznaczna odpowiedniość pomiędzy wzbudzeniami realnego układu i wzbudzeniami idealnego układu nieoddziałujących cząstek, pod warunkiem, że wzbudzenia mają dostatecznie niską energię. Dlatego też słuszniejsze jest nazwanie tych wzbudzeń realnego układu odpowiednio kwazicząstkami lub też kwaziantycząstkami, a ogólnie — kwazicząstkami. Koncepcja kwazicząstek, uzasadniona tutaj w odniesieniu do układów cząstek Fermiego, znajduje powszechne zastosowanie we współczesnej fizyce fazy skondensowanej, a więc i ciała stałego. Mówiąc ogólnie, kwazicząstki to wyidealizowany układ gazowy, którego wzbudzenia o niskiej energii odpowiadają we wzajemnie jednoznaczny sposób wzbudzeniom układu realnego. Koncepcja kwazicząstek jest koncepcją naczelną w dalszej części tego artykułu.

kwazicząstki i kwaziantycząstki

Trochę o fononach; wzbudzenia w kryształach kwantowych

Dotychczas rozpatrywaliśmy wyłącznie wzbudzenia układów cząstek Fermiego, przy czym wzbudzenia ich podlegały również statystyce Fermiego. Nasze rozważania pozwoliły na wprowadzenie ogólnej koncepcji kwazicząstki. Charakterystycznym przykładem kwazicząstek typu Bosego, i to niezależnie od statystyki atomów tworzących kryształ, są fonony. (Omówiono je w artykule „Dynamika sieci krystalicznej"). Wiadomo stamtąd, że w przybliżeniu harmonicznym fonony są stabilne, tzn. że ich liczba przy ustalonym wektorze falowym i polaryzacji jest stała w czasie. Wiadomo również, że efekty anharmoniczne prowadzą do niestabilności fononów i, z drugiej strony, również do przesunięcia ich poziomów energii, przy czym wzbudzenia są w dalszym ciągu dobrze określone, zaś zmiany poziomów energetycznych niewielkie, jeżeli człony anharmoniczne są dostatecznie słabe. Jest to spełnione przy dostatecznie małej amplitudzie drgań atomów, tj. gdy $A \ll a$, przy czym A jest amplitudą drgań, zaś a — stałą sieci. Jeśli mechanika klasyczna opisywałaby kryształ, to odpowiednio małe amplitudy zgadzałyby się z odpowiednio małymi energiami wzbudzeń, przy zerowym wzbudzeniu amplituda byłaby zerowa. Dlatego też, przy dostatecznie niskich temperaturach, człony anharmoniczne byłyby pomijalnie małe. Do tego ujęcia wnosi istotne „ale" kwantowość zjawiska, ze względu na drgania zerowe atomów, zachodzące nawet w temperaturze zera bezwzględnego. Jeżeli ich amplituda jest znacznie mniejsza od stałej sieci a, to jesteśmy bliscy obrazu klasycznego i przy odpowiednio niskich temperaturach przybliżenie harmoniczne ma sens. Jeżeli jednak ta amplituda jest porównywalna z a — to przybliżenie harmoniczne nie będzie dobre nawet w temperaturze zera bezwzględnego; ten rodzaj kryształów nazywa się kryształami kwantowymi. Typowym przykładem kryształów kwantowych są kryształy obydwu izotopów helu, o masach atomowych 3 i 4; kryształy takie istnieją w temperaturach rzędu paru stopni Kelvina, przy ciśnieniach powyżej 25 hPa (^{4}He) i 30 hPa (^{3}He). Ogólnie — im mniejsza masa atomowa, tym większa

fonony

kryształy kwantowe

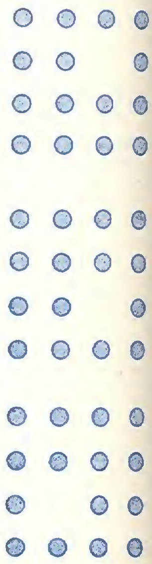
tendencja do kwantowości kryształu; tendencję tę można ocenić jakościowo nawet przy użyciu modelu harmonicznego, jeśli obliczyć amplitudę drgań zerowych.

Ponieważ dla kryształów niekwantowych model harmoniczny powinien być dobry przy odpowiednio niskich temperaturach, a w modelu harmonicznym fonony są stabilne — więc przy tak niskich temperaturach wzbudzenia będą dobrze określone dla wszystkich wektorów falowych $\vec{k}$ i polaryzacji λ. Podobna sytuacja może wystąpić również wówczas, gdy efekty anharmoniczne są istotne, a więc przy wyższych temperaturach, bądź też dla kryształów kwantowych. W tych obydwu wypadkach fonony również istnieją, choć przejawiają bardziej złożoną, bardziej kwazicząstkową strukturę. Warto dodać, że niezależnie od struktury kwazicząstek fononowych, z faktu że liczba fononów w danym stanie $(\vec{k}, \lambda)$ może być nieograniczona, wynika, że są to cząstki Bosego, niezależnie od statystyki atomów kryształu.

Pozostańmy nadal przy kryształach kwantowych. Odmienny charakter wzbudzeń fononowych nie jest ich jedyną cechą specyficzną. Odmiennie niż w innych niekwantowych kryształach przebiega tam również proces dyfuzji. Jest to proces samodyfuzji, czyli przypadkowego błądzenia atomów wśród pozostałych atomów kryształu, identycznych z błądzącymi. Atomy błądzące wykorzystują przy tym głównie istnienie niezajętych miejsc w sieci krystalicznej, zwanych lukami. Dyfuzja polega na zajęciu luki przez sąsiedni atom, co tworzy lukę w odpowiednim sąsiednim węźle, kolejnym zajęciu tej luki przez sąsiedni atom, co powoduje powstanie luki w nowym miejscu, itd. Jeśli potraktować lukę jako antycząstkę (str. 558, zwróćmy uwagę, że tam była mowa o poziomach energetycznych, tu zaś o miejscach w sieci) to można cały proces opisać jako dyfuzję luki (rys. 1).

luka kwantowa

Wszystkie czyste substancje, poza ^{3}He i ^{4}He, w odpowiednio niskich temperaturach przechodzą w stan stały, również przy zerowym ciśnieniu zewnętrznym. Wynika stąd, że kryształ z wszystkimi węzłami regularnej sieci krystalicznej zajętymi przez atomy ma najniższą energię. Powoduje to powstanie barier energii potencjalnej między węzłami sieci. Bariery takie uniemożliwiają przeskok atomu do sąsiedniego miejsca, nawet jeżeli jest tam luka, dopóki atom nie uzyska, dzięki wzbudzeniu termicznemu energii wystarczającej do pokonania bariery potencjału. Obniżaniu temperatury towarzyszy zatem zmniejszanie się możliwości zarówno tworzenia luk, jak i pokonywania bariery potencjału, a więc zmniejszanie się współczynnika dyfuzji. (Rozważania te są słuszne przy traktowaniu atomu niekwantowo). Mechanika kwantowa dopuszcza również, oprócz dozwolonego klasycznie skoku nad barierą potencjału, tunelowe przejście przez barierę, tj. przejście atomu o energii mniejszej od wysokości bariery. Efekty kwantowe są oczywiście tym istotniejsze, im dyfundujący atom jest lżejszy (nikt bowiem dotychczas — nie bez racji — nie kwantował równania ruchu beczek). Stąd, ten podbarierowy mechanizm dyfuzji kwantowej, dotyczy, jeśli nie brać pod uwagę nierozsądnie małych prawdopodobieństw, jedynie atomów najlżejszych, takich jak izotopy helu o liczbie masowej 3 i 4 (wodór z reguły nie występuje w postaci atomowej, zaś jego cząsteczka, ze względu na rozmiary, dyfunduje w sposób utrudniony). Do dyfuzji jest jednak potrzebne występowanie luk, które powinny zanikać przy zerowej temperaturze bezwzględnej. Czy zatem brak luk w kryształach ^{3}He i ^{4}He w dostatecznie niskich temperaturach powstrzymuje opisany tutaj proces podbarierowy? Okazuje się, że nie. Rzecz w tym, że zdolność luki do dyfundowania pod barierą potencjału może prowadzić do istnienia luk w równowagowym krysztale kwantowym nawet w zerowej temperaturze bezwzględnej. Spróbujmy wykazać tę możliwość.

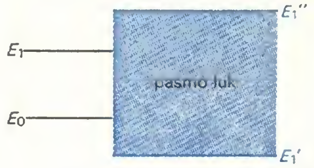


Rys. 1. Kolejne etapy dyfuzji luki w kwadratowej sieci płaskiej

dyfuzja kwantowa

Dla luki kwantowej, mogącej swobodnie dyfundo-

wać, kryształ jest periodycznym polem sił. Ruch w takim polu opisuje pasmowa teoria ciał stałych. Energia luki ulega rozszczepieniu w całe pasmo energetyczne, a stan kwantowy luki jest opisywany przez wektor kwazipędu. Przy dostatecznie silnym rozszczepieniu energii luki, dno pasma może znaleźć się poniżej najniższej energii kryształu bez luk — i wtedy wytworzenie pewnej liczby luk staje się energetycznie dogodne. Popatrzmy na rys. 2 ilustrujący tę możliwość. Jego lewa część przedstawia sytuację klasyczną. Energia E_0 oznacza energię kryształu bez luk,



Rys. 2

E_1 — energię kryształu z jedną luką. W wyniku dyfuzji podbarierowej poziom E_1 rozszczepia się w całe pasmo dozwolonych energii o maksimum E_1'' i minimum E_1', co pokazuje prawa część rysunku. Jeśli dno pasma znajdzie się poniżej E_0 (energia $E_1' < E_0$), to wytworzenie pewnej liczby luk staje się energetycznie dogodne. Dzięki temu luki powinny się znajdować również w równowagowym krysztale przy zerowej temperaturze bezwzględnej. Należy dodać, że przy określaniu struktury takich kryształów nie stwierdzono by istnienia zlokalizowanych luk, byłyby więc one „rozmazane" po całym krysztale. Można powiedzieć, że funkcja rozkładu gęstości atomów w krysztale byłaby nadal periodyczna, z tym że prawdopodobieństwo znalezienia atomu w węźle sieci krystalicznej byłoby mniejsze od jedności. Ponieważ luki kwantowe opisuje teoria pasmowa, można więc stanowi luki przypisać wektor kwazipędu. Taka luka jest specyficzną kwazicząstką, której występowanie jest charakterystyczne dla kryształów kwantowych.

luka jako kwazicząstka

Prawdopodobieństwo dyfuzji kwantowej maleje wraz ze wzrostem temperatury, głównie ze względu na zderzenia luk z fononami. Dzięki podobnemu mechanizmowi zderzeń elektronów z fononami opór elektryczny metali maleje wraz z temperaturą. Dla kryształów kwantowych, w których luki istnieją również w zerze bezwzględnym, współczynnik dyfuzji (analog przewodnictwa, nie zaś oporu metali) powinien wzrastać ze spadkiem temperatury, przy dostatecznie niskich temperaturach — gdy dyfuzja kwantowa odgrywa główną rolę. Jest to akurat odwrotnie niż przy normalnej dyfuzji, dla której, wraz ze wzrostem temperatury, wzrasta zarówno liczba wytworzonych luk jak i prawdopodobieństwo nadbarierowego wzbudzenia atomów, a więc i współczynnik dyfuzji. Dlatego współczynnik dyfuzji, malejący wraz z temperaturą przy odpowiednio wysokich temperaturach, staje się rosnącą funkcją temperatury przy temperaturach odpowiednio niskich. Fakt ten, niezrozumiały w klasycznym modelu dyfuzji, został potwierdzony doświadczalnie na kryształach helu.

Omówiona luka kwantowa nie jest jedyną kwazicząstką mogącą występować w kryształach kwantowych. Mogą tu również wystąpić stany związane dwu lub więcej luk; podobnie, lekki atom domieszki w międzywęźlu kryształu kwantowego, dzięki zjawisku tunelowemu, zachowuje się jak kwazicząstka, której można przypisać kwazipęd. Dodajmy jeszcze, że podany tu opis dyfuzji w krysztale kwantowym i związanych z tym wzbudzeń został zaproponowany przez I. M. Lifszyca i A. F. Andrejewa.

Plazmony

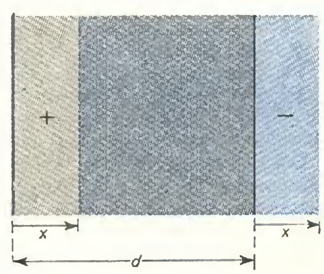
Aby zrozumieć wzbudzenia plazmowe rozważmy ekranujące działanie elektronów. Mechanizm tego zjawiska jest niezwykle prosty. Wyobraźmy sobie, że do elektrycznie obojętnego ośrodka zawierającego elektrony swobodne, w takim np. sensie jak elektrony w metalu, wprowadzono nadmiarowy ładunek dodatni. Ładunek ten przyciąga elektrony, dzięki czemu ich gęstość w otoczeniu ładunku jest coraz większa. Inne cząstki naładowane, a więc jony, mają znacznie większą masę i są zlokalizowane, wobec czego zmiana gęstości ich ładunku jest znacznie trudniejsza. Wzrost ładunku w pobliżu ładunku wprowadzonego kompensuje jego działanie na dużych odległościach. Podany mechanizm ekranowania statycznego może ulec istotnej modyfikacji, gdy wprowadzony nadmiarowy ładunek porusza się — mówimy wtedy o ekranowaniu dynamicznym. Jeśli wprowadzony nadmiarowy ładunek jest ładunkiem ujemnym, to elektrony zostają odepchnięte. Wokół wprowadzonego ładunku powstaje nadmiar ładunku dodatniego, który z kolei przyciąga elektrony itd., co przekonuje nas, że mechanizm ekranowania działa również w stosunku do ładunków ujemnych.

ekranowanie statyczne i dynamiczne

Wyobraźmy sobie teraz, że nadmiarowy ładunek, np. dodatni, został nagle wprowadzony do ośrodka. Przyciągane przez ten ładunek elektrony zaczęły się poruszać w jego kierunku. Po nabraniu pewnej prędkości elektrony będą miały skłonność do poruszania się na większą odległość, niż wynikałoby to z ich równowagowego rozkładu w obecności ładunku nadmiarowego. Ulegną one z kolei działaniu sił elektrostatycznych działających w przeciwnym kierunku. I znowu, ze względu na niezerową prędkość poruszania się i bezwładność elektronów, będą one miały skłonność do dotarcia za daleko. Powtarzanie się tego mechanizmu prowadzi do pojawienia się drgań zwanych drganiami plazmy; ich mechanizm jest nierozerwalnie związany z ekranowaniem.

drgania plazmy

Spróbujmy otrzymać wzór na częstość tych drgań. W tym celu założymy, że ładunek dodatni, który ma znacznie mniejsze możliwości poruszania się niż elektrony, jest całkowicie nieruchomy i że został jednorodnie „rozmazany" po całym krysztale. Pierwsze z tych założeń jest idealizacją istniejącego stanu, drugie zaś związane jest z faktem, że interesują nas wzbudzenia o bardzo długiej fali, dla których niewielkie (w skali atomowej) niejednorodności rozkładu ładunku nie odgrywają żadnej roli. Załóżmy również, że nierównowagowy rozkład ładunku został spowodowany przez przesunięcie naładowanej warstwy o grubości d na niewielką odległość x ($x \ll d$, rys. 3).



Rys. 3

Warstwa taka zachowuje się jak kondensator płaski. Powstaje przy tym stałe pole elektryczne $\mathcal{E}$, działające w takim kierunku, aby przywrócić elektryczną obojętność całej warstwy. Natężenie pola elektrycznego określa wzór

$$\mathcal{E} = -4\pi Nex,$$

w którym N jest koncentracją, tj. liczbą elektronów swobodnych w jednostce objętości, a e — ładunkiem elektronu. Nazwa „elektrony swobodne" została tu użyta nie dla określenia elektronów nieoddziałujących ze sobą, ale elektronów mogących swobodnie przepływać. Wzór na natężenie pola elektrycznego wynika ze wzoru na pojemność kondensatora płaskiego przy uwzględnieniu, że gęstość ładunku powierzchniowego na obu okładkach kondensatora wynosi odpo-

wienio $\pm Nex$. Ruch elektronu wewnątrz warstwy opisuje równanie

$$ma = e\mathcal{E}, \quad \text{czyli} \quad m(d^2x/dt^2) = -4\pi Ne^2x,$$

w którym m jest masą elektronu. Jego rozwiązaniem jest ruch harmoniczny o częstości

$$\omega_p = (4\pi Ne^2/m)^{1/2},$$

zwanej częstością plazmową. Wzbudzenia gęstości elektronów swobodnych, w podanym wyżej sensie, nazywamy plazmonami. Jako kwazicząstki będące wynikiem kwantowania ruchu oscylatora podlegają one statystyce Bosego. Proste oszacowanie wielkości $h\omega_p/2\pi$ przy założeniu koncentracji elektronów charakterystycznych dla metali pokazuje, że jest ona rzędu energii Fermiego. Nie jest to więc wzbudzenie nisko leżące i jego wzbudzenie termiczne jest wysoce nieprawdopodobne, ponieważ $kT \ll E_F$, aż do temperatury topnienia metalu włącznie. Dlatego też statystyka tych wzbudzeń nie odgrywa istotnej roli przy termicznym wzbudzeniu plazmonów, ponieważ kwestia, czy w danym stanie może istnieć jeden plazmon czy więcej, w sytuacji gdy ich prawie zupełnie nie ma, jest raczej pytaniem akademickim. Zdanie to nie sugeruje bynajmniej, że plazmony nie są obserwowane, wzbudzają się one np. przy przechodzeniu cząstek naładowanych przez metal, lub też przy odbiciu od jego powierzchni. Eksperymenty takie wykazują, że plazmony są z reguły wzbudzeniami dobrze określonymi.

częstość plazmonów

Dotychczasowe rozważania dotyczyły sytuacji granicznej — długich fal, czyli zerowej wartości wektora falowego. Wnikliwsza analiza wykazuje, że częstość plazmonów zależy od wektora falowego, a więc, że plazmony podlegają dyspresji. Główna poprawka do wyrażenia na ω_p ma następującą postać w granicy długofalowej: $\gamma hk^2/2\pi m$, gdzie k^2 oznacza kwadrat długości wektora $\vec{k}$, a γ jest wielkością bezwymiarową rzędu jedności ($\gamma \approx 0{,}5$ dla Be, Mg i Al). Warto dodać, że w teorii uwzględniającej oddziaływanie elektronów między sobą, otrzymuje się również wyrażenie na ω_p przy $\vec{k} = 0$ takiej samej postaci, ale tym razem m oznacza masę krystaliczną elektronu, zależną od oddziaływania elektronu z siecią krystaliczną, a niezależną od oddziaływania międzyelektronowego.

Przy opisie plazmonów powoływaliśmy się głównie na przykłady wzbudzeń w metalach. Jednak wzbudzenia takie istnieją we wszystkich ciałach stałych zawierających swobodne nośniki, a więc również w półprzewodnikach. Warto zastanowić się w tym miejscu nad elementarnością wzbudzeń. Dla wzbudzeń w fazie skondensowanej, a więc w ciele stałym lub w cieczach kwestia ta, podobnie jak elementarność w fizyce cząstek, nie jest raz na zawsze rozwiązana, całkowicie jednoznaczna i nie sprawiająca kłopotów. Zaproponowane tutaj kryterium elementarności wzbudzeń zwraca uwagę na dwa aspekty zagadnienia — na niemożność złożenia danego wzbudzenia z pewnych innych wzbudzeń oraz — możliwość występowania wzbudzenia w innych, bardziej złożonych wzbudzeniach. W szczególności, jeśli przy tworzeniu takiego wzbudzenia istotne jest oddziaływanie między wzbudzeniami składającymi się na nie i we wzbudzeniu tym bierze udział duża liczba wzbudzeń składowych, to wzbudzenie takie nazywa się kolektywnym. Nietrudno zauważyć, że kwazicząstki elektronowe i fonony są wzbudzeniami elementarnymi. Jeśli mowa o plazmonach, to można je traktować jako wzbudzenia kolektywne, złożone z elementarnych wzbudzeń kwazicząstkowych i kwaziantycząstkowych. Przytoczone klasyczne wyprowadzenie wyrażenia na częstość plazmową ani nie potwierdza, ani nie odrzuca tego punktu widzenia. Warto jednak zwrócić uwagę na to, że dowolna, dostatecznie długofalowa zmiana gęstości przestrzennej elektronów może być wywołana powstaniem przewagi kwazicząstek w jednym obszarze przestrzennym, kwaziantycząstek zaś — w innym, oraz na dominującą rolę oddziaływania elektronów przy drganiach plazmy. Mówienie o kwazicząstkach, czy też o kwaziantycząstkach wymaga dostatecznie dobrego rozróżnienia ich kwazipędu, aby można było stwierdzić, czy są to wzbudzenia spod, czy też znad powierzchni Fermiego, czyli Δp musi być znacznie mniejsza od pędu Fermiego $p_F \sim h/a$, gdzie a jest stałą sieci. Z drugiej strony $\Delta x \, \Delta p \geqslant h/4\pi$ zgodnie z zasadą nieokreśloności, co daje razem oszacowanie rozmiarów liniowych wspomnianych obszarów przestrzennych: $\Delta x \gg a$. Ponadto, jeśli mamy precyzyjnie określać rozchodzenie się fali, to Δx musi być wyraźnie mniejsze od długości fali λ, a więc jeśli $\lambda \gg a$, to opis mikroskopowy, w języku kwazicząstek elektronowych, jest równoważny opisowi makroskopowemu w języku zmiany gęstości przestrzennej elektronów.

elementarność wzbudzeń

wzbudzenia kolektywne

Ekscytony i polarony

Omówimy pokrótce jeszcze dwa rodzaje wzbudzeń — ekscytony i polarony, które z punktu widzenia podanej tu definicji nie są wzbudzeniami elementarnymi, mimo iż ekscytony często zalicza się do wzbudzeń elementarnych. Ekscytony są stanami związanymi elektronu i dziury; występują one głównie w półprzewodnikach, półmetalach i kryształach molekularnych. Natomiast polarony, będące stanami związanymi elektronu lub dziury z paroma fononami, występują z reguły w kryształach jonowych i półprzewodnikach polarnych. W tym ostatnim wypadku polaron powstały z elektronu i polaron powstały z dziury mogą związać się również w specyficzny ekscyton.

ekscyton

Wyobraźmy sobie elektron i dziurę. Dziurze można przypisać ładunek przeciwny do ładunku elektronu, stąd elektron i dziura będą się przyciągały. Załóżmy na chwilę, że periodyczne pole kryształu nie istnieje. Wtedy ruch elektronu i dziury można rozłożyć na dwie składowe: ruch środka masy — jednostajny i prostoliniowy — oraz ruch względny. Przy działaniu siły przyciągania kulombowskiego ruch względny może odbywać się po elipsie, paraboli i hiperboli. W ostatnich dwóch sytuacjach cząstki oddalają się nieograniczenie od siebie, co odpowiada stanowi niezwiązanemu. Przy ruchu eliptycznym odległości między cząstkami są ograniczone od góry; taki ruch podlega kwantowaniu identycznemu jak w wypadku atomu wodoru. Może tu wprawdzie zabraknąć, mimo lekkich elektronów i ciężkich dziur w półprzewodnikach, tak wyróżnionej cząstki centralnej jak proton w atomie wodoru. Całkowita energia układu cząstki i dziury jest sumą energii ruchu środka masy i energii ruchu względnego. Pierwsza z nich zależy w sposób ciągły od pędu ($E = P^2/2M$, gdzie M jest sumą mas dziury i elektronu, a P — sumą ich pędów), natomiast druga z nich tworzy znaną serię poziomów wodoru, tyle że z innym współczynnikiem proporcjonalności przed $-1/n^2$, $n = 1, 2, \ldots$. Stąd energia całkowita układu ma postać: $-\beta/n^2 + P^2/2M$. W warunkach działania periodycznego pola kryształu podział ruchu na dwie niezależne składowe — ruch środka masy oraz ruch względny — staje się w zasadzie niemożliwy; rozkład taki można przeprowadzić jedynie w pobliżu ekstremum pasm energetycznych. Zatem zależność energii ekscytonu od głównej liczby kwantowej n jak i od pędu $\vec{P}$ ulegnie skomplikowaniu tym bardziej, że $\vec{P}$, ze względu na działanie periodycznego pola kryształu, jest teraz kwazipędem. Dzięki temu możemy mówić o pasmach ekscytonowych. Ponieważ przy przestawieniu wszystkich liczb kwantowych opisujących dwie jednakowe cząstki Fermiego (fermiony) funkcja falowa zmienia znak, a przy analogicznym przedstawieniu liczb kwantowych cząstek Bosego (bozonów) — pozostaje niezmieniona, więc przy przestawieniu dwóch zespołów liczb kwantowych opisujących układy składające się z dwóch fermionów — funkcja również nie zmienia znaku, co

pasma ekscytonowe

dowodzi, że ekscytony podlegają statystyce Bosego. Ze względu na zerowy wypadkowy ładunek ekscytony nie biorą udziału w transporcie ładunku, wpływają natomiast na własności optyczne kryształów. Rozróżnia się zazwyczaj 2 rodzaje ekscytonów: ekscytony Frenkla, występujące z reguły w kryształach molekularnych, w których elektron i dziura są praktycznie biorąc zlokalizowane w molekule, oraz ekscytony Wanniera-Motta, o wysokiem stopniu delokalizacji, występujące z reguły w półprzewodnikach.

ekscytony Frenkla

ekscytony Wanniera–Motta

Warto zwrócić uwagę na to, że energie ekscytonów odpowiadają energiom z przerwy energetycznej, a więc niedopuszczalnym dla kwazicząstek elektronowych. Energia wzbudzenia układu elektron + dziura, jeśli obydwie kwazicząstki znajdują się blisko ekstremów pasma przewodnictwa i pasma walencyjnego, jest bowiem bliska wielkość przerwy energetycznej E_g. Do tego dochodzi ujemna energia oddziaływania kulombowskiego elektronu i dziury, co powoduje, że energia ekscytonu jest mniejsza od E_g. Z drugiej jednak strony, energia ta nie może być ujemna, wskazywałoby to bowiem na fakt, że wiązanie się elektronów i dziur w ekscytony — tzw. przejście ekscytonowe — obniżałoby energię układu, a więc układ byłby nietrwały. Zjawisko przejścia ekscytonowego pojawia się niekiedy w półmetalach. Podane oszacowanie energii ekscytonu nie zabrania istnieć ekscytonom o energii większej od E_g. Można jednak wykazać, że takie ekscytony przestają być wzbudzeniami dobrze określonymi, gdyż ulegają rozpadowi na wzbudzenia elektronowe, a więc dysocjację ekscytonu. Mimo nieelementarności ekscytonów z punktu widzenia przyjętej definicji, można je traktować jako składowe bardziej skomplikowanych wzbudzeń.

przejście ekscytonowe

W specyficznych warunkach, przy dużej gęstości ekscytonów w tzw. kroplach ekscytonowych generowanych w półprzewodnikach impulsem laserowym, istotne stają się tak oddziaływania międzyekscytonowe, jak i wzbudzenia kolektywne układu ekscytonów. Ponieważ ekscytony podlegają statystyce Bosego, oddziaływania międzyekscytonowe mogą wręcz doprowadzić do przejścia kropli ekscytonowej w stan nadciekły podobnie jak to się dzieje z helem ^{4}He.

polarony

Dodajmy parę słów o polaronach. Elektron oddziałuje z jonami sieci, zatem pojawienie się zmiany w równowagowej gęstości elektronów musi powodować zmianę w rozkładzie gęstości jonów, czyli polaryzować sieć. Taką zmianę rozkładu gęstości jonów można uzyskać przez nałożenie na siebie pewnej liczby fal wychyleń jonów z położenia równowagi o różnej długości fali — tj. przez związanie elektronu z pewną liczbą fononów. Podobnie jak w wypadku ekscytonów wyróżniamy tu polarony o małym promieniu, dobrze zlokalizowane, oraz polarony o dużym promieniu, rzędu wielu stałych sieci krystalicznej. Proces wiązania elektronów z fononami powinien prowadzić do zwiększenia masy powstałej kwazicząstki, co jest intuicyjnie jasne. Analogicznemu procesowi „obrastania fononami" może ulegać nie tylko elektron, ale i dziura. Stąd, jak już wspomniano, wynika możliwość występowania polaronowego ekscytonu, będącego stanem związanym polaronowej dziury i polaronowego elektronu. Wzbudzenia takie można obserwować w półprzewodnikach polarnych, o dużej biegunowości wiązań. Rozważania podobne jak dla ekscytonów wykazują fermionowy charakter polaronów.

Uzupełniając poprzedni opis można powiedzieć, że w wyniku polaryzowania sieci przez elektron powstaje jama potencjału wychwytująca elektron. Polaron nie jest jedyną kwazicząstką o tym charakterze w fazie skondensowanej. Wśród kwazicząstek podobnego typu można wymienić np. fluktuon i fazon. Pierwszy z nich to elektron wychwycony przez jamę potencjału wytworzoną w wyniku fluktuacji składników w pobliżu elektronu w stopie nieuporządkowanym; drugi polega na wychwyceniu elektronu przez obszar fazy ferromagnetycznej powstały w paramagnetyku.

fluktuon i fazon

Czy da się zamknąć listę kwazicząstek? Polarytony

Ostatnie dwie kwazicząstki omówiliśmy w skrócie niemalże telegraficznym. Nie od rzeczy będzie pytanie o kryteria wyboru kwazicząsteczek omówionych w tym artykule. Liczba rozmaitych wzbudzeń w fazie skondensowanej, elementarnych lub nie — w podanym wyżej sensie, jest tak wielka, że trudno byłoby znaleźć fizyka do skompletowania listy wzbudzeń „na dzisiaj" na podstawie tego, co aktualnie wie, nie zaś na podstawie dodatkowej lektury. Zrezygnowaliśmy tu więc z omawiania tych wzbudzeń, przy których dublowalibyśmy fragmenty innych artykułów. Dotyczy to fononów (→ Dynamika sieci krystalicznej), magnonów (→ Teoria magnetyzmu) oraz specyficznych kwazicząstek elektronowych charakterystycznych dla nadprzewodników (→ Nadprzewodnictwo). Ze względu na specyfikę zagadnienia pominęliśmy również rozmaite wzbudzenia metali w zewnętrznym silnym polu magnetycznym (fale spinowe w metalach normalnych, w stałym zewnętrznym polu magnetycznym, fale cyklotronowe, helikony itd.). W artykule tym położyliśmy nacisk raczej na wzbudzenia elementarne niż kolektywne i, z drugiej strony, na wzbudzenia powszechnie występujące, zdając sobie oczywiście sprawę z nieostrości tego ostatniego kryterium. Było ono wprawdzie zlekceważone przy opisie wzbudzeń w kryształach kwantowych, ale usprawiedliwiała nas wtedy niezwykłość sytuacji fizycznej.

Ponieważ kontynuowanie przeglądu kwazicząstek w formie wyliczenia przypominającego książkę telefoniczną jest raczej pozbawione sensu, zamkniemy listę omówieniem polarytonu. Jest to wzbudzenie tym różniące się jakościowo od już omówionych lub wzmiankowanych, że stanowi ono stan związany fotonu z którymś ze wzbudzeń, takich jak fonon optyczny, ekscyton, magnon itp. Omówimy pokrótce polarytony fononowe i polarytony ekscytonowe. Musimy najpierw wyjaśnić, w jaki sposób fala elektromagnetyczna może oddziaływać z fononem optycznym lub ekscytonem. Otóż wektor elektryczny fali elektromagnetycznej wywołuje polaryzację elektryczną ośrodka. Ze względu na falowy charakter ruchu wektora elektrycznego polaryzacja ta ma również charakter falowy. Warto przy tym zwrócić uwagę, że polaryzacja taka, przynajmniej w ośrodkach o niezbyt silnej anizotropii, jest poprzeczna, tzn. wektor polaryzacji jest prostopadły do kierunku rozchodzenia się fali, ponieważ fala elektromagnetyczna jest poprzeczna. Zastanówmy się, jakiego typu wzbudzenia mogą być indukowane przez wektor elektryczny. Jeżeli sieć składa się z co najmniej dwóch podsieci, obsadzonych przez atomy różnych pierwiastków, to na atomach poszczególnych podsieci będzie znajdował się nieco inny ładunek elektryczny. Fonon optyczny o zerowym wektorze falowym odpowiada jednorodnemu przesunięciu podsieci względem siebie tak, aby środek masy całego kryształu pozostawał nieruchomy. Przy niecałkowitym skompensowaniu ładunku elektrycznego na atomach każdej podsieci odpowiada to zmianie momentu dipolowego kryształu, a więc jego polaryzacji elektrycznej. Możemy stąd wywnioskować, że elektryczne pole fali elektromagnetycznej będzie wzbudzało fonony optyczne o polaryzacji poprzecznej, ze względu na poprzeczny charakter fal elektromagnetycznych. Natomiast fonony akustyczne odpowiadają drganiom wszystkich podsieci w fazie, co nie prowadzi do zmiany momentu dipolowego kryształu, a zatem pole elektryczne nie wpływa na fonony akustyczne. Oddziaływanie między falą elektromagnetyczną a polaryzacją dynamiczną sieci ma charakter przyciągania, bowiem sieć polaryzuje się w taki sposób, aby zmniejszyć energię łącznego układu promieniowania + sieć. Zgodnie z zasadami mechaniki kwantowej istnieje zatem możliwość powstania stanu związanego porzecznej polaryzacji sieci, tj. poprzecz-

oddziaływanie fali elektromagnetycznej z fononem lub ekscytonem

polarytony fononowe

nie spolaryzowanych fononów optycznych, z kwantami pola elektromagnetycznego, czyli fotonami. Wiązanie to będzie szczególnie silne przy częstościach promieniowania elektromagnetycznego bliskich częstości fononów optycznych. Wynika to z faktu, że przy zbliżonych do siebie częstościach pola zewnętrznego i fononów optycznych mamy do czynienia z rezonansem i polaryzacja jest szczególnie duża. Biorąc typowy rząd wielkości częstości fononów optycznych ω_0 i oceniając na tej podstawie długość fali promieniowania elektromagnetycznego ($\lambda = 2\pi c/\omega_0$) znajdujemy $\lambda \approx 10^4$ nm. Fonony optyczne będą zatem wiązać najsilniej fale bardzo długie w porównaniu z odległością międzyatomową (rzędu 0,1 nm) i polaryton fononowy będzie kwazicząstką o wysokim stopniu delokalizacji. Opisane kwazicząstki odgrywają istotną rolę w optycznych własnościach tych kryształów, dla których wzbudzenie fononu optycznego zmienia moment dipolowy kryształu.

polarytony ekscytonowe

Przejdźmy teraz do polarytonów ekscytonowych. Zasada wiązania fali elektromagnetycznej z ekscytonami jest taka sama, jak w wypadku fononów optycznych, ponieważ ekscyton jest stanem związanym kwazicząstek naładowanych. Powstała kwazicząstka będzie z wielu względów podobna do opisanego poprzednio polarytonu fononowego. Wprowadzenie koncepcji polarytonu ekscytowanego zmieniło w sposób istotny opis własności optycznych półprzewodników. Przedtem uważano, że fala elektromagnetyczna o częstości mniejszej od częstości odpowiadającej szerokości przerwy energetycznej może wzbudzać jedynie ekscytony. Obecnie sądzi się, że taka fala elektromagnetyczna wzbudza również polarytony ekscytonowe, co komplikuje opis własności optycznych półprzewodników. Jest to fakt, który został potwierdzony doświadczalnie. Dodajmy jeszcze, że analogicznie do wzmiankowanych tu polarytonów magnonowych wektor pola magnetycznego fali elektromagnetycznej sprzęga się z dynamiczną polaryzacją magnetyczną, a więc magnonami.

wykrywanie wzbudzeń

Warto jeszcze na zakończenie powiedzieć trochę o wykrywaniu wzbudzeń w układach fizycznych. Załóżmy, że na układ działa pole zewnętrzne o określonej częstości i określonym wektorze falowym. Wektor falowy $\vec{k}$ i częstość kołowa ω pomnożone przez $h/2\pi$ dają odpowiednio pęd i energię kwantu pola. Załóżmy, że pole to wzbudza kwazicząstki określonego typu. Jeśli układ fizyczny jest ciałem stałym, to obowiązuje prawo zachowania kwazipędu, z dokładnością do wektora sieci odwrotnej pomnożonego przez $h/2\pi$, oraz prawo zachowania energii. Jeśli więc będą spełnione równości $E(\vec{k}_1)+h\omega(\vec{k})/2\pi = E(\vec{k}_2)$ oraz $\vec{k}_1+\vec{k} = \vec{k}_2+\vec{K}_n$, gdzie $E(\vec{k})$ energia kwazicząstek zależna od wektora falowego, a $\vec{K}_n$ — wektor sieci odwrotnej, to układ będzie pochłaniał energię pola. Dla absolutnie stabilnych kwazicząstek niespełnienie którejkolwiek z tych równości powoduje niemożność pochłaniania energii przez układ w reakcji: kwazicząstka + kwant energii → inna kwazicząstka, spełnienie zaś obydwu tych równości prowadzi do wystąpienia nieskończenie cienkiej linii absorpcji. Dla dobrze określonych wzbudzeń kwazicząstkowych energie będą nieco rozmyte, zgodnie z zasadą nieokreśloności dla energii, $\Delta E\tau \geqslant h/2\pi$, i wystarczy, aby pierwsze z wymienionych równań było spełnione jedynie w przybliżony sposób, z dokładnością rzędu $h/2\pi\tau$. Powoduje to, że nieskończenie cienka linia absorpcji zmienia się w wąską linię o szerokości rzędu $h/2\pi\tau$ w skali energii. Dla wzbudzeń, które nie są dobrze określone, linie absorpcji stają się liniami szerokimi. Wspomniane procesy absorpcji wraz z procesami rozpraszania, dającymi się wyrazić symbolicznie w sposób następujący: kwazicząstka + kwant energii → inną kwazicząstkę + inny kwant energii, spełniającymi odpowiednie prawa zachowania energii i kwazipędu, są podstawowymi metodami zdobywania informacji o widmie kwazicząstek. Warto zaznaczyć, że oprócz rozpraszania kwantów stosuje się również rozpraszanie cząstek, np. rozpraszanie neutronów, dające podstawowe informacje o widmie fononów.

wzbudzenia powierzchniowe

Omawialiśmy tu wzbudzenia związane ze wzbudzeniami wnętrza trójwymiarowego kryształu. Nieomal każde z tych wzbudzeń ma swój analog związany z powierzchnią kryształu. Wzbudzenia takie, jak zwykle w fizyce powierzchni (→ Stany powierzchniowe w ciałach stałych), są opisywane dwuwymiarowym wektorem kwazipędu, określonym przez sieć płaską, odwrotną do powierzchni kryształu, stanowiącej jego granicę. W niektórych wypadkach rozważania nad wzbudzeniami powierzchniowymi niewiele się różnią od rozważań nad wzbudzeniami objętościowymi.

M. I. Kaganow *Elektrony, fonony, magnony...* Warszawa 1978; M. I. Kaganow, I. M. Lifszyc, *Kwaziczasticy*, Moskwa 1976; W. Wardzyński, *Polarytony*, Post. Fiz. 27, 477 (1976).

Stany powierzchniowe w ciałach stałych

Jacek Łagowski i Andrzej Morawski

Powierzchnia rozumiana jako ograniczenie bryły ciała stałego lub, w złożonych układach, jako granica pomiędzy różnymi fazami pełni istotną rolę w wielu zjawiskach, będących przedmiotem równoczesnego zainteresowania szeregu dziedzin nauki. Spośród szczególnie aktualnych zagadnień związanych z powierzchnią wymienić można badania mechanizmów funkcjonowania układów biologicznych, zjawiska katalizy itp.

Zainteresowanie powierzchnią ciała stałego, a zwłaszcza procesami fizycznymi zachodzącymi na granicy złącza metal–półprzewodnik datuje się od 1876 r., kiedy to odkryto zjawisko fotowoltaiczne w tego rodzaju złączach. Znane już wtedy było prostujące działanie kontaktu metal–półprzewodnik (1874 r.). Jednak dopiero prace z lat pięćdziesiątych XX w. przyczyniły się do istotnego postępu w dziedzinie wiedzy o powierzchni. Zapoczątkowany w latach dwudziestych naszego stulecia burzliwy rozwój fizyki ciała stałego pozwala w stopniu zadawalającym opisywać i rozumieć własności wnętrza kryształów. Postęp w technologii umożliwiający otrzymywanie dużych, jednorodnych monokryształów zbliżył rezultaty badań doświadczalnych do teoretycznych. Teoretyczny model krystalicznego ciała stałego zakłada translację komórki elementarnej aż do nieskończoności; nie uwzględnia zatem ograniczenia przestrzennego kryształu. Wprowadzenie takiego ograniczenia, tzn. powierzchni, pociąga za sobą daleko idące konsekwencje. Jak wykazał I. E. Tamm (1932 r.), występujące na powierzchni zaburzenie periodyczności sieci krystalicznej i w związku z tym zaburzenie periodyczności potencjału na granicy kryształ–próżnia prowadzi do pojawienia się, nawet w idealnych kryształach półprzewodnikowych, dozwolonych elektronowych stanów energetycznych w obszarze energii wzbronionych (tzw. stanów powierzchniowych Tamma). Zagadnieniem tym zajmowało się następnie wielu fizyków. Do szczególnego postępu wiedzy o stanach powierzchniowych dopro-

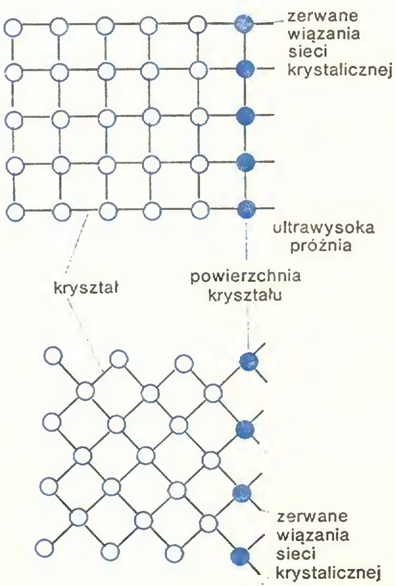
stany powierzchniowe Tamma

wadziły prace W. Shockleya. Wynalezienie w 1948 r. tranzystorów przez J. Bardeena, W. Brattaina i W. Shockleya, przy okazji badań powierzchni półprzewodników, w istotny sposób wzmogło zainteresowanie fizyką półprzewodników i w szczególności problematyką stanów powierzchniowych, mających ogromne znaczenie w działaniu kontaktów metal–półprzewodnik, złączy *p-n*, tranzystorów, a zwłaszcza układów scalonych. Dotychczas jednak nie zostały rozwiązane zasadnicze problemy poznawcze i fizyka powierzchni i stanów powierzchniowych nadal znajduje się w stadium intensywnego rozwoju.

Powierzchnie czyste i powierzchnie rzeczywiste

Pochodzenie (natura) stanów powierzchniowych oraz zjawiska nimi uwarunkowane wiążą się ściśle z charakterem powierzchni kryształu. Najprostszy w sensie fizycznym i chemicznym układ stanowi tzw. powierzchnia czysta, czyli zewnętrzna powierzchnia atomowa (jonowa) kryształu, będąca zakończeniem sieci krystalicznej (rys. 1). Najlepszą ilustracją takiej powierzchni jest powierzchnia kryształu rozłupanego w ultrawysokiej próżni. Istnienie w kryształach pewnych płaszczyzn sieciowych, tzw. płaszczyzn łupliwości, umożliwia odsłonięcie płaszczyzn zawie-

powierzchnia czysta

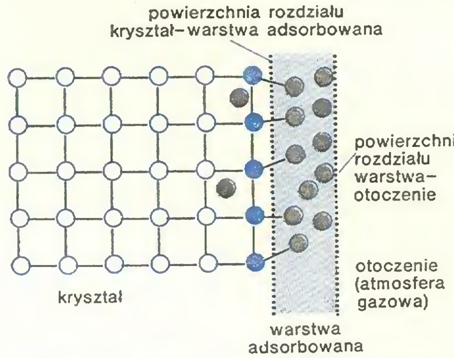


Rys. 1. Idealna powierzchnia czysta — schemat dwuwymiarowy. W zależności od powierzchni krystalograficznej atom powierzchniowy ma zerwane jedno wiązanie (np. powierzchnia 111 germanu) lub dwa wiązania (np. powierzchnia 100 germanu)

rających tylko jeden rodzaj atomów (jonów). Na czystość powierzchni odsłoniętych po przełupaniu wpływa wtedy tylko czystość całego kryształu, tzn. przypadkowe domieszki przedostające się do monokryształu w procesie hodowania. Z powierzchniami czystymi można mieć do czynienia jedynia w warunkach ultrawysokiej próżni ($p < 10^{-8}$ Pa) w ograniczonym czasie. W próżni rzędu 10^{-8} Pa już po paru godzinach występuje dostrzegalne pokrycie powierzchni adsorbowanymi molekułami, co zmienia charakter powierzchni. Powierzchnię czystą można również otrzymać usuwając warstwy absorbowane bombardowaniem jonami, elektronami, a w niektórych wypadkach również wygrzewaniem kryształu w ultrawysokiej próżni.

powierzchnie rzeczywiste

Powierzchnie, na których znajdują się warstwy adsorbowane, noszą nazwę powierzchni rzeczywistych. Warstwy te mogą być silnie związane z kryształem, tzw. warstwy chemisorbowane (np. tlen tworzący na powierzchni krzemu tlenki), lub warstwy słabo związane (np. cząsteczki wody), powstające wskutek fizycznej adsorpcji. Zwykle oba rodzaje warstw występują równocześnie. Powierzchnia rzeczywista obejmuje trzy elementy (rys. 2): powierzch-



Rys. 2. Powierzchnia rzeczywista — schemat dwuwymiarowy. W praktyce, z uwagi na proces dyfuzji w głąb kryształu, rozdział kryształ-warstwa adsorbowana nie musi być ostry

nię rozdziału kryształ–warstwa adsorbowana, warstwę adsorbowaną, zewnętrzną powierzchnię rozdziału warstwa–otoczenie.

wpływ obróbki powierzchni

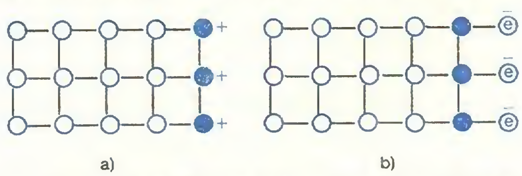
Przy powszechnie stosowanej obróbce kryształów obejmującej cięcie, polerowanie i trawienie chemiczne (lub elektrolityczne), które usuwa warstwę przypowierzchniową uszkodzoną przy obróbce mechanicznej, uzyskuje się powierzchnie rzeczywiste. Jeden z rodzajów powierzchni rzeczywistych stanowią powierzchnie celowo pokryte warstwami izolatora, przewodnika lub jednego i drugiego. Pewne ich rodzaje, tzw. struktury warstwowe (np. struktury MIS — metal–izolator–półprzewodnik) wykorzystywane są powszechnie jako elementy elektroniczne.

W dotychczas opanowanych zastosowaniach praktycznych ciał stałych mamy do czynienia jedynie z powierzchniami rzeczywistymi. Znaczenie badań powierzchni czystych określają względy poznawcze. Powierzchnie te stanowią najprostszy układ odniesienia, bez którego trudno określić jednoznacznie mechanizmy zjawisk zachodzących na dowolnej powierzchni.

Stany powierzchniowe na powierzchni czystej

Atomowa powierzchnia kryształu jest utworzona przez ostatnią płaszczyznę atomów (jonów) sieci krystalicznej (rys. 1). Atomy tworzące tę płaszczyznę są pozbawione części swoich sąsiadów niezbędnych do skompletowania wiązań sieci krystalicznej. W poglądowych modelach atomistycznych uważa się, że stany powierzchniowe pochodzą od zerwanych wiązań atomów powierzchni, które oddają elektron (stany donorowe, rys. 3a) lub też zostają uzupełnione przyłączeniem elektronu (stany akceptorowe, rys. 3b). W modelu takim liczba stanów powierzchniowych wiąże się ściśle z liczbą atomów na powierzchni i jest rzędu $10^{14}/cm^2$. Ze względu na to, że ruch ele-

stany donorowe i akceptorowe



Rys. 3. Stany powierzchniowe związane z zerwanymi wiązaniami: a) donorowe, b) akceptorowe. W wypadku (a) atom powierzchni oddaje swój elektron uczestniczący w wiązaniu, ładując się dodatnio; w wypadku (b) przyłącza elektron uzupełniający zerwane wiązanie

ktronów jest możliwy po powierzchni (przejście z jednego atomu na drugi), a niemożliwy prostopadle do niej, mówi się, że elektrony w stanach powierzchniowych są swobodne w kierunkach równoległych do powierzchni, a zlokalizowane prostopadle do niej.

modele atomistyczne powierzchni

Modele atomistyczne powierzchni, aczkolwiek nie dają ścisłego ilościowego opisu stanów powierzchniowych i procesów elektronowych z nimi związanych, stanowią podstawę do jakościowej interpretacji wielu zjawisk, w tym w szczególności fizykochemicznej aktywności powierzchni, zależności prędkości wzrostu kryształów od orientacji powierzchni, różnej odporności mechanicznej powierzchni itp.

Znacznie precyzyjniejszy opis stanów powierzchniowych na czystej powierzchni można uzyskać na gruncie teorii kwantowej. Jednak mimo już ponad czterdziestoletnich badań i niewątpliwych postępów ciągle jeszcze nie jest możliwy pełny ilościowy opis elektronowej struktury powierzchni rzeczywistego kryształu. Wielka złożoność zagadnienia zmusza do przyjmowania upraszczających założeń dość daleko odbiegających od sytuacji istniejącej w rzeczywistym krysztale. Podstawowym założeniem upraszczającym jest założenie, że symetria translacyjna zachowuje się aż do powierzchni. Przy tak postawionym zagadnieniu powierzchnia jest traktowana jako obszar przejścia od periodycznie zmieniającej się energii potencjalnej elektronu wewnątrz kryształu (→ Struktura elektronowa ciał stałych) do stałej wartości energii potencjalnej elektronu w otaczającej kryształ próżni. Obszar przejścia stanowi równocześnie barierę potencjału uniemożliwiającą odpływ elektronów z kryształu do próżni. Znając rozwiązanie równania Schrödingera w obu ośrodkach, tj. znając funkcje falowe elektronów w krysztale i w próżni, można określić strukturę elektronową na powierzchni.

struktura elektronowa na powierzchni

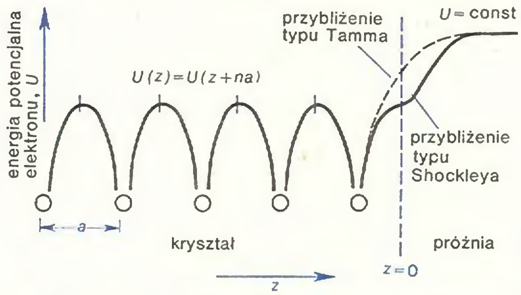
Wynikiem takiej analizy są dwa ogólne wnioski: 1) dla energii elektronu leżącej w obrębie dozwolonych pasm energetycznych powierzchnia nie wprowadza żadnych dodatkowych ograniczeń na stany energetyczne elektronu w krysztale; 2) w obszarze energii wzbronionych (w przerwie energetycznej) pojawiają się pewne poziomy energetyczne, zwane stanami powierzchniowymi, dozwolone dla elektronów. Obydwa te wnioski są niezwykle istotne. A mianowicie wniosek pierwszy oznacza, że struktura pasmowa cienkich kryształów (coraz powszechniej stosowanych do produkcji przyrządów półprzewodnikowych) zawiera wszystkie cechy struktury pasmowej nieskończonego kryształu; z wniosku zaś drugiego wynika, że obecność powierzchni (bariery potencjału na powierzchni) manifestuje się analogicznie jak obecność domieszek w krysztale, wprowadzając lokalne poziomy energetyczne w przerwie wzbronionej.

stany powierzchniowe

Przyjęty powszechnie dla poziomów energetycznych na powierzchni termin stany powierzchniowe wiąże się ściśle z charakterem funkcji falowych opisujących elektrony na powierzchni. Funkcje falowe tych elektronów zanikają wykładniczo w głąb kryształu. Oznacza to, że elektrony nie mogą przemieszczać się w kierunku prostopadłym do powierzchni, a zatem są na niej zlokalizowane. Omówione powyżej ogólne, jakościowe wnioski otrzymuje się bez względu na kształt bariery potencjalnej przy powierzchni — konieczne jest tylko jej istnienie. Chcąc podać pełniejszy, ilościowy opis stanów powierzchniowych, tzn. określić funkcje falowe elektronów w stanach powierzchniowych, położenie energetyczne (wartości energii), gęstość tych stanów, należy uwzględnić przebieg energii potencjalnej elektronów przy powierzchni. Rozwiązanie tego zagadnienia w sposób ścisły nie jest możliwe. Niemożność pełnego opisu teoretycznego stanów powierzchniowych wynika z: 1) nieznajomości rzeczywistego kształtu energii potencjalnej na powierzchni kryształów, 2) występowania przy powierzchni zaburzenia struktury krystalograficznej — przegrupowania atomów na płaszczyźnie sieciowej ograniczającej kryształ i na płaszczyznach z nią sąsiadujących, 3) trudności rachunkowych w określeniu funkcji falowych we wnętrzu kryształu odpowiadających energiom wzbronionym. Najtrudniejsze do pokonania wydają się problemy 1 i 2. Warunkiem ich rozwiązania jest uzyskanie danych doświadczalnych o strukturze powierzchni w mikro- i makroskali. Duże nadzieje wiąże się tutaj z dalszym rozwojem nowoczesnych metod badawczych opartych na dyfrakcji elektronów (np. metody LEED — dyfrakcji powolnych elektronów) oraz spektroskopii jonowej, tzw. spektroskopii elektronów Augera, wykorzystującej czułe na lokalną strukturę powierzchni zjawisko emisji elektronów z kryształu na skutek oddziaływania z jonami gazów szlachetnych.

Uzyskiwane z metod przybliżonych parametry stanów powierzchniowych w decydujący sposób zależą od przebiegu energii potencjalnej na powierzchni. Ze względów historycznych rozróżnia się dwa typy przybliżenia przebiegu energii potencjalnej, które posłużyły Tammowi (1932 r.) i Shockleyowi (1939 r.) do sformułowania pierwszych teorii stanów powierzchniowych. Tamm przyjął niesymetryczny przebieg energii potencjalnej w otoczeniu ostatniej warstwy atomowej, Shockley natomiast zakładał symetryczny przebieg potencjału aż do ostatniej warstwy atomowej (rys. 4). W rezultacie Tamm uzyskał jeden poziom energetyczny leżący pomiędzy każdymi dwoma pas-

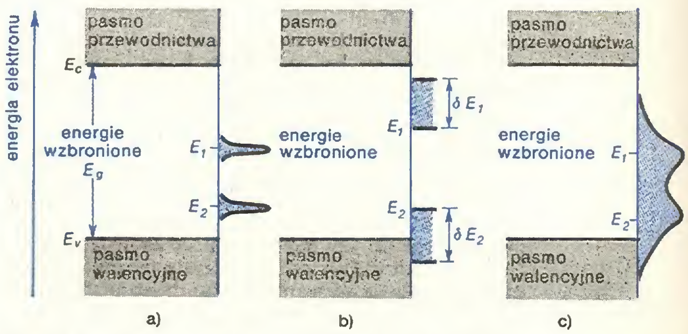
przybliżenie Tamma



Rys. 4. Przejście od periodycznie zmiennej energii potencjalnej elektronu U wewnątrz kryształu do stałej energii w próżni dla idealnego kryształu i idealnej powierzchni. Linia przerywana przedstawia przybliżenie Tamma z niesymetrycznym zakończeniem energii potencjalnej elektronu na powierzchni ($z = 0$). W przybliżeniu Shockleya energia elektronu zmienia się symetrycznie aż do powierzchni; a stała sieci krystalicznej

mami energii dozwolonych — są to tzw. stany powierzchniowe Tamma, podczas gdy Shockley uzyskał dwa poziomy energetyczne leżące w pobliżu środka przerwy energetycznej, tzw. stany powierzchniowe Shockleya. Jakkolwiek obydwa modele odbiegają znacznie od sytuacji w krysztale rzeczywistym, odegrały jednakże bardzo ważną rolę w rozwoju całej fizyki stanów powierzchniowych. Na podstawie dotychczasowych badań teoretycznych można dzisiaj przynajmniej w sposób jakościowy określić możliwe rodzaje

przybliżenie Shockleya



Rys. 5. Rodzaje stanów powierzchniowych na powierzchni czystej: a) poziomy dyskretne o położeniach energetycznych E_1 i E_2, b) pasma o szerokościach δE_1 i δE_2, c) stany o ciągłym rozkładzie gęstości w funkcji energii; E_c, E_v — energia odpowiadająca dolnej krawędzi pasma przewodnictwa i wierzchołkowi pasma walencyjnego

stanów powierzchniowych na powierzchni atomowo czystej. Rezultaty te przedstawione są na rys. 5. Tak więc możliwe są dyskretne poziomy energetyczne o położeniu energetycznym E_1 i E_2, wąskie pasma o krawędziach przy energii E_1 i E_2 (szerokościach energetycznych δE_1 i δE_2) oraz ciągły rozkład stanów powierzchniowych z maksimami gęstości stanów przy energiach E_1 i E_2.

Stany powierzchniowe na powierzchni rzeczywistej

Zagadnienie stanów powierzchniowych na powierzchni rzeczywistej jest znacznie bardziej skomplikowane niż na powierzchni atomowo czystej, tak że teoria stanów powierzchniowych w tym wypadku nie została dotychczas w sposób zadowalający sformułowana. Adsorbowana na zewnętrznej płaszczyźnie atomowej kryształu warstwa w sposób zasadniczy zmienia przebieg energii potencjalnej elektronu. W wyniku adsorbcji zerwane wiązania zostają wysycone (rys. 2), co powoduje zanik stanów powierzchniowych charakterystycznych dla powierzchni czystych. Pojawiają się natomiast inne stany powierzchniowe pochodzące od niedopasowania warstwy adsorbowanej z płaszczyzną atomową, domieszek gromadzących się na granicy rozdziału kryształ–warstwa, domieszek i defektów w warstwie oraz molekuł osiadających na granicy rozdziału warstwa adsorbowana–otoczenie. Parametry stanów powierzchniowych, takie jak położenie energetyczne, gęstość, przekrój czynny na wychwyt elektronu czy dziury itp., zależą od powierzchni krystalograficznej, obróbki kryształu oraz atmosfery gazowej otaczającej kryształ.

stany szybkie i powolne

Zazwyczaj stany powierzchniowe występujące na powierzchni rzeczywistej dzieli się na dwie grupy — tzw. stany szybkie i powolne. Podział ten dokonywany jest ze względu na czas wymiany ładunku pomiędzy stanami powierzchniowymi i wnętrzem kryształu. Dla stanów szybkich czas ten jest rzędu mikrosekund, dla powolnych zaś — milisekund lub dłuższy. Stany szybkie o gęstości nie przewyższającej 10^{12} na cm^2 zlokalizowane są na ogół na granicy rozdziału kryształ-warstwa adsorbowana; powolne — wewnątrz warstwy adsorbowanej i na granicy rozdziału warstwa –otoczenie (liczba tych stanów może osiągać 10^{13} na cm^2). Parametry stanów powierzchniowych uzyskuje się doświadczalnie analizując udział stanów powierzchniowych w zjawiskach fizycznych, takich jak przewodnictwo powierzchniowe, efekt polowy i fotonapięcie powierzchniowe.

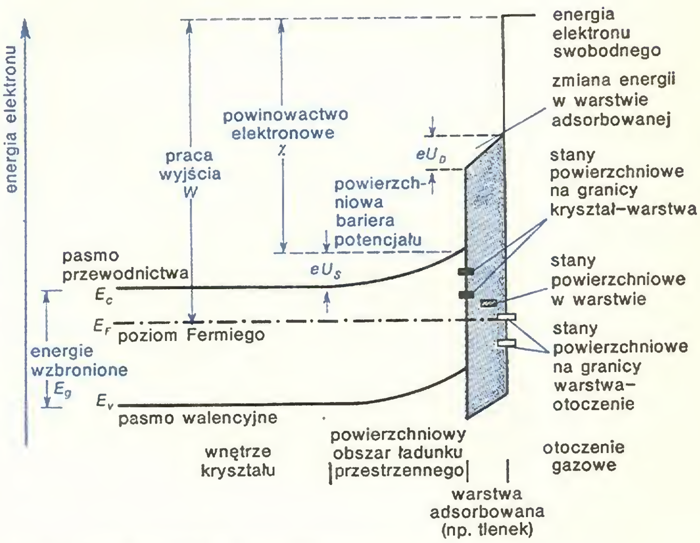
Udział stanów powierzchniowych w zjawiskach fizycznych

Obecność stanów powierzchniowych, ładunku elektrycznego w nich zgromadzonego oraz zmiana ładunku w stanach powierzchniowych przejawia się w wielu zjawiskach fizycznych zachodzących w krysztale. Wśród tych zjawisk istnieją takie, w których powierzchnia daje pewien przyczynek do obserwowanego w całym krysztale efektu, np. udział przewodnictwa powierzchniowego w całkowitym przewodnictwie kryształu, udział rekombinacji powierzchniowej w fotoprzewodnictwie itp., oraz takie, które nie istniałyby, gdyby nie było stanów powierzchniowych, np. fotonapięcie powierzchniowe. Doświadczalne badanie tych zjawisk jest źródłem informacji o parametrach stanów powierzchniowych, takich jak położenie energetyczne, gęstość, przekroje czynne na wychwyt elektronów i dziur itp. Przy wyjaśnieniu wpływu powierzchni na elektronowe zjawiska zachodzące w krysztale szczególnie ważna jest koncepcja przypowierzchniowego obszaru ładunku przestrzennego.

przewodnictwo i rekombinacja powierzchniowa

Przypowierzchniowy obszar ładunku przestrzennego

Koncepcję przypowierzchniowego obszaru ładunku wysunęli J. Bardeen, W. Shockley i W. Brattain. Stany powierzchniowe traktuje się jako zlokalizowane na powierzchni naładowane centra. Oznacza to, że powierzchnia jest elektrycznie naładowana. Aby kryształ pozostał elektrycznie obojętny, sumaryczny ładunek powierzchni musi być zrównoważony równym co do wartości ładunkiem przeciwnego znaku w sąsiadującym z powierzchnią wnętrzem kryształu. W ten sposób w krysztale mamy do czynienia z trzema



Rys. 6. Schemat energetyczny powierzchni rzeczywistej i przypowierzchniowego obszaru ładunku przestrzennego; prostokątami zaznaczono różne rodzaje stanów powierzchniowych. Podano również podstawowe wielkości charakteryzujące powierzchnię

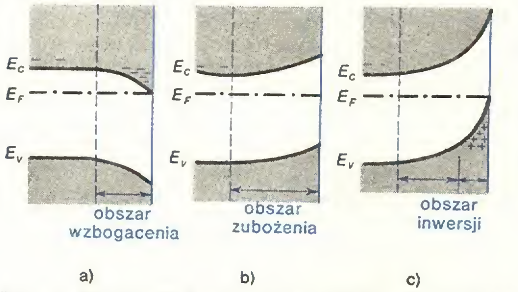
Przypowierzchniowy obszar ładunku przestrzennego w półprzewodnikach

Typ przewodnictwa elektrycznego we wnętrzu kryształu	Ładunek w stanach powierzchniowych	Charakterystyka	
		nazwa obszaru	typ przewodnictwa elektrycznego
Elektronowe	ujemny	obszar zubożenia (słabo przewodzący)	elektronowe, koncentracja elektronów mniejsza niż we wnętrzu
	ujemny o dużej wartości	obszar inwersji (przewodzący)	dziurowe
	dodatni	obszar wzbogacenia (silnie przewodzący)	elektronowe, koncentracja elektronów większa niż we wnętrzu
Dziurowe	ujemny	obszar wzbogacenia (silnie przewodzący)	dziurowe, koncentracja dziur większa niż we wnętrzu
	dodatni	obszar zubożenia (słabo przewodzący)	dziurowe, koncentracja dziur mniejsza niż we wnętrzu
	dodatni o dużej wartości	obszar inwersji (przewodzący)	elektronowe

elementami: naładowaną elektrycznie powierzchnią, przypowierzchniowym obszarem ładunku przestrzennego — neutralizującym ładunek powierzchni, oraz elektrycznie obojętnym wnętrzem kryształu (rys. 6).

obszary zubożenia, inwersji i wzbogacenia

W półprzewodnikach koncentracje elektronów (lub dziur) są stosunkowo niewielkie i przypowierzchniowy obszar ładunku przestrzennego rozciąga się na głębokość od 10^{-5} cm do 10^{-3} cm. W metalach jest on rzędu 10^{-7} cm i w większości zjawisk nie odgrywa znaczącej roli. Zależnie od ładunku w stanach powierzchniowych rozróżnia się trzy obszary ładunku powierzchniowego — tzw. obszar zubożenia, inwersji i wzbogacenia, scharakteryzowane w tabeli i przedstawione na rys. 7.



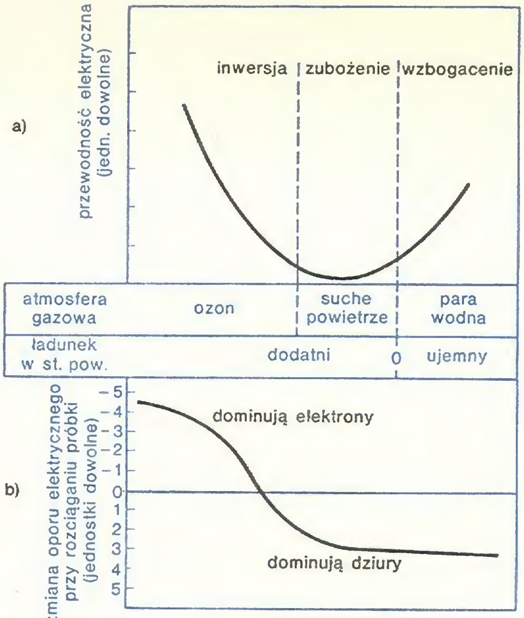
Rys. 7. Przypowierzchniowy obszar ładunku przestrzennego w półprzewodniku typu n (przewodnictwo elektronowe): a) obszar wzbogacenia — koncentracja elektronów większa niż we wnętrzu kryształu, b) obszar zubożenia — koncentracja elektronów mniejsza niż we wnętrzu kryształu, c) obszar inwersji — zmiana typu przewodnictwa; we wnętrzu — typ n, w POŁP — typ p; E_c, E_v energia odpowiadająca dolnej krawędzi pasma przewodnictwa i wierzchołkowi pasma walencyjnego, E_F poziom Fermiego. Obszar inwersji jest zawsze oddzielony od objętości obszarem zubożenia

pole elektryczne w obszarze przypowierzchniowym

Obszar przypowierzchniowy może mieć zupełnie inne własności fizyczne niż wnętrze kryształu. Pole elektryczne może w nim osiągnąć wartości rzędu 10^5 V/cm. Może to być obszar izolujący lub silnie przewodzący. Ruchliwość nośników prądu jest w nim z reguły znacznie niższa niż we wnętrzu kryształu (z uwagi na rozpraszanie przez powierzchnię). Inne są własności rekombinacyjne. W obszarach silnej inwersji i wzbogacenia istotnej modyfikacji ulega struktura pasmowa. Charakterystyczne dla wnętrza kryształu kwaziciągłe pasma dozwolonych energii ulegają kwantyzacji w kierunku prostopadłym do powierzchni, tzn. zachowując ciągłość w kierunkach równoległych do powierzchni rozpadają się na układ poziomów dyskretnych dla kierunku ruchu elektronu (dziury) prostopadłego do powierzchni.

Przewodnictwo powierzchniowe

Jedną z najbardziej rozpowszechnionych metod eksperymentalnych badania stanów powierzchniowych opartych na procesach zachodzących w przypowierzchniowym obszarze ładunku przestrzennego stanowi pomiar przewodnictwa elektrycznego cienkich płytek w funkcji atmosfery gazowej (rys. 8). W metodzie tej poddaje się analizie przewodność powierzchniową materiału zdefiniowaną jako $\sigma_s = {}^1/_2 d(\sigma - \sigma_{obj})$, gdzie σ — całkowita (mierzona) przewodność płytki, a σ_{obj} — przewodność objętościowa materiału, czyli — gdy płytka jest dostatecznie cienka — przewodność odpowiadająca zerowemu ładunkowi na powierzchni, d — grubość płytki; czynnik $^1/_2$ pochodzi z uwzględnienia dwóch powierzchni płytki. Pokazany na rys. 8a przebieg przewodności pozwala określić całkowity ładunek w stanach powierzchniowych odpowiadający danej atmosferze. Na krzywej można wyodrębnić silnie przewodzące obszary przypowierzchniowe: obszar inwersji, w którym dominuje przewodnictwo elektronowe, oraz obszar wzbogacenia o przewodnictwie dziurowym. Obszary te przedzielone są obsza-



Rys. 8. Zmiana przewodnictwa elektrycznego cienkich próbek Ge typu p (przewodnictwo dziurowe) przy zmianach atmosfery gazowej oraz zmiana ich oporu elektrycznego przy rozciąganiu

rem zubożenia przewodzącym słabiej niż wnętrze płytki.

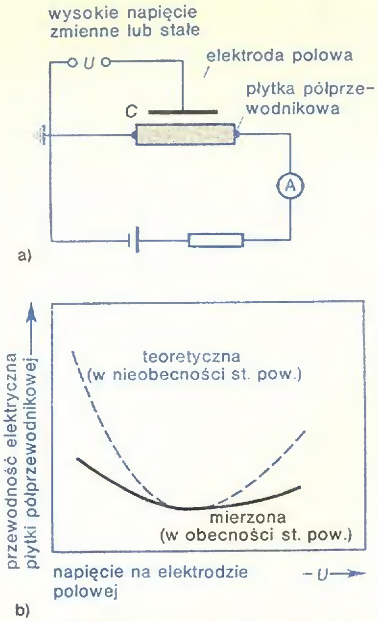
pomiar piezooporu

Informacji o przejściu pod działaniem atmosfery od dominacji elektronów w przypowierzchniowym obszarze ładunku do dominacji dziur dostarcza pomiar piezooporu (zmiany oporu wywołanej mechanicznym naprężeniem) cienkich płytek garmanowych lub krzemowych (rys. 8b). W germanie i krzemie elektrony wnoszą do piezooporu przyczynek przeciwnego znaku niż dziury, stąd też całkowity piezoopór cienkiej płytki zmienia znak przy zmianie atmosfery otaczającej płytkę sygnalizując przejście od dominacji elektronów do dominacji dziur.

Należy podkreślić, że przejście pod działaniem atmosfery przez wszystkie trzy rodzaje ładunku powierzchniowego możliwe jest jedynie w nielicznych wypadkach. W ogromnej większości materiałów półprzewodnikowych obszar ten jest obszarem zubożenia o małej przewodności, trudnej do wydzielenia z całkowitej przewodności cienkich płytek. Ponadto w przewodnictwie powierzchniowym, obok udziału swobodnych elektronów i dziur w przypowierzchniowym obszarze ładunku przestrzennego, możliwe jest również przewodnictwo po samych stanach powierzchniowych, którego nie można określić wskutek braku modeli teoretycznych stanów powierzchniowych (a tym samym i mechanizmu przewodnictwa po stanach powierzchniowych).

Efekt polowy

W określaniu położeń energetycznych i koncentracji stanów powierzchniowych na powierzchniach rzeczywistych, a niekiedy również na powierzchniach czystych, pomocny jest efekt polowy. Jest to efekt powstający, gdy do kryształu półprzewodnikowego przyłoży się zewnętrzne pole elektryczne w sposób pojemnościowy. Płaska elektroda przewodząca (najczęściej metalowa) tworzy z płytką półprzewodnikową kondensator płaski o pojemności C (rys. 9a). Przyłożone napięcie U indukuje na elektrodzie i w próbce ładunek przeciwnego znaku o wartości $Q_{ind} = CU$. Gdyby nie było stanów powierzchniowych, cały ładunek elektryczny indukowany w próbce przejawiałby się jako ładunek przewodzących prąd swobodnych



Rys. 9. Układ do pomiaru efektu polowego oraz zależność przewodności elektrycznej cienkiej płytki germanowej typu *p* od napięcia na elektrodzie polowej (wykres wykonany jest w funkcji $-U$, gdyż znak ładunku indukowanego w próbce jest przeciwny niż ładunku na elektrodzie polowej)

nośników (elektronów lub dziur) $Q_{przew} = Q_{ind}$. W rzeczywistości duża część indukowanego ładunku zostaje wychwycona i unieruchomiona w centrach na powierzchni próbki (czyli w stanach powierzchniowych) i $Q_{przew}/Q_{ind} < 1$; różnica $Q_{ind} - Q_{przew}$ jest ładunkiem wychwyconym przez stany powierzchniowe. O wartości Q_{przew} wnioskuje się ze zmian przewodności płytki w funkcji napięcia U przyłożonego do elektrody polowej (rys. 9b).

zastosowanie w tranzystorze polowym

Przebiegi czasowe zmian przewodności przy włączeniu i wyłączeniu napięcia na elektrodzie umożliwiają określenie parametrów dynamicznych stanów powierzchniowych (przekroju czynnego na wychwyt elektronów i dziur) oraz badanie mechanizmów wymiany ładunków między wnętrzem kryształu a stanami powierzchniowymi (np. zjawisk tunelowych na powierzchni). Efekt polowy ma także ogromne znaczenie praktyczne — został on wyzyskany w tranzystorze polowym.

powierzchniowy efekt piezoelektryczny

W półprzewodnikach o własnościach piezoelektrycznych (kryształy polarne o strukturze nie mającej środka symetrii) wykorzystuje się wersję efektu polowego, w której modulacja pola elektrycznego przy powierzchni odbywa się przez przyłożenie do kryształu jednoosiowych naprężeń mechanicznych (tzw. powierzchniowe efekty piezoelektryczne). Przyłożone naprężenia powodują powstawanie w krysztale polaryzacji elektrycznej. Różnica w stosunku do klasycznej wersji efektu polowego polega na wykorzystaniu pól wewnętrznych, a nie zewnętrznych, co umożliwia śledzenie w eksperymencie zmian pracy wyjścia.

praca wyjścia

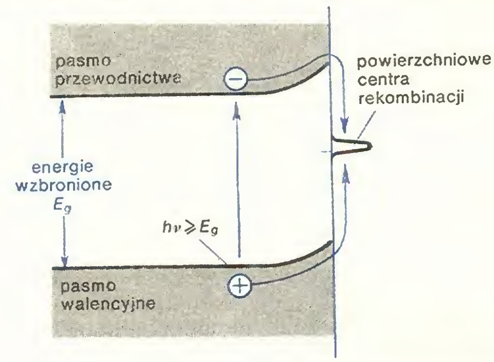
Praca wyjścia W, zdefiniowana jako różnica energii pomiędzy położeniem poziomu Fermiego w półprzewodniku a energią elektronu w próżni, jest sumą powinowactwa elektronowego χ, odległości energetycznej poziomu Fermiego E_F od dna pasma przewodnictwa E_c i energii eU_s wynikającej z istnienia powierzchniowej bariery potencjału (rys. 6). W ustalonych warunkach — przy określonej temperaturze i atmosferze otaczającej kryształ (χ i $E_F - E_c$ mają wówczas stałe wartości) zmiany pracy wyjścia spowodowane jednoosiowym naprężeniem pozwalają śledzić zmiany bariery powierzchniowej U_s, a więc wielkości bezpośrednio związanej z ładunkiem w stanach powierzchniowych.

Zjawiska fotoelektryczne wewnętrzne

Stany powierzchniowe odgrywają szczególnie istotną rolę w zjawiskach fotoelektrycznych wewnętrznych w ciałach stałych, np. w fotoprzewodnictwie i w powstawaniu fotonapięcia. Stanowią one bowiem dodatkowy kanał rekombinacyjny, a także biorą bezpośrednio udział w pochłanianiu promieniowania elektromagnetycznego padającego na kryształ. Wielkość obserwowanego efektu zależy od generacji, czyli ilości dodatkowo wytwarzanych swobodnych nośników — i od procesu odwrotnego, tzn. rekombinacji (anihilacji) — zaniku generowanych nośników. Parametrem opisującym rekombinację par elektron–dziura, a zatem i fotoefekt przy stałej generacji jest czas życia generowanych par. Czas ten dla określonego kryształu jest tym dłuższy, im mniej jest centrów rekombinacyjnych (poziomów lokalnych w przerwie wzbronionej). Stany powierzchniowe, które są dodatkowym, obok istniejących we wnętrzu kryształu, kanałem rekombinacyjnym zmniejszają czas życia generowanych nośników.

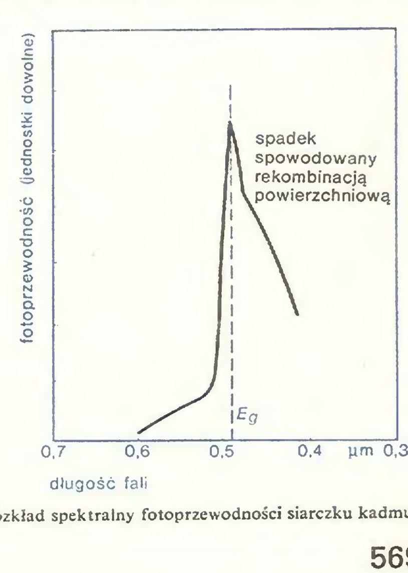
rekombinacja powierzchniowa

Rekombinacja w stanach powierzchniowych (rys. 10) wpływa na zależność czasu życia od grubości kryształu bądź od głębokości wnikania światła. Jeśli światło jest słabo absorbowane



Rys. 10. Przykład działania stanów powierzchniowych jako centrów rekombinacji. Pary elektron–dziura, generowane światłem o energii $h\nu$ większej od przerwy energetycznej E_g, anihilują szybciej w obecności stanów powierzchniowych

(energia padających fotonów odpowiada podstawowej krawędzi absorpcji), generacja par jest jednorodna w całej objętości kryształu i w bardzo grubych płytkach rekombinacja powierzchniowa nieznacznie zmienia czas życia. Natomiast w cienkich kryształach (rzędu kilkunastu μm) procesy generacji i rekombinacji zachodzą przy powierzchni i na czas życia duży wpływ ma rekombinacja powierzchniowa. Tak więc



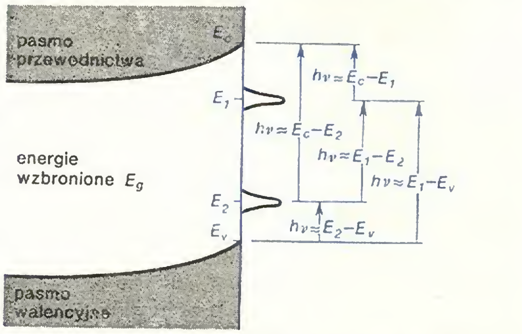
Rys. 11. Rozkład spektralny fotoprzewodności siarczku kadmu

gdy płytki są cienkie — przy generacji jednorodnej obserwuje się silną zależność czasu życia od grubości płytki. Jeśli współczynniki absorpcji są duże, światło pochłaniane jest w obszarze przypowierzchniowym (w przybliżeniu w warstwie o grubości równej odwrotności współczynnika absorpcji), a zatem grubość płytki jest mało istotnym parametrem. Wynika to stąd, że grubość kryształu jest zawsze co najmniej kilka razy większa od głębokości wnikania światła. Przy generacji przypowierzchniowej szczególnie uprzywilejowana jest więc rekombinacja przez stany powierzchniowe i prawie nie zależy od grubości kryształu a jedynie od stanu powierzchni (ilości centrów rekombinacyjnych na powierzchni). Przejawem silnej rekombinacji powierzchniowej jest występowanie maksimum w rozkładzie spektralnym fotonapięcia, fotoprzewodności itp. Przy energii fotonów większej od przerwy energetycznej, w obszarze wzrastającego nadal współczynnika absorpcji, zamiast spodziewanego wzrostu np. fotoprzewodności obserwuje się jej zmniejszenie, spowodowane rekombinacją powierzchniową (rys. 11).

ujemny wpływ rekombinacji powierzchniowej

Rekombinacja powierzchniowa jest zjawiskiem ujemnie wpływającym na pracę wszystkich fotoelementów, a zwłaszcza fotoogniw i złącz laserowych. Na przykład przy niewłaściwej obróbce powierzchni gęstość stanów powierzchniowych może tak wzrosnąć, że wskutek rekombinacji powierzchniowej wielkość fotonapięcia będzie znacznie mniejsza od spodziewanej, wynikającej z własności kryształu i wielkości generacji.

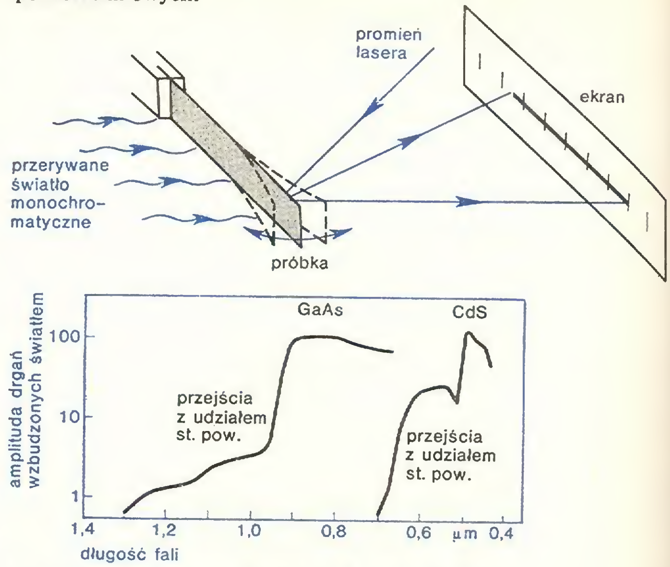
Udział stanów powierzchniowych w pochłanianiu światła padającego na kryształ związany jest z indukowanymi światłem przejściami elektronowymi przedstawionymi na rys. 12. Jest to pochłanianie nieznaczne, bezpośrednio mierzalne jedynie przy wykorzystaniu wielokrotnego odbicia światła od powierzchni. W sposób pośredni można je wykryć dzięki towarzyszącym tym przejściom zmianom koncentracji nośników ładunku w próbce, modyfikacji procesów rekombinacji lub zmianie obsadzenia stanów powierzchniowych. Stosowane w praktyce metody sprowadzają się w zasadzie do pomiarów fotonapięcia w funkcji energii fotonów w obszarze $h\nu < E_g$.



Rys. 12. Schematyczne przedstawienie przejść elektronowych do i ze stanów powierzchniowych przy absorpcji fotonów o odpowiedniej energii $h\nu$

spektroskopia fotonapięcia powierzchniowego

W najnowszej z tych metod, tzw. spektroskopii fotonapięcia powierzchniowego, wykorzystuje się zjawisko zmiany potencjału powierzchniowego (zmiany spadku napięcia na przypowierzchniowym obszarze ładunku przestrzennego — rys. 6) przy oświetleniu, wynikające ze zmiany ładunku elektrostatycznego w stanach powierzchniowych pod wpływem przejść indukowanych światłem. W kryształach o właściwościach piezoelektrycznych, możliwa jest do zrealizowania niezwykle prosta wersja spektroskopii fotonapięcia powierzchniowego, polegająca na analizie wzbudzonych światłem drgań mechanicznych cienkich płytek (rys. 13). Wywołana światłem zmiana ładunku w stanach powierzchniowych wywołuje zmianę pola elektrycznego w przypowierzchniowym obszarze ładunku przestrzennego, co prowadzi do mechanicznego odkształcenia cienkich płytek. Pomiar amplitudy drgań w funkcji energii fotonów pozwala m.in. na bezpośrednie określenie położenia energii stanów powierzchniowych.

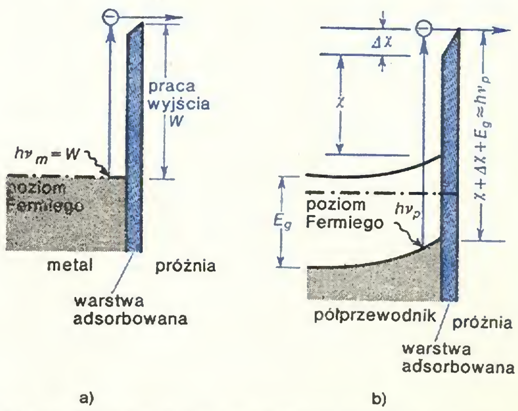


Rys. 13. Wzbudzanie drgań cienkich płytek piezoelektrycznych przerywanym strumieniem światła monochromatycznego i rejestrowanie ich jako odchylenie odbitego od kryształu promienia lasera oraz typowe zależności amplitudy drgań kryształów arsenku galu i siarczku kadmu (grubości 20 μm) od długości fali światła pobudzającego drgania

Fotoemisja (zjawisko fotoelektryczne zewnętrzne)

Jeśli energia fotonów padających na powierzchnię jest dostatecznie duża, obserwuje się efekt wybijania elektronów z materiału do otoczenia (zjawisko fotoemisji). W przypadku metali próg fotoemisji (minimalna energia fotonów, przy której obserwuje się efekt) pozwala określić pracę wyjścia, czyli różnicę pomiędzy energią elektronu w próżni i energią poziomu Fermiego w materiale. W przypadku izolatorów i niezwyrodniałych półprzewodników, poziom Fermiego leży w przerwie energetycznej, stąd też obecność elektronów na poziomie Fermiego może wynikać jedynie z istnienia poziomów lokalnych i ich koncentracja jest niewielka. Próg fotoemisji pojawia się zatem nie przy energii fotonów równej pracy wyjścia, lecz przy energii $h\nu_p = \chi + \Delta\chi + E_g$ (rys. 14), odpowiadającej wybijaniu elektronów z pasma wa-

próg fotoemisji



Rys. 14. Emisja elektronów przy absorpcji fotonów: a) z metalu, b) z półprzewodnika; $h\nu_m$ i $h\nu_p$ oznaczają energie progowe odpowiednio metalu i półprzewodnika

lencyjnego. Porównanie wielkości $h\nu_p$ i pracy wyjścia (mierzonej innymi metodami) pozwala wnioskować jak jest położony poziom Fermiego na powierzchni kryształu. Położenie poziomu Fermiego zdeterminowane jest ładunkiem elektrycznym w stanach powierzchniowych. Badania fotoemisji w ultrawysokiej próżni, połączone z oczyszczaniem powierzchni, pozwalają wnioskować o skoku energii potencjalnej, $\Delta\chi$, na warstwie adsorbowanej. W ostatnim okresie szczególne zainteresowanie wzbudzają precyzyjne pomiary fotoemisji wywołanej fotonami o energiach poniżej energii progowych. Elektrony mogą być wybijane do otoczenia ze stanów powierzchniowych zlokalizowanych w przerwie energetycznej.

Fononowe stany powierzchniowe

Obok omówionych wyżej elektronowych stanów powierzchniowych wzrasta zainteresowanie drganiami sieci krystalicznej (→ Dynamika sieci krystalicznej) o amplitudzie zanikającej wykładniczo przy przejściu od powierzchni w głąb kryształu — tzw. fononowymi stanami powierzchniowymi. Inaczej, są to fonony zlokalizowane na powierzchni kryształu. Podobnie jak elektronowe stany powierzchniowe, występują na powierzchniach rzeczywistych i czystych i ich natura wiąże się zarówno ze strukturą powierzchni, jak też z adsorbowanymi na powierzchni molekułami. Fonony powierzchniowe występują zarówno jako drgania normalne akustyczne, jak też optyczne, a ich energie leżą zazwyczaj w obszarze energii wzbronionych we wnętrzu kryształu (poniżej energii objętościowych fononów optycznych i powyżej maksymalnej energii objętościowych fononów akustycznych). Obok roli w określaniu np. termicznych własności kryształów lub procesów rozproszeniowych, badania fononów powierzchniowych są uzupełnieniem do badań elektronowych stanów powierzchniowych — obie problematyki opierają się na tych samych modelach struktury powierzchni i oddziaływania z otaczającą atmosferą.

S. G. Davison, J. D. Levine *Surface States*, New York 1970 (ros. Moskwa 1973); D. R. Frankl *Electrical Properties of Semiconductor Surfaces*, Oxford 1967; A. Many, Y. Goldstein, N. B. Grover *Semiconductor Surfaces*, Amsterdam 1971; C. G. Scott, C. E. Reed, *Surface Physics of Phosphors and Semiconductors*, London 1975.

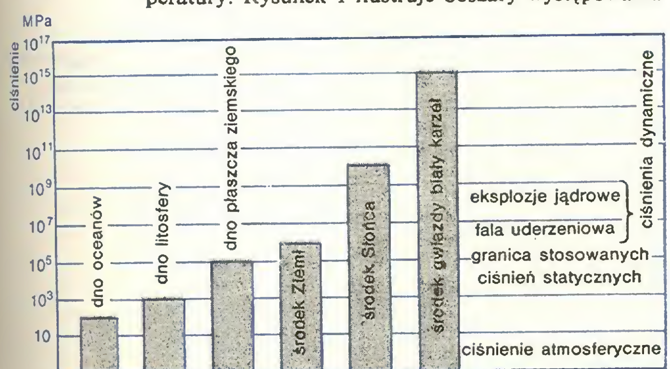
Wysokie ciśnienia

Tadeusz Suski

Układ poddany działaniu ciśnienia zmienia się pod różnymi względami. Gaz lub para może zamieniać się w ciecz, ciecz — w ciało stałe lub odwrotnie, struktura ciała stałego może ulec przeobrażeniu w wyniku zmiany układu atomów. Ciśnienie może wpływać na szybkość i sposób krystalizacji, stopień rozpuszczalności jednej substancji w drugiej, może też wpływać na zmianę szybkości zachodzenia wielu reakcji chemicznych.

wysokie ciśnienia w przyrodzie

Geologia i geofizyka bada rezultaty działania ciśnienia w procesie tworzenia się skał i minerałów na Ziemi i we Wszechświecie. Opisuje zjawiska zachodzące we wnętrzu Ziemi czy innych planet. Zasadnicze znaczenie dla poznania tych zjawisk ma możliwość odtworzenia w laboratorium warunków, które decydują o zachodzeniu w przyrodzie badanych procesów. Najważniejszymi parametrami określającymi te warunki są wysokie ciśnienia i towarzyszące im wysokie temperatury. Rysunek 1 ilustruje obszary występowania wysokich ciśnień w przyrodzie. Dla porównania przedstawiono górne granice ciśnień osiąganych w laboratoriach.



Rys. 1. Występowanie wysokich ciśnień w przyrodzie

Inżynieria materiałowa koncentruje się na wykorzystaniu wysokich ciśnień do otrzymywania nowych materiałów. Specjalnie prowadzone procesy technologiczne umożliwiają uzyskanie substancji o bardzo korzystnych własnościach, np. o wyjątkowej wytrzymałości mechanicznej czy też odporności termicznej. Wysokie ciśnienia umożliwiają również szersze stosowanie technologii obróbki plastycznej materiałów, co prowadzi do znacznej poprawy własności uzyskiwanych elementów.

Metody otrzymywania wysokich ciśnień

Do otrzymywania wysokich ciśnień w laboratoriach stosuje się aparaturę różnych typów i konstrukcji. Różnice w ich budowie zależą od wielkości uzyskiwanych ciśnień, zakresu temperatur, objętości komory, aktywności chemicznej stosowanych materiałów. Specjalne warunki eksperymentalne, jak na przykład konieczność prowadzenia badań elektrycznych, optycznych, magnetycznych czy rentgenowskich, narzucają dodatkowe wymagania co do ich rozwiązań konstrukcyjnych.

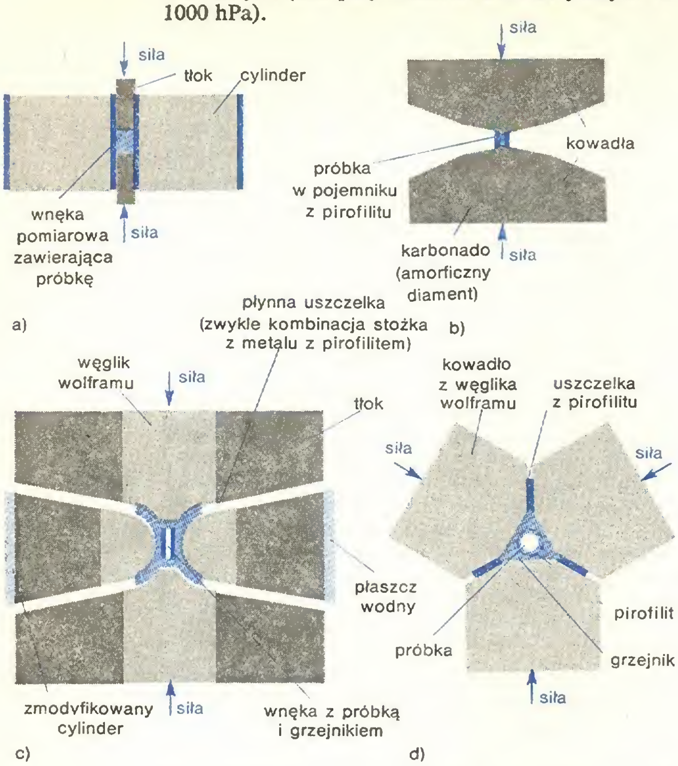
W klasyfikacji aparatury wysokociśnieniowej często stosuje się podział wg rodzaju ośrodka przekazującego ciśnienia i odpowiednio dzieli się je wtedy na aparaturę gazową, cieczową oraz aparaturę, w której ośrodkiem przekazującym ciśnienie jest ciało stałe. Doświadczenia wymagające wysokiej precyzji pomiaru prowadzone są na ogół w ciśnieniowych urządzeniach gazowych. Natomiast tam, gdzie głównym celem jest uzyskanie jak najwyższych ciśnień, stosuje się aparaturę, w której ośrodkiem przekazującym ciśnienie jest ciało stałe. Przy doborze najodpowiedniejszego urządzenia ciśnieniowego konieczny jest na ogół kompromis pomiędzy wygodą w prowadzeniu badań a trudnościami wytworzenia wysokiego ciśnienia.

aparatura gazowa

Największą jednorodność ciśnienia gwarantują aparatury gazowe. Sprężanie gazów utrudnia jednak ich duża ściśliwość oraz konieczność uszczelnienia aparatury. Trudności te zmniejszają się oczywiście przy zastosowaniu urządzeń cieczowych oraz urządzeń z ciałem stałym. Ze względu na dużą ściśliwość gazów sprężanie ich odbywa się w kilku etapach w cylindrach z ruchomymi tłokami (rys. 2a; il. 59, tabl. 16). Sprężony gaz przekazuje się elastyczną kapilarą do specjalnie przystosowanej dla danego rodzaju pomiarów komory. Wygląd jednej z takich komór przedstawia il. 57 na tabl. 16.

Maksymalne ciśnienie w aparaturach gazowych nie przekracza na ogół 1500 MPa. W tym zakresie ciśnień metody badawcze są tak rozwinięte, że możliwe jest

przeprowadzenie doświadczeń tak jak w warunkach normalnych (tzn. przy ciśnieniu atmosferycznym ok. 1000 hPa).



Rys. 2. Podstawowe typy urządzeń do wytwarzania wysokich ciśnień statycznych: a) układ tłok–cylinder, b) kowadło Bridgmanna (otaczający próbkę pirofilit jest ośrodkiem przekazującym ciśnienie i stanowi izolację termiczną między kowadłami a tłokiem i wewnętrznie grzaną próbką), c) urządzenie wielokowadłowe Halla, d) urządzenia typu „belt"

aparatura cieczowa

Aparaturę cieczową można stosować w znacznie mniejszym zakresie temperatur niż aparaturę gazową. W niskich temperaturach ciecze zestalają się już przy bardzo małych ciśnieniach. W wysokich temperaturach ciecz staje się chemicznie niestabilna. Utrudnione jest również prowadzenie badań optycznych, gdyż ciecze są znacznie mniej przezroczyste niż gazy. Urządzenia cieczowe są natomiast znacznie prostsze w konstrukcji i łatwiejsze w użyciu niż gazowe. Sprężanie odbywa się tu na ogół również w cylindrze z ruchomym tłokiem, jednak ze względu na mniejszą ściśliwość wystarcza jeden etap sprężania. Ponadto w aparaturze cieczowej trudności z zapewnieniem szczelności są znacznie mniejsze niż w aparaturze gazowej. Za pomocą komór cieczowych można uzyskać ciśnienie 3 a nawet 5 GPa.

aparatura z ciałem stałym

Aparatura ciśnieniowa z ośrodkiem stałym, nazywana też często aparaturą kwazihydrostatyczną, umożliwia uzyskanie najwyższych ciśnień (rys. 2b, c, d). Wytworzone ciśnienia są jednak niejednorodne, a bardzo mała objętość komory roboczej ogranicza możliwości eksperymentalne do obserwacji jedynie dużych, jakościowych efektów najczęściej związanych z przemianami fazowymi. W urządzeniach takich jak przedstawione na rys. 2b możliwe jest obecnie uzyskanie ciśnień do 100 GPa.

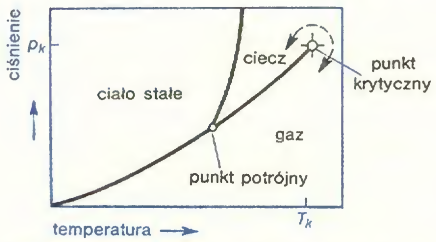
Do uzyskania największych ciśnień stosuje się technikę ciśnień dynamicznych. Polega ona na wykorzystaniu fali uderzeniowej powstającej w wyniku eksplozji (→ Fale uderzeniowe), co pozwala uzyskiwać ciśnienia rzędu 10^4 GPa. Wadą ograniczającą możliwości stosowania tej techniki jest trudność kontrolowania warunków doświadczenia, ponieważ ciśnienie działa w okresach mikrosekundowych. Ponadto fala uderzeniowa powoduje wzrost temperatury do kilku tysięcy stopni.

Procesy zachodzące przy wysokim ciśnieniu

reguła Le Chateliera

Interpretację zachowania się substancji poddanych działaniu wysokich ciśnień ułatwia reguła Le Chateliera. Mówi ona, że układ w stanie równowagi poddany działaniu sił zewnętrznych dąży do zminimalizowania efektu działania przyłożonych sił. Dlatego też ciśnienie dąży do zahamowania wszelkich procesów prowadzących do wzrostu objętości układu i faworyzuje reakcje, których wynikiem jest zmniejszenie tej objętości i, wiążący się z tym, wzrost gęstości.

Regułę Le Chateliera można zilustrować rozważając równanie stanu $f(p, \varrho, T) = 0$, wiążące termodynamiczne parametry stanu: ciśnienie p, gęstość ϱ i temperaturę T. Równanie stanu można przedstawić jako powierzchnię w przestrzeni trójwymiarowej o współrzędnych p, ϱ, T. Każdy z punktów tej powierzchni odpowiada stanowi, w którym sąsiadujące fazy znajdują się w równowadze. Na wykresie (rys. 3) można wyróżnić trzy oddzielne obszary odpowiadające trzem stanom skupienia substancji: gazowi, cieczy i ciału stałemu. Faza gazowa i stała znajdują się w równowadze wzdłuż krzywej sublimacji: faza stała i ciekła — wzdłuż krzywej topnienia (krzepnięcia), a gazowa i ciekła — wzdłuż krzywej wrzenia (kondensacji). Każdy



Rys. 3. Równanie stanu $f(p, \rho, T) = 0$. Rzut powierzchni p, ρ, T na płaszczyznę pT

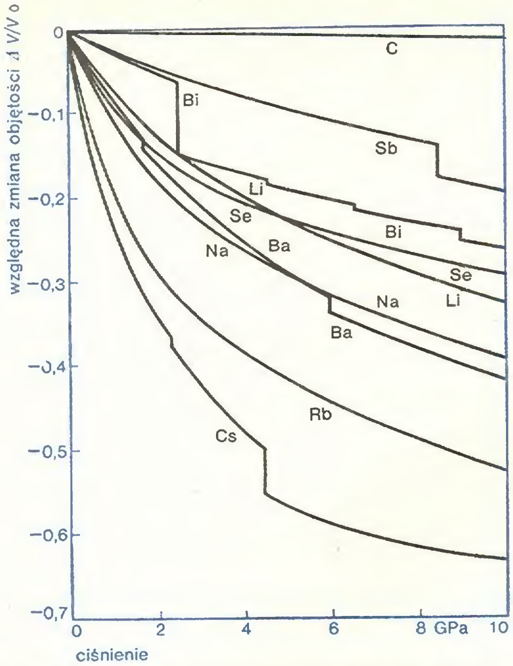
punkt tych trzech krzywych odpowiada stanowi (określonemu przez ciśnienie i temperaturę), w którym mogą istnieć dwie fazy, a punkt potrójny jest charakterystyczny jako jedyny punkt współistnienia wszystkich trzech faz. Należy zauważyć, że krzywa wrzenia nie istnieje dla ciśnień i temperatur wyższych niż krytyczne (p_k, T_k). Oznacza to, że można przeprowadzić ciecz w gaz w sposób ciągły, nie przekraczając linii przejścia fazowego, zwiększając ciśnienie i temperaturę powyżej wartości krytycznych.

wpływ ciśnienia na objętość

W gazie, przy temperaturze powyżej krytycznej, wzrost ciśnienia prowadzi do zmniejszenia odległości międzyatomowych lub międzycząsteczkowych. Przy ciśnieniach rzędu 10^4 MPa objętość zmniejsza się tysiąckrotnie w stosunku do objętości w warunkach normalnych. Poniżej punktu krytycznego, zgodnie z równaniem stanu konkretnej substancji, może nastąpić przemiana w fazę ciekłą. Wzrost ciśnienia do około 5 GPa powoduje 20–50% zmniejszenie objętości cieczy. Ciśnienie podwyższa temperaturę topnienia substancji, które krzepną ze zwiększeniem gęstości fazy stałej (rys. 3). W wypadku substancji takich jak woda, mających fazę stałą o mniejszej gęstości niż ciekła, temperatura krzepnięcia ulega obniżeniu ze wzrostem ciśnienia. Na przykład temperatura krzepnięcia wody przy 200 MPa wynosi −20°C. Powyżej 200 MPa woda krystalizuje w innych fazach stałych o większej gęstości niż ciecz, a temperatura krzepnięcia wzrasta ze wzrostem ciśnienia. Przy 4500 MPa krzepnięcie zachodzi w temperaturze 190°C.

Ciała stałe są mniej ściśliwe niż ciecze, przy czym ściśliwość zarówno jednych jak i drugich maleje ze wzrostem ciśnienia. Oprócz ciągłych zmian objętości obserwuje się również zmiany skokowe (rys. 4) związane z przemianami fazowymi. Nie istnieje ogólna reguła pozwalająca przewidzieć dla danej substancji liczbę możliwych faz i rodzaj przemian fazowych w warunkach działania wysokiego ciśnienia.

Dopiero wszechstronne określenie mikroskopowej budowy substancji (symetria rozkładu atomów, ich rodzaj, charakter i wielkość oddziaływań między nimi) pomaga zrozumieć jej zachowanie się w warunkach działania ciśnienia.



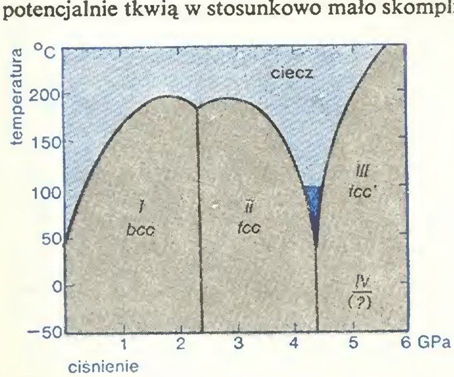
Rys. 4. Zależność objętości różnych ciał stałych od ciśnienia

Z punktu widzenia atomowej struktury materii każda substancja stanowi układ elektronów i rdzeni atomowych. Liczba rdzeni atomowych w molu substancji jest rzędu liczby Avogadra, tzn. 10^{23}. W wielu zagadnieniach fizyki ciała stałego rozważa się niezależnie zachowanie się elektronów i sieci krystalicznej. W innych wypadkach konieczne jest wzięcie pod uwagę oddziaływań między tymi dwoma układami. Istnieją przemiany fazowe, w których zachodzą zmiany w strukturze układu elektronów przy zachowaniu symetrii rozkładu atomów, tzw. izotypia. Mechanizm przemian fazowych ze zmianą struktury krystalicznej polegający na przesunięciach atomów, czyli zmianie nie tylko ich odległości ale i symetrii, prowadzi do tzw. przemian polimorficznych. Mogą być one zapowiadane przez charakterystyczne zmiany w strukturze energetycznej drgań cieplnych atomów (fononów). Opis i interpretacja wielu przemian fazowych wymaga jednak dokładnego zbadania obu podukładów (elektrony, rdzenie atomowe) wraz z ich wzajemnymi oddziaływaniami.

przemiany polimorficzne

Na przykładzie cezu poddanego działaniu wysokiego ciśnienia pokażemy możliwości modyfikacji, które potencjalnie tkwią w stosunkowo mało skomplikowanym układzie atomów. Cez w wysokich ciśnieniach wykazuje zarówno polimorficzną przemianę fazową jak i izostrukturalną (bez zmiany symetrii rozkładu atomów) przemianę elektronową, poza tym staje się nadprzewodnikiem. Badania metodami dyfrakcji nautronów i promieni rentgenowskich prowadzone w temperaturze pokojowej wykazały trzy wartości ciśnienia, przy których występuje skokowa zmiana objętości tego metalu. Pierwsza osobliwość występuje przy 2,37 GPa (rys. 5). Przy tym ciśnieniu następuje polimorficzna przemiana fazowa. Towarzyszy jej zmiana objętości o około 0,6%; stosunek objętości komórki elementarnej cezu po przemianie do objętości tej komórki przy ciśnieniu atmosferycznym wynosi $V/V_0 = 0{,}63$. Przed przemianą fazową atomy tworzą sieć przestrzennie centrowaną (*bcc*), po przemianie — płasko centrowaną (*fcc*).

cez pod wysokim ciśnieniem



Rys. 5. Wykres fazowy cezu

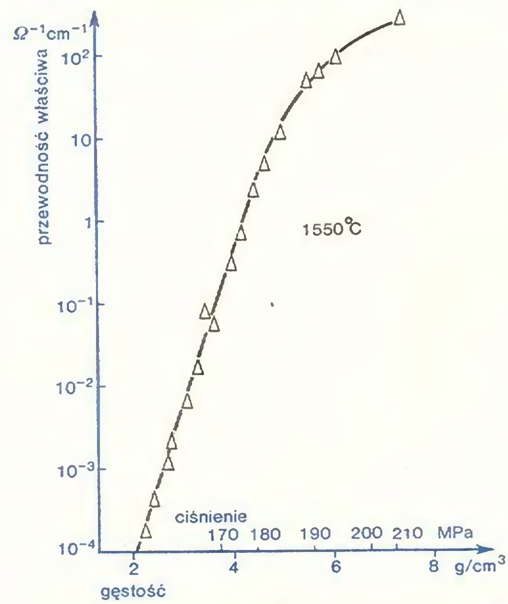
Następna nieciągłość występuje przy 4,22 GPa; skok objętości wynosi ok. 9% ($V/V_0 = 0{,}45$). Można przyjąć, że jest to przemiana izostrukturalna pierwszego rodzaju. Struktura nowej fazy jest również płasko centrowana. W rezultacie działania ciśnienia zmniejszeniu ulegają atomy cezu. Uważa się, że przemiana ta polega na zmianie stanu elektronowego cezu z $6s \rightarrow 5d$. Elektron z powłoki o większym promieniu zostaje „wciśnięty" w powłokę o mniejszym promieniu.

Trzecia przemiana Cs III do Cs IV zachodzi przy 4,27 GPa; objętość maleje o 2,4% ($V/V_0 = 0{,}41$). Struktura fazy Cs IV nie jest znana. Powyższe zmiany są także wykazywane w pomiarach oporu elektrycznego w funkcji ciśnienia. Skokową zmianę właściwości cezu obserwuje się ponadto przy ok. 12 GPa. Obniżenie temperatury do 1,5 K przy tym ciśnieniu powoduje pojawienie się fazy nadprzewodzącej cezu.

Elektrony i ciśnienie

Wskutek zbliżania się do siebie atomów tworzących pierwotnie układ nie oddziałujących ze sobą cząstek, z dyskretnych poziomów energetycznych elektronów formują się pasma energetyczne (→ Struktura elektronowa ciał stałych). Szerokość przerwy pomiędzy pasmami, szerokość samego pasma oraz nakładanie się pasm zależą od charakteru oddziaływań między atomami kryształu. Działanie ciśnienia modyfikuje te oddziaływania i prowadzi do obniżenia wielkości obszarów energii zabronionej między pasmami (np.

wpływ ciśnienia na przerwę energetyczną

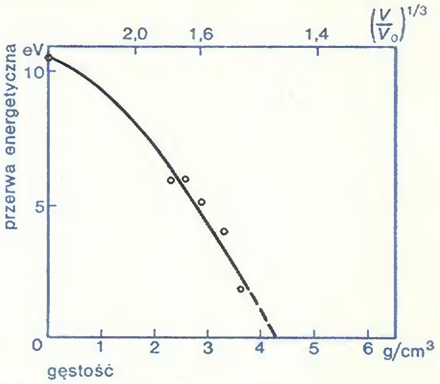


Rys. 6. Przewodność właściwa par rtęci w funkcji ciśnienia i gęstości

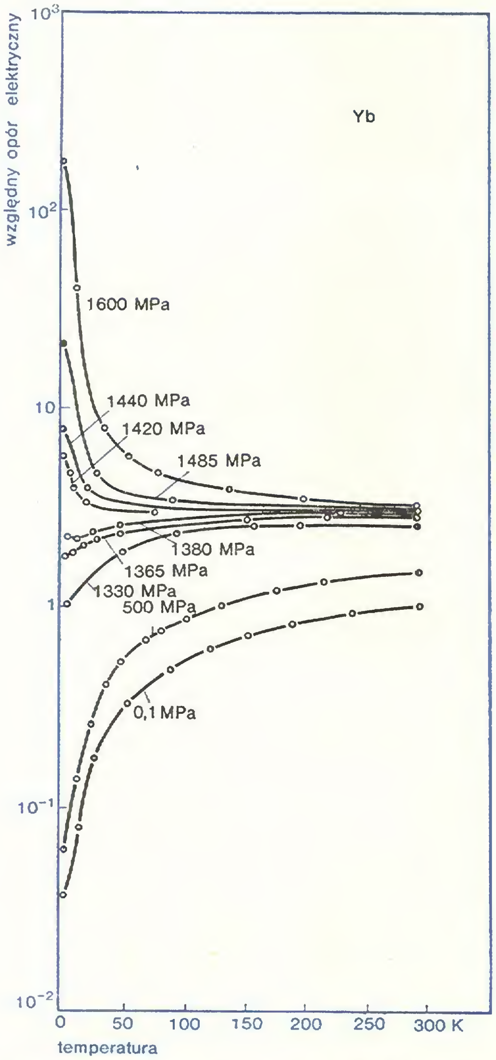
między pasmem walencyjnym i przewodnictwa w półprzewodnikach) albo wzrostu wielkości nakładania się tych pasm, a więc do wzrostu własności metalicznych badanej substancji.

pomiar przewodności elektrycznej

Zachowanie się większości substancji potwierdza ten ogólny wniosek. Obserwacja przewodnictwa elektrycznego par metali poddanych działaniu ciśnienia



Rys. 7. Zależność przerwy energetycznej rtęci od gęstości; V_0 objętość rtęci w 0°C i przy ciśnieniu 100 MPa



Rys. 8. Zależność oporu elektrycznego iterbu od temperatury, przy różnych ciśnieniach. Przy ciśnieniu około 1400 MPa widać zmianę charakteru funkcji $R = f(T)$ z zależności odpowiadającej metalom na zależność odpowiadającą półprzewodnikom

wykazuje takie tendencje. Wzrost gęstości par w warunkach powyżej temperatury krytycznej (nie ma możliwości utworzenia się fazy ciekłej w wyniku kondensacji par) prowadzi do powstania struktury pasmowej, pojawienia się przerwy energetycznej, a następnie jej zamknięcia. W rezultacie powoduje to ciągłe przejście od własności dielektrycznych przez półprzewodnikowe do metalicznych. Przeprowadzono wiele doświadczeń obejmujących pomiary przewodności elektrycznej, siły termoelektrycznej, efektu Halla oraz absorpcji optycznej ciekłej i gazowej rtęci. Zmiana gęstości par rtęci od 2 do 6 g/cm³ przy ciśnieniu nie przekraczającym 200 MPa odpowiadała zmianie przewodności elektrycznej o sześć rzędów. Rys. 6 przedstawia zależność przewodności par rtęci w funkcji ciśnienia i gęstości w temperaturze 1550°C, a rys. 7 ilustruje zależność wielkości przerwy energetycznej od zmiany gęstości spowodowanej działaniem wysokiego ciśnienia.

wzrost charakteru metalicznego

Przy bardzo wysokich ciśnieniach obowiązuje ogólna zasada stwierdzająca, że działaniu ciśnienia towarzyszy wzrost charakteru metalicznego substancji. Przy ciśnieniach pośrednich obserwuje się jednak również zmianę własności w kierunku przeciwnym. Ciśnienie wywołuje przemiany metal–półprzewodnik czy metal–izolator. W ten sposób zachowuje się np. iterb (rys. 8), wapń, stront i bizmut. Przy określonych wartościach ciśnienia i temperatury w metalach tych „otwiera się" przerwa energetyczna. Warto również wspomnieć o zachowaniu się kryształów półprzewodnikowych. Na przykład w germanie pod wpływem działania ciśnienia przerwa energetyczna początkowo rośnie, następnie przechodzi przez maksimum i od ciśnień rzędu 4000 MPa ulega zmniejszeniu. Podobne wyniki uzyskiwano przy badaniach antymonku galu GaSb. Natomiast zachowanie się fosforku galu GaP potwierdza ogólną tendencję wzrostu właściwości metalicznych wraz z ciśnieniem, i to w pełnym zakresie przykładanego ciśnienia.

Sieć krystaliczna i ciśnienie

Zmiana odległości między tworzącymi ciało stałe atomami powoduje zmianę ich wzajemnych oddziaływań, a w konsekwencji zmianę częstości ich drgań termicznych (ω_i). Gdy substancja poddana jest równomiernemu ściskaniu, jej objętość ulega zmniejszeniu, atomy drgają wokół położeń o większym udziale sił odpychających (→ Dynamika sieci krystalicznej). Wskutek tego powinna znacznie wzrosnąć częstość ω_i. Potwierdzają ten wniosek wyniki pomiarów przy wysokich ciśnieniach. Wzrost ω_i obserwowano w substancjach z wiązaniami jonowymi, kowalencyjnymi oraz w kryształach molekularnych. Odbiciem tych zmian mikroskopowych jest wzrost współczynników sprężystości ciał stałych przy wzroście ciśnienia.

wpływ ciśnienia na sprężystość

Obserwowane w niektórych kryształach obniżenie częstości jednego czy kilku typów drgań termicznych atomów wraz ze wzrostem ciśnienia można traktować jako zapowiedź utraty stabilności sieci krystalicznej. W rezultacie następuje polimorficzna przemiana fazowa; zmienia się symetria rozkładu atomów oraz ich wzajemne odległości. Sytuacja taka powstaje w wielu kryształach ferroelektrycznych, na przykład w $PbTiO_3$, SbSi, a także w kryształach półprzewodnikowych, na przykład CdS.

Występowanie większej lub mniejszej ściśliwości oraz polimorficznych przemian fazowych różnorodnych substancji związane jest z możliwościami zwiększenia gęstości upakowania atomów czy cząsteczek w strukturze substancji. Przypomnieć tu można przemianę fazową w cezie przy 2,37 GPa, podczas której przebudowuje się struktura *bcc* w bardziej gęstą *fcc*. Rodzaj wiązań chemicznych decyduje o istnieniu takiej a nie innej struktury krystalicznej. Struktura kryształów molekularnych (np. kryształy gazów szlachetnych o słabych wiązaniach międzycząsteczkowych

modyfikacja struktur krystalicznych

i małej gęstości) ulega łatwiej zmianom pod wpływem ciśnienia. Dla lodu, który jest również kryształem molekularnym, określono osiem modyfikacji krystalicznych, istniejących w warunkach różnych ciśnień i temperatur. Kryształy z wiązaniami metalicznymi są na ogół mniej ściśliwe od kryształów molekularnych. Podobnie jest dla struktur łańcuchowych (parafina), czy warstwowych (grafit). Struktury te wykazują znaczne zróżnicowanie ściśliwości w różnych kierunkach krystalograficznych. Silne wiązania kowalencyjne ograniczają możliwość zwiększenia upakowania pod wpływem ciśnienia — substancje z tymi wiązaniami należą do najmniej ściśliwych, np. diament jest ponad 200 razy mniej ściśliwy niż cez.

Nadprzewodnictwo

Głównym mikroskopowym mechanizmem nadprzewodnictwa jest oddziaływanie przyciągające między elektronami, realizowane za pośrednictwem sieci krystalicznej (oddziaływanie elektron–fonon). Im większe jest to oddziaływanie, tym wyższa temperatura T_k przejścia do stanu nadprzewodzącego (→ Nadprzewodnictwo). Ciśnienie — przez modyfikacje struktur krystalicznych lub własności elektronowych, a także oddziaływania elektron–fonon — wywiera bardzo istotny wpływ na zjawisko nadprzewodnictwa.

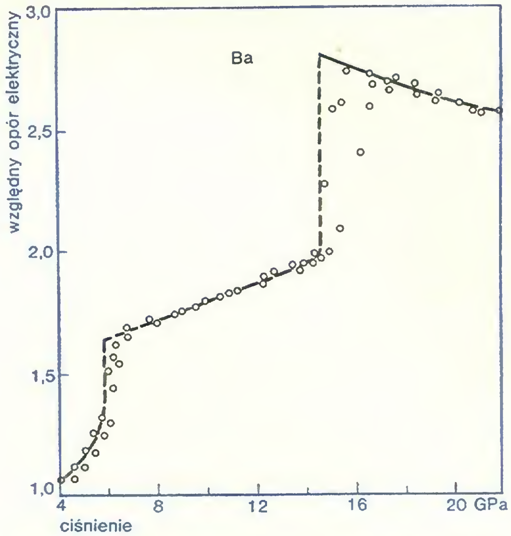
Ogromna liczba metali, związków metalicznych, a także związków i pierwiastków o charakterze półprzewodnikowym w pewnych warunkach może występować w fazie nadprzewodzącej. Często warunkiem koniecznym pojawienia się nadprzewodnictwa jest działanie ciśnienia hydrostatycznego (rys. 9). Indukowane ciśnieniem nadprzewodniki grupują się głównie

metale indukowane ciśnieniem

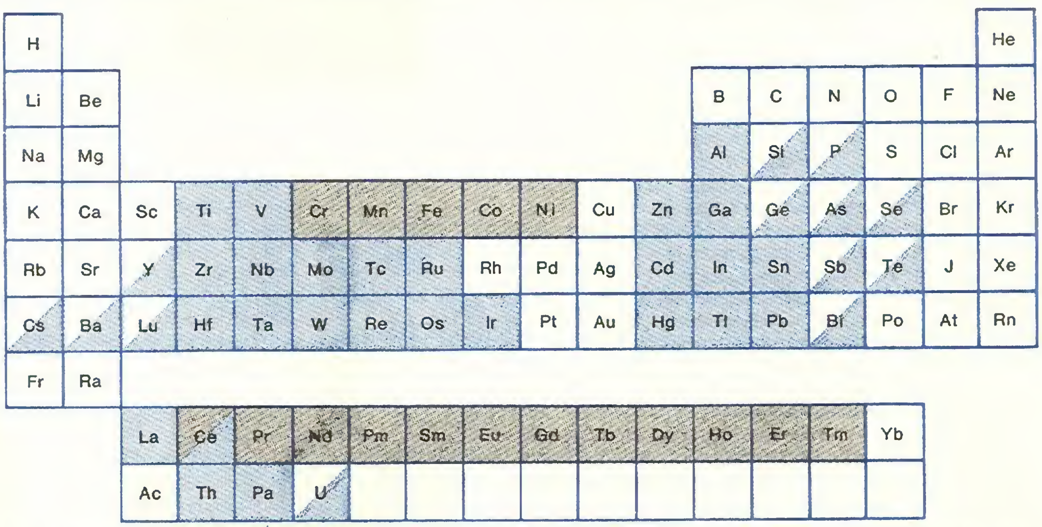
w dwóch miejscach układu okresowego. Pierwiastki z prawej strony układu Mendelejewa przy podwyższaniu ciśnienia wykazują na ogół wzrost własności metalicznych. Te indukowane ciśnieniem metale, np. Ge i Si, przy wysokich ciśnieniach są bez wyjątku nadprzewodnikami, podobnie jak inne metale sąsiadujące w układzie okresowym (Zn, In, Sn), w których elektronami przewodnictwa są elektrony *s* i *p*. Z lewej strony sytuacja jest bardziej skomplikowana. Indukowane ciśnieniem nadprzewodniki bar i cez mają właściwości metali przejściowych, a więc i potencjalne możliwości by stać się nadprzewodnikami. Mają one niezlokalizowane elektrony *s*, a pod wpływem ciśnienia pojawia się dodatkowy udział silnie zlokalizowanych elektronów *d* i *f*.

W wielu substancjach poddanych wysokiemu ciśnieniu pojawiają się dwie lub więcej nadprzewodzące fazy wysokociśnieniowe. W barze przy 5,5 GPa pojawia się faza nadprzewodząca (temperatura przejścia $T_k \approx 1{,}3$ K), druga przemiana fazowa zachodzi przy ciśnieniu około 14 GPa z $T_k \approx 5$ K (rys. 10). Faza nadprzewodząca pojawia się w cerze przy $p \approx 5$ GPa. Cer jest obecnie jedynym znanym pierwiastkiem wy-

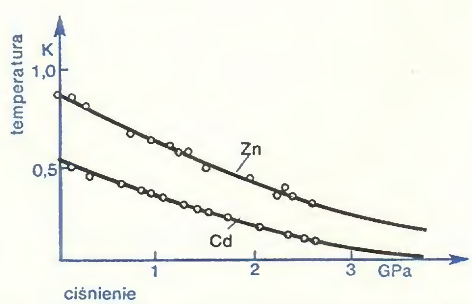
fazy nadprzewodzące



Rys. 10. Zależność względnego oporu elektrycznego baru od ciśnienia. Przy ciśnieniu 5,5 GPa i ciśnieniu 14 GPa zachodzą przemiany fazowe. Powstałe fazy mają własności nadprzewodzące



Rys. 9. Układ okresowy pierwiastków: kolor niebieski oznacza pierwiastki nadprzewodzące; kolor szary — pierwiastki magnetyczne; kratki w połowie niebieskie odpowiadają pierwiastkom nadprzewodzącym w wysokich ciśnieniach



Rys. 11. Zależność temperatury przejścia nadprzewodzącego cynku i kadmu od ciśnienia. Zmniejszenie temperatury przy wzroście ciśnienia jest wynikiem obniżenia oddziaływania elektron–fonon

kazującym jednocześnie uporządkowanie magnetyczne i nadprzewodzące (rys. 9).

wzrost lub obniżenie temperatury przejścia

Działanie ciśnienia może prowadzić zarówno do wzrostu temperatury przejścia nadprzewodzącego, jak również jej obniżenia. Tak na przykład obniżenie T_k i zanik nadprzewodnictwa dla cynku i kadmu (rys. 11) w warunkach działania ciśnienia jest wynikiem obniżenia wielkości oddziaływania elektron–fonon. Pomiary przeprowadzone w szerokim zakresie ciśnień wskazują, że T_k wielu metali maleje liniowo ze wzrostem ciśnienia ($dT_k/dp < 0$). Taki przebieg jest spowodowany głównie wzrostem częstości drgań termicznych atomów wraz ze wzrostem ciśnienia.

Zastosowanie wysokich ciśnień w inżynierii materiałowej

Uważa się, że każdy materiał istniejący w przyrodzie może być wytworzony w warunkach laboratoryjnych. Diament, wysokociśnieniowa modyfikacja węgla, jest w temperaturze pokojowej i przy ciśnieniach poniżej 1200 MPa fazą metastabilną termodynamicznie. Struktura tej metastabilnej fazy utrzymuje się nieskończenie długo. Dopiero poddanie diamentu działaniu temperatury powyżej 1000°C, przy niskich ciśnieniach, prowadzi do przemiany diament–grafit. Dla grafitu ciśnienie 20 GPa w temperaturze pokojowej nie jest wystarczające do zmiany w diament. Aby zaszła taka przemiana, konieczne jest jednoczesne działanie wysokiej temperatury i ciśnienia.

syntetyczny diament

Od wielu lat podejmowano różne próby otrzymania syntetycznego diamentu, m.in. metodą podgrzewania związków węgla do wysokich temperatur przy jednoczesnym stosowaniu wysokich ciśnień. Dokonano tego po raz pierwszy w Stanach Zjednoczonych (General Electric, 1955 r.). Sukces osiągnięto dzięki zastosowaniu ciśnienia rzędu 10 GPa (znacznie wyższych niż dotychczas stosowane) jednocześnie z wysokimi temperaturami (ok. 1000°C). Wykorzystano dodatkowo działanie katalityczne stopionych metali (np. tantalu). Katalizatory ułatwiają zrywanie silnych wiązań międzyatomowych wewnątrz warstw tworzących grafit. Katalizator spełnia funkcję rozpuszczalnika dla węgla. Z tak wytworzonego roztworu wolny węgiel krystalizuje w postaci diamentu. Następuje to w wyniku obniżenia stopnia rozpuszczalności diamentu w stosunku do rozpuszczalności grafitu w tych samych warunkach, tzn. w warunkach działania wysokiej temperatury i ciśnienia. Należy dodać, że uzyskiwano diament w procesie przemiany grafit–diament również bez użycia katalizatorów. Proces taki wymaga jednak stosowania ciśnień powyżej 12 GPa i temperatur w granicach 2500–3000°C.

Poszukiwania nowych materiałów doprowadziły do uzyskania wielu ciekawych substancji, wytworzonych w warunkach wysokich i ultrawysokich ciśnień, stabilnych przy ciśnieniu atmosferycznym. Udało się na przykład „zagęścić" kwarc (SiO_2). Dwie jego fazy — koezyt i stiszowit — są substancjami gęstszymi odpowiednio o 10% i 6%. Wytworzenie stiszowitu wymaga stosowania ciśnień rzędu 10 GPa przy 600°C.

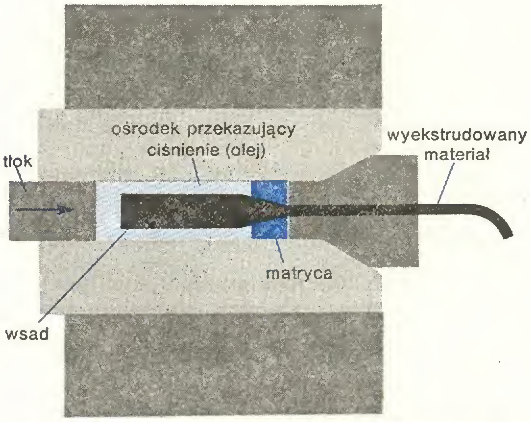
koezyt i stiszowit

Obie laboratoryjnie wytworzone formy SiO_2 zostały wykryte w kraterach meteorytów. Koezyt został także rozpoznany jako produkt podziemnej eksplozji jądrowej. Fakty te dają wyobrażenie o możliwościach laboratoriów wysokociśnieniowych. Można je porównywać z tym co istnieje w warunkach naturalnych, w procesach zderzeń gigantycznych meteorytów z powierzchnią Ziemi, czy podczas bardzo silnych wybuchów nuklearnych.

Stosując wysokie ciśnienia w procesie syntezy różnych materiałów uzyskuje się materiały o własnościach szczególnie cennych ze względu na ich zastosowanie. W wielu wypadkach są to materiały nie mające odpowiedników w przyrodzie. Człowiek sterując procesami technologicznymi ma większe możliwości zmian — i to niezależnych — temperatury i ciśnienia. W warunkach naturalnych procesy wzrostu i obniżania temperatury są ściśle związane z odpowiednimi wzrostami i redukcją ciśnień. Przykładem materiału syntetycznego nie mającego odpowiednika naturalnego jest regularna odmiana azotku boru BN (borazon). Związek ten krystalizuje w procesie podobnym do stosowanego przy wytwarzaniu diamentu i tworzy materiał o strukturze regularnej. Odznacza się ogromną odpornością termiczną i dużą twardością.

hydroekstruzja

Wysokie ciśnienia znalazły wiele ważnych zastosowań w dziedzinie obróbki plastycznej materiałów zwanej hydrostatycznym wytłaczaniem lub hydroekstruzją. W metodzie tej wykorzystuje się zjawisko podwyższenia plastyczności przy wysokim ciśnieniu. Materiał wyjściowy wraz z ośrodkiem przenoszącym ciśnienie (np. olej) zawarty jest w grubościennym cylindrze. Cylinder ten z jednej strony zamyka matryca z otworem, z drugiej ruchomy tłok. Gdy wskutek przesuwania tłoka wzrośnie wystarczająco ciśnienie



Rys. 12. Zasada procesu hydrostatycznego wytłaczania (hydroekstruzji)



Rys. 13. Elementy uzyskiwane w procesie hydroekstruzji (produkty firmy Asea)

ośrodka ciśnieniowego, materiał (sprężany hydrostatycznie) przechodzi przez otwór w matrycy zmieniając swój kształt i zmieniając przekrój (rys. 12). Proces hydroekstruzji ma wiele zalet w porównaniu z innymi technikami obróbki plastycznej. Oto niektóre z nich: proces ten odbywa się w znacznie niższych temperaturach; własności mechaniczne uzyskiwanego materiału wykazują dużą jednorodność w całym przekroju. Ciśnienie hydrostatyczne zwiększa ciągliwość niektórych twardych i kruchych materiałów do takiego stopnia, że ich deformacja następuje znacznie łatwiej (np. Be, V, Zr); różnorodność kształtów matrycy pozwala uzyskiwać prawie dowolne kształty ekstrudowanych elementów (rys. 13), przy czym uzyskuje się dużą dokładność wymiarów produktu. Istnieje możliwość powlekania jednych materiałów innymi, należy jedynie stosować odpowiednio przygotowane wsady.

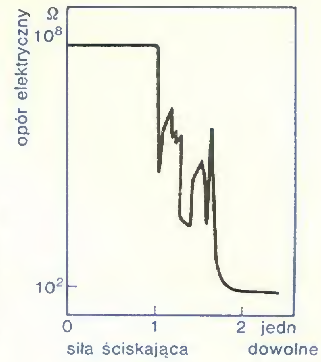
Metaliczny wodór?

Jeżeli się spojrzy na pierwszą kolumnę układu okresowego, zwraca uwagę odrębność wodoru w stosunku do reszty pierwiastków z tej kolumny. Są one metalami, a tylko wodór ma własności dielektryczne (jest izolatorem). Odrębność ta od dawna inspirowała wiele prac teoretycznych. Wielu fizyków przewidywało, że działanie odpowiednio wysokiego ciśnienia winno przeprowadzać dielektryczny wodór w metal.

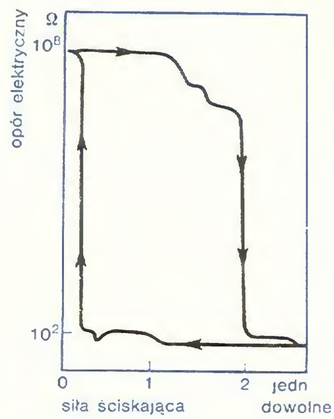
Ogromne nakłady finansowe na badania dotyczące tego zagadnienia związane są głównie z nadzieją rozbudzoną przez przewidywania naukowców co do własności nadprzewodzących metalicznego wodoru. Wyjątkowo cenną własnością ma być bardzo wysoka temperatura przejścia nadprzewodzącego, pomiędzy 200–300 K. Możliwość zachowania stanu nadprzewodzącego nawet w 200 K (−73°C) otwierałaby zupełnie nowe perspektywy przed energetyką. Przewiduje się również, że metaliczny wodór mógłby pełnić funkcję wysokokalorycznego paliwa nie wprowadzającego zanieczyszczeń.

własności nadprzewodzące

Obiecujące rezultaty prowadzące do otrzymania metalicznego wodoru uzyskano w Instytucie Fizyki Wysokich Ciśnień Radzieckiej Akademii Nauk. Naukowcy z tego Instytutu wykorzystali gigantyczną prasę hydrauliczną współdziałającą z kowadłami diamentowymi ochłodzonymi do 4,2 K. Jedno z nich jest płaskie, drugie ma kształt bardzo płaskiego stożka. Takie rozwiązanie służy uzyskaniu ciśnień rzędu 100 GPa w centralnej części kowadeł. Przez obszar ten przepuszczono bardzo czysty gazowy wodór, doprowadzając do jego zestalenia w postaci cienkiej warstwy. Podczas eksperymentu mierzono prąd elektryczny płynący w obwodzie, w którym znajdowały się kowadła i warstwa zestalonego wodoru. Pojawienie się prądu w tym obwodzie powinno świadczyć o przekroczeniu ciśnienia krytycznego. W opisywanym eksperymencie opór elektryczny zmalał o sześć rzędów (od 10^8 do $10^2\,\Omega$) przy różnych ciśnieniach leżących między 100 i 300 GPa (rys. 14). Eksperymentatorzy uważają, że ten skok odpowiada przemianie dielektrycznego wodoru w wodór metaliczny. Następnie zaczęto się zastanawiać nad możliwościami stabilizacji fazy



Rys. 14. Zależność między oporem elektrycznym i siłą działającą na kowadła w procesie otrzymywania metalicznego wodoru ($T = 4{,}2$ K). Zależność odpowiada wzrostowi ciśnienia



Rys. 15. Histereza zależności oporu elektrycznego od działającej na kowadła siły w procesie otrzymywania metalicznego wodoru ($T = 4{,}2$ K). Strzałki wskazują kierunek zmian ciśnienia

metalicznej wodoru, jego istnieniem w warunkach normalnych oraz możliwościami otrzymania większej ilości tej substancji. Znanych jest wiele substancji, które istnieją w fazie metastabilnej po usunięciu działania ciśnienia (syntetyczny diament). Uczeni radzieccy uważają, że bardzo obiecujące z tego punktu widzenia jest występowanie histerezy ciśnieniowej. Otóż wodór nie powraca do fazy dielektrycznej, gdy redukuje się ciśnienie. Pozostaje metaliczny nawet w chwili, gdy ciśnienie wynosi około połowy wartości krytycznej (rys. 15). Szuka się więc sposobów powiększenia tej histerezy do wartości gwarantującej stabilność uzyskanego nadprzewodnika w warunkach zbliżonych do normalnych.

histereza ciśnieniowa

C. C. Bradley *High Pressure Methods in Solid State Research*, New York 1969; R. S. Bradley (ed.) *Advances in High Pressure Research*, London 1966; P. W. Bridgman *The Physics of High Pressure*, London 1952; H. G. Drickamer, C. W. Frank *Electronic Transition and the High Pressure Chemistry and Physics of Solids*, London 1973; S. W. Popowa, N.A. Bendeliani *Wysokije dawlenija*, Moskwa 1974.

MAGNETYZM

Teoria magnetyzmu

Jerzy Mielnicki i Bogusław Mrygoń

Teoria magnetyzmu obejmuje teorię właściwości magnetycznych izolowanych cząstek elementarnych, atomów i molekuł. a także zbiorów atomów lub molekuł w postaci gazu, cieczy lub ciała stałego. Właściwości magnetyczne zbioru atomów wynikają z właściwości magnetycznych izolowanych atomów oraz z charakteru oddziaływań między pojedynczymi atomami. Natura tych oddziaływań ma decydujące znaczenie w kształtowaniu się określonych właściwości magnetycznych ciał złożonych z wielu elementarnych

nośników magnetyzmu. W tym sensie ferromagnetyzm, który występuje w niektórych kryształach, jest zjawiskiem kolektywnym, ponieważ warunkiem koniecznym pojawienia się ferromagnetyzmu w ciele stałym są odpowiednie oddziaływania między elektronami. Jednolity opis właściwości magnetycznych elementarnych nośników magnetyzmu oraz zbiorów tych nośników z uwzględnieniem oddziaływań możliwy jest jedynie w ujęciu teorii kwantowej. Podamy tu zarys kwantowej teorii magnetyzmu powłoki elektronowej atomu oraz układu atomowych momentów magnetycznych w ciałach stałych o budowie krystalicznej.

Magnetyzm atomowy

Wiele podstawowych właściwości fizycznych ciał stałych wyjaśnić można znając zachowanie się elektronów na zewnętrznych powłokach atomów izolowanych. Każdy elektron ma własny moment pędu określony spinową liczbą kwantową s równą $^1/_2$. Dla spinu $s = {}^1/_2$ jedyne dwa możliwe rzuty wektora własnego momentu pędu na dowolnie wybrany kierunek zewnętrznego pola magnetycznego mają tę samą wartość i określone są przez magnetyczną spinową liczbę kwantową m_s:

kwantowa liczba magnetyczna

$$s^{[z]} = m_s\hbar;$$

m_s może przyjmować tylko dwie wartości $\pm{}^1/_2$; $\hbar = 1{,}0545887 \cdot 10^{-34}$ J · s — stała Plancka podzielona przez 2π. Ze spinem elektronu związany jest moment magnetyczny, którego dwa możliwe rzuty na kierunek zewnętrznego pola magnetycznego są równe

moment magnetyczny elektronu

$$\mu_s^{[z]} = 2m_s\mu_B. \tag{1}$$

Wielkość $\mu_B = e\hbar/2mc$, gdzie e oznacza ładunek elektronu, m — jego masę spoczynkową, c — prędkość światła, nazywana jest magnetonem Bohra i odgrywa rolę jednostki momentu magnetycznego w skali atomowej. Wartości bezwzględne wektorów spinowego momentu pędu i momentu magnetycznego wyrażają się wzorami:

$$|\vec{P_s}| = \sqrt{s(s+1)}\hbar, \tag{2}$$

$$|\vec{\mu_s}| = 2\sqrt{s(s+1)}\,\mu_B. \tag{3}$$

Określając właściwości magnetyczne powłoki elektronowej atomu, oprócz momentów magnetycznych związanych ze spinem elektronów należy uwzględnić również momenty magnetyczne związane z ruchem orbitalnym elektronów. Pewną rolę odgrywa również magnetyzm jądra atomowego. Momenty magnetyczne nukleonów są jednak około 1000 razy mniejsze od momentów magnetycznych elektronów. Przykładem efektów związanych z momentem magnetycznym jądra może być nadsubtelna struktura widm atomowych (→ Spektroskopia atomowa) bądź jądrowy rezonans magnetyczny.

Wielkość wektora momentu pędu związanego z orbitalnym ruchem elektronu wyraża się wzorem:

$$|\vec{P_l}| = \sqrt{l(l+1)}\hbar. \tag{4}$$

Orbitalna liczba kwantowa l przyjmuje wartości

$$l = 0, 1, 2, \ldots, (n-1),$$

gdzie n oznacza główną liczbę kwantową. Z orbitalnym momentem pędu związany jest moment magnetyczny

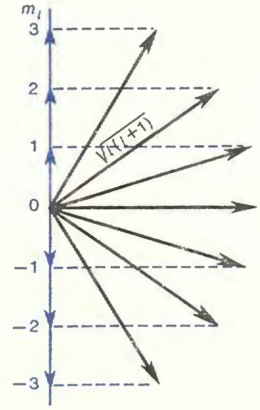
orbitalny moment magnetyczny

$$|\vec{\mu_l}| = \sqrt{l(l+1)}\;\mu_B \tag{5}$$

Z mechaniki kwantowej wynika bez dodatkowych założeń, że rzuty wektora orbitalnego momentu pędu na kierunek zewnętrznego pola mogą przyjmować tylko ściśle określone wartości. Możliwe rzuty orbitalnego momentu magnetycznego określone są przez zespół magnetycznych orbitalnych liczb kwantowych m_l, które mogą przyjmować $(2l+1)$ wartości, a mianowicie

$$m_l = -l,\ -(l-1), \ldots, -1, 0, 1, \ldots, (l-1), l.$$

Mówimy więc, że wektor orbitalnego momentu pędu, a tym samym związany z nim moment magnetyczny, jest skwantowany przestrzennie (rys. 1). Wartości



Rys. 1. Kwantowanie przestrzenne orbitalnego momentu pędu dla $l = 3$

rzutu orbitalnego momentu magnetycznego elektronu na kierunek wyróżniony przez zewnętrzne pole wynoszą

$$\mu_l^{[z]} = m_l\mu_B. \tag{6}$$

Gdy w atomie znajduje się więcej niż jeden elektron, suma wektorowa momentów pędu oraz suma momentów magnetycznych poszczególnych elektronów daje w wyniku wypadkowy moment pędu i moment magnetyczny atomu. Możliwe są dwa sposoby złożenia momentów pędu elektronów w atomie. Całkowity moment pędu atomu można otrzymać przez dodanie najpierw spinowego i orbitalnego momentu poszczególnych elektronów, a następnie zsumowanie wypadkowych momentów pędu wszystkich elektronów. Drugi sposób, prowadzący do innych wartości energii poziomów energetycznych atomu, polega na zsumowaniu najpierw wszystkich momentów spinowych, a następnie momentów orbitalnych poszczególnych elektronów. Całkowity moment pędu atomu otrzymuje się przez dodanie wypadkowego momentu spinowego $\vec{P_S}$ i wypadkowego momentu orbitalnego $\vec{P_L}$. O wyborze sposobu obliczania całkowitego momentu pędu atomu decyduje wielkość oddziaływania pomiędzy momentem spinowym i orbitalnym elektronu. Zwykle sprzężenie spin-orbita jest słabe i słuszny jest drugi sposób sumowania momentów pędu. Całkowity moment pędu atomu określamy liczbą kwantową I (liczbę tę oznacza się zazwyczaj literą J). Wartości bezwzględne odpowiednich wektorów momentu pędu atomu określone są następującymi wzorami:

całkowity moment pędu atomu

$$|\vec{P_S}| = \sqrt{S(S+1)}\;\hbar,$$

$$|\vec{P_L}| = \sqrt{L(L+1)}\;\hbar,$$

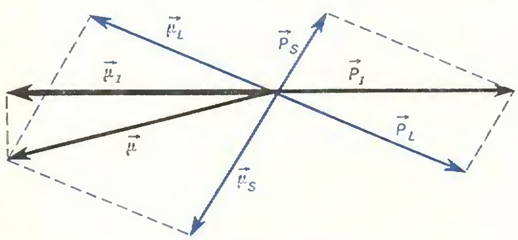
$$|\vec{P_I}| = \sqrt{I(I+1)}\,\hbar.$$

Ponieważ wypadkowy moment spinowy i moment orbitalny całkowicie zapełnionej powłoki elektronowej atomu równe są zeru, przy obliczaniu wypadkowych momentów atomu uwzględnia się jedynie momenty elektronów z powłok niezapełnionych. Z momentami pędu atomu, określonymi liczbami kwantowymi S, L i I, związane są odpowiednie momenty magnetyczne atomu μ_S, μ_L i μ_I. Zarówno momenty pędu jak i momenty magnetyczne atomu są skwanto

moment magnetyczny atomu

wane analogicznie do momentu pojedynczego elektronu. Liczbom kwantowym S, L i I przyporządkowane są odpowiednie magnetyczne liczby kwantowe m_S, m_L i m_I. Stosując atomowe liczby kwantowe można wyrazić za pomocą wzorów (1) i (6) wartości możliwych rzutów spinowego i orbitalnego momentu magnetycznego atomu na kierunek zewnętrznego pola. Wartości bezwzględnego spinowego i orbitalnego momentu magnetycznego atomu określone są wzorami analogicznymi do wzorów (3) i (5).

Z porównania wzorów (2) i (3) oraz (4) i (5) wynika, że spinowy i orbitalny moment magnetyczny niejednakowo wyrażają się przez odpowiednie momenty pędu. Dla momentu spinowego stosunek momentu magnetycznego do mechanicznego jest dwa razy większy niż taki sam stosunek dla momentu orbitalnego. Z uwagi na ten fakt kierunek wypadkowego momentu magnetycznego atomu $\vec{\mu}$ nie pokrywa się z kierunkiem wypadkowego momentu mechanicznego $\vec{P}_I$. Ponieważ ładunek elektronu jest ujemny, wektory $\vec{\mu}_S$ i $\vec{\mu}_L$ są odpowiednio antyrównoległe do wektorów $\vec{P}_S$ i $\vec{P}_L$. W skali, w której długość wektora $\vec{\mu}_L$ równa się długości wektora $\vec{P}_L$, długość wektora $\vec{\mu}_S$ jest równa podwójnej długości wektora $\vec{P}_S$. Graficzny obraz sumowania momentów mechanicznych i magnetycznych atomu przedstawiony jest na rys. 2. Wypadkowy moment magnetyczny $\vec{\mu}$ tworzy z wektorem $\vec{P}_I$ kąt różny od 180°. Można powiedzieć, że wektor $\vec{\mu}$ wykonuje precesję dookoła kierunku $\vec{P}$. Skuteczny moment magnetyczny atomu $\vec{\mu}_I$ równy jest rzutowi wektora $\vec{\mu}$ na kierunek wektora $\vec{P}_I$. Wykonując odpowiednie obliczenia trygonometryczne znajdujemy



Rys. 2. Sumowanie mechanicznych i magnetycznych momentów atomu

skuteczny moment magnetyczny atomu

$$|\vec{\mu}_I| = g\sqrt{I(I+1)}\mu_B, \tag{7}$$

gdzie

$$g = 1 + \frac{I(I+1)+S(S+1)-L(L+1)}{2I(I+1)} \tag{8}$$

jest współczynnikiem rozszczepienia spektroskopowego. Możliwe wartości rzutów wypadkowego momentu magnetycznego atomu $\vec{\mu}_I$ na kierunek zewnętrznego pola określone są wzorem:

$$\mu_I^{[z]} = g\, m_I\, \mu_B; \tag{9}$$

magnetyczna liczba kwantowa atomu przyjmuje wartości

$$m_I = -I, -(I-1), ..., (I-1), I.$$

W przypadkach granicznych, kiedy całkowity moment $\vec{P}_I = \vec{P}_S$, tzn. gdy $L = 0$, czynnik rozszczepienia spektroskopowego $g = 2$. W drugim skrajnym przypadku $\vec{P}_I = \vec{P}_L$, tzn. gdy $S = 0$, $g = 1$. Istnienie momentu magnetycznego atomu i omówiona powyżej jego kwantyzacja przestrzenna zostały potwierdzone w doświadczeniach Sterna i Gerlacha, w których badano odchylenie strumienia atomów w niejednorodnym polu magnetycznym.

Diamagnetyzm

Najbardziej naturalnym sposobem klasyfikacji substancji pod względem ich właściwości magnetycznych jest podanie wielkości charakteryzującej reakcję ośrodka na zewnętrzne pole magnetyczne H. Taką wielkością jest z definicji podatność magnetyczna

podatność magnetyczna

$$\chi = M/H;$$

M jest namagnesowaniem ośrodka, tzn. momentem magnetycznym przypadającym na jednostkę objętości. Substancje o dodatniej podatności nazywamy paramagnetykami, a o ujemnej — diamagnetykami.

powszechność diamagnetyzmu

Zjawisko diamagnetyzmu jest powszechne i właściwe wszystkim bez wyjątku ciałom. Nie zawsze jednak zjawisko to można zaobserwować, ponieważ słaby efekt diamagnetyczny bywa często przesłonięty przez silniejszy efekt paramagnetycznego namagnesowania. Diamagnetyzm można obserwować łatwo i bezpośrednio wówczas, gdy atomy, jony lub cząsteczki mają wypadkowy moment magnetyczny równy zeru. Orbity elektronowe w atomach można wyobrażać sobie jako kołowe obwody elektryczne o zerowym oporze, w których płynie prąd o stałym natężeniu. Przyłożenie zewnętrznego pola magnetycznego powoduje zmianę przechodzącego przez te obwody strumienia indukcji magnetycznej. W orbitach elektronowych indukują się wówczas dodatkowe prądy, które płyną tak długo, jak długo działa zewnętrzne pole. Zgodnie z prawem Lenza momenty magnetyczne indukowanych prądów kompensują zmianę strumienia indukcji magnetycznej, a więc mają kierunki przeciwne do pola zewnętrznego. Wobec tego substancja złożona z atomów o wypadkowym momencie magnetycznym równym zeru, umieszczona w zewnętrznym polu magnetycznym magnesuje się, przy czym zwrot wektora namagnesowania jest przeciwny do wektora natężenia pola. Podatność diamagnetyków jest zwykle bardzo mała, rzędu -10^{-6}, i nie zależy od temperatury.

obserwacja diamagnetyzmu

Do diamagnetyków należą wszystkie gazy szlachetne, niektóre metale (np. cynk, złoto, rtęć), niektóre niemetale (np. krzem, fosfor, siarka) oraz wiele związków organicznych.

Paramagnetyzm

Zjawisko paramagnetyzmu występuje tylko w tych materiałach, w których atomy lub molekuły mają stały moment magnetyczny. Zastanówmy się, jakie jest namagnesowanie ośrodka zawierającego N atomów w jednostce objętości, z których każdy ma moment magnetyczny $\vec{\mu}_I$, umieszczonego w zewnętrznym polu magnetycznym o natężeniu $\vec{H}$. Zakładamy, że atomy nie oddziałują ze sobą. Namagnesowanie ośrodka jest wynikiem orientacji momentów magnetycznych atomów wzdłuż kierunku przyłożonego pola magnetycznego pod wpływem tego pola. Drgania cieplne przeciwdziałają porządkującemu działaniu pola na momenty magnetyczne.

stały moment magnetyczny

Jak już wiemy, moment atomu o liczbie kwantowej I ma $(2I+1)$ możliwych orientacji w przestrzeni, a możliwe składowe tego momentu w kierunku pola określone są wzorem (9). Prawdopodobieństwo ω danej orientacji określone jest wzorem Boltzmanna:

$$\omega \sim \exp(\mu_I^{[z]}H/k_BT),$$

gdzie k_B oznacza stałą Boltzmanna, T — temperaturę bezwzględną. Iloczyn $-\mu_I^{[z]}H$ wyraża energię oddziaływania atomu z przyłożonym polem magnetycznym. Namagnesowanie ośrodka równe jest

namagnesowanie ośrodka

$$M = N\langle\mu_I^{[z]}\rangle, \tag{10}$$

gdzie $\langle\mu_I^{[z]}\rangle$ oznacza wartość średnią rzutu momentu magnetycznego na kierunek pola, którą można obliczyć znając wartości energii oddziaływania między atomem i polem, odpowiadające poszczególnym orientacjom momentu $\vec{\mu}_I$. Wartość $\langle\mu_I^{[z]}\rangle$ dana jest wyrażeniem:

$$\langle\mu_I^{[z]}\rangle = \frac{\sum\limits_{-I}^{+I} \mu_I^{[z]} \exp(\mu_I^{[z]} H/k_B T)}{\sum\limits_{-I}^{+I} \exp(\mu_I^{[z]} H/k_B T)}. \qquad (11)$$

Podstawiając zamiast $\mu_I^{[z]}$ prawą stronę równania (9), oraz wykonując sumowanie we wzorze (11) otrzymujemy:

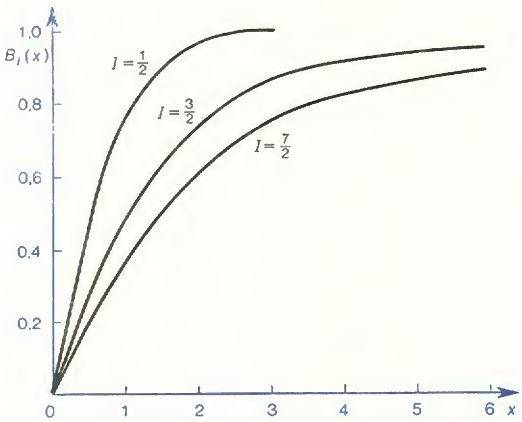
$$\langle\mu_I^{[z]}\rangle = g\mu_B I B_I(x), \qquad (12)$$

gdzie

$$x = \frac{g\mu_B IH}{k_B T},$$

funkcja Brillouina

a $B_I(x)$ nazywana jest funkcją Brillouina. Na rysunku 3 podane są wykresy $B_I(x)$ dla kilku wartości I. Dla $x \ll 1$ (tzn. dla małych natężeń zewnętrznego



Rys. 3. Wykres funkcji Brillouina $B_I(x)$ dla $I = {}^1/_2, {}^3/_2, {}^7/_2$

pola magnetycznego) funkcję $B_I(x)$ można przedstawić w przybliżonej postaci:

$$B_I(x) = \frac{I+1}{3I}\,x; \quad x \ll 1. \qquad (13)$$

Podstawiając prawe strony równości (12) i (13) do równania (10) otrzymujemy zależność namagnesowania ośrodka od pola H:

$$M = \frac{Ng^2\mu_B\, I(I+1)}{3k_B T} H. \qquad (14)$$

Porównując z definicją podatności magnetycznej $\chi = M/H$, otrzymujemy:

$$\chi = \mathrm{const}/T. \qquad (15)$$

Curie prawo

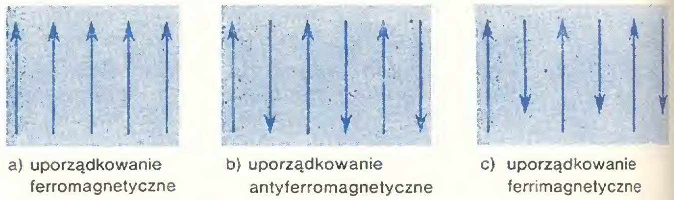
Ośrodki, dla których słuszne jest to równanie, znane jako prawo Curie, nazywamy paramagnetykami. Prawo Curie mówi, że podatność paramagnetyka nie zależy od zewnętrznego pola w zakresie małych jego natężeń ($x \ll 1$) i jest odwrotnie proporcjonalna do temperatury. Z rysunku 3 wynika, że dla dużych wartości x, co odpowiada fizycznej sytuacji $T \to 0$, funkcja Brillouina dąży do wartości asymptotycznej. Oznacza to, że dla $T \to 0$ namagnesowanie ośrodka osiąga wartość maksymalną, zwaną nasyceniem, nawet przy bardzo małych wartościach zewnętrznego pola.

przykłady paramagnetyków

Zgodność kwantowej teorii paramagnetyzmu z wynikami doświadczalnymi dla wielu ośrodków wykazujących właściwości paramagnetyczne jest doskonała. Przykładami ośrodków paramagnetycznych mogą być niektóre metale i pierwiastki oraz ich sole, gazowy tlenek azotu NO, wolne rodniki organiczne, jak np. trójfenylometyl $C(C_6H_5)_3$.

Spontaniczne uporządkowanie magnetyczne

Znany jest fakt, że pewne kryształy zawierające atomy z różnym od zera momentem magnetycznym wykazują w dostatecznie niskich temperaturach spontaniczne namagnesowanie w nieobecności zewnętrznego pola magnetycznego. Istnienie makroskopowego momentu magnetycznego w krysztale przy $H = 0$ sugeruje, że w wyniku wzajemnego oddziaływania momenty magnetyczne atomów tworzą regularną strukturę (uporządkowanie). Rysunek 4 przedstawia



Rys. 4. Niektóre możliwe uporządkowania układów momentów magnetycznych

trzy najbardziej typowe możliwe uporządkowania momentów magnetycznych w kryształach wykazujących spontaniczne namagnesowanie.

uporządkowanie ferromagnetyczne

Z danych doświadczalnych wynika, że dla kryształów ferromagnetycznych współczynnik rozszczepienia spektroskopowego g równy jest w przybliżeniu 2. Oznacza to, że ferromagnetyzm uwarunkowany jest nie orbitalnymi, a spinowymi magnetycznymi momentami atomów. Jak już wspomniano, momenty magnetyczne powłok zapełnionych są równe zeru. Elektrony walencyjne kolektywizują się przy powstawaniu stanu metalicznego tworząc paramagnetyczny gaz elektronów prawie swobodnych (→ Metale). Wobec tego ferromagnetyzm możliwy jest jedynie w kryształach zawierających atomy pierwiastków przejściowych, tzn. atomy mające niezapełnioną wewnętrzną powłokę elektronową. Są to metale przejściowe grupy żelaza z niezapełnioną powłoką 3*d* oraz pierwiastki ziem rzadkich z niezapełnioną powłoką 4*f*. Istnienie niezapełnionej powłoki wewnętrznej jest warunkiem koniecznym ale niedostatecznym powstania uporządkowania ferromagnetycznego w krysztale. Z metali grupy żelaza tylko żelazo, nikiel i kobalt są ferromagnetykami. Z metali ziem rzadkich tylko gadolin, dysproz i erb wykazują właściwości ferromagnetyczne. Ferromagnetyzm powstaje w wyniku oddziaływania wymiennego elektronów z niezapełnionych powłok, przy czym oddziaływanie to jest efektem kwantowym. Kwantowy opis tego oddziaływania podany został przez W. Heisenberga. W modelu Heisenberga energia oddziaływania wymiennego określona jest następującym wzorem:

energia oddziaływania wymiennego

$$U = -2J \sum_{i=1}^{N} \sum_{j=1}^{z} \vec{S}_i \vec{S}_j, \qquad (16)$$

gdzie N jest całkowitą liczbą spinów w krysztale, indeksy i i j numerują węzły sieci krystalicznej, a $\vec{S}_i$ i $\vec{S}_j$ oznaczają spiny atomów w i-tym i j-tym węźle sieci, traktowane jako klasyczne wektory. Wzór (16) uwzględnia oddziaływanie każdego spinu tylko z najbliższymi sąsiadami w liczbie z — stanowi to tzw.

przybliżenie najbliższych sąsiadów. Sumowanie we wzorze (16) obejmuje każdą parę spinów w krysztale tylko raz. Miarą oddziaływania wymiennego jest tzw. całka wymiany J — wielkość fizyczna nie mająca analogii w fizyce klasycznej.

W celu ilustracji wzoru (16) rozpatrzymy parę oddziałujących ze sobą spinów $\vec{S_i}$ oraz $\vec{S_j}$. Na podstawie wzoru (16) energia tego oddziaływania jest równa:

$$U = -2J\vec{S_i}\,\vec{S_j} = -2JS_iS_j\cos\varphi, \qquad (17)$$

uporządkowanie antyferromagnetyczne i ferrimagnetyczne

gdzie φ jest kątem pomiędzy $\vec{S_i}$ i $\vec{S_j}$. Energia układu fizycznego w warunkach równowagi termodynamicznej osiąga zawsze minimum. Gdy $J > 0$, najniższa wartość energii U odpowiada wartości kąta $\varphi = 0°$. Stan, któremu odpowiada najniższa energia, nazywamy stanem podstawowym. Dla $J > 0$ stanem podstawowym jest więc uporządkowanie typu ferromagnetycznego, tzn. spiny $\vec{S_i}$ i $\vec{S_j}$ są skierowane równolegle (rys. 4a). Dla $J < 0$ najniższa energia odpowiada antyrównoległemu ustawieniu spinów ($\varphi = 180°$). Możemy wyróżnić wtedy dwa rodzaje uporządkowania — uporządkowanie antyferromagnetyczne (rys. 4b) oraz ferrimagnetyczne (rys. 4c). W pierwszym przypadku wartości bezwzględne spinów $\vec{S_i}$ i $\vec{S_j}$ są takie same i wypadkowy spin uporządkowania antyferromagnetycznego równy jest zeru. W uporządkowaniu ferrimagnetycznym ze względu na różne wartości bezwzględne spinów $\vec{S_i}$ i $\vec{S_j}$ wypadkowy spin jest różny od zera.

oddziaływanie wymienne

Stany podstawowe realizowane są jedynie w temperaturze zera bezwzględnego. Przy wzroście temperatury, na skutek drgań termicznych sieci krystalicznej idealne uporządkowanie spinów ulega coraz większemu zaburzeniu i układ N spinów przechodzi w jeden z możliwych stanów wzbudzonych o energii większej od energii stanu podstawowego. Znajomość energii tych stanów pozwala określić wiele właściwości termodynamicznych kryształów ferromagnetycznych.

Omówimy dwie metody teoretycznego opisu układu oddziałujących ze sobą spinów w temperaturach wyższych od zera. Najprostsza i najstarsza teoria spontanicznego uporządkowania magnetycznego układu N spinów w sieci krystalicznej opiera się na koncepcji tzw. pola molekularnego. Jedno z nowszych i ściślejszych ujęć teorii posługuje się pojęciem tzw. fal spinowych.

Teoria pola molekularnego

Wzór (16) zastosowany do jednego atomu znajdującego się w i-tym węźle sieci krystalicznej ma postać:

$$U_1 = -2J\vec{S_i}\sum_{j=1}^{z}\vec{S_j}. \qquad (18)$$

Oddziaływanie wymienne spinu $\vec{S_i}$ z najbliższymi sąsiadami możemy zastąpić oddziaływaniem spinu $\vec{S_i}$ z hipotetycznym polem magnetycznym $\vec{H_e}$, które nazywamy polem molekularnym. Pojęcie pola molekularnego zostało wprowadzone po raz pierwszy w klasycznej teorii ferromagnetyzmu przez P. Weissa, który założył *ad hoc* istnienie pola magnetycznego wewnątrz kryształu ferromagnetycznego. W ujęciu kwantowym pole molekularne ma sens przybliżenia modelowego, w którym oddziaływanie wymienne wyrażamy w postaci oddziaływania momentu magnetycznego z efektywnym polem zewnętrznym. W takim ujęciu energię U_1 można zapisać w postaci:

$$U_1 = -g\mu_B\vec{S_i}\,\vec{H_e}. \qquad (19)$$

Wartość natężenia H_e musi być oczywiście taka, aby energia wyrażona wzorem (18) była równa energii wyrażonej wzorem (19). Wobec tego, porównując te dwa wzory, otrzymujemy

$$H_e = \frac{2J}{g\mu_B}\sum_{j=1}^{z}\vec{S_j}. \qquad (20)$$

wartość średnia spinu

Zakładamy następnie, że każdy spin $\vec{S_j}$ możemy zastąpić przez wartość średnią $\langle S_j^{[z]}\rangle$. Jest to zasadnicze przybliżenie, na którym opiera się metoda pola molekularnego. Ponieważ wszystkie atomy są identyczne i równoważne, wartość średnia $\langle S_j^{[z]}\rangle$ związana jest z całkowitym namagnesowaniem kryształu zgodnie ze wzorem

$$M = Ng\mu_B\langle S_j^{[z]}\rangle. \qquad (21)$$

Wobec tego

$$H_e = \frac{2zJ}{g\mu_B}\langle S_j^{[z]}\rangle = \frac{2zJ}{Ng^2\mu_B^2}M. \qquad (22)$$

Energia U_1 określona wzorem (19) ma analogiczną postać do wzoru na energię momentu magnetycznego umieszczonego w zewnętrznym polu. Wobec tego, przeprowadzając podobne rozumowanie jak dla ośrodka paramagnetycznego, otrzymujemy wzór na namagnesowanie ferromagnetyka analogiczny do wzoru (12):

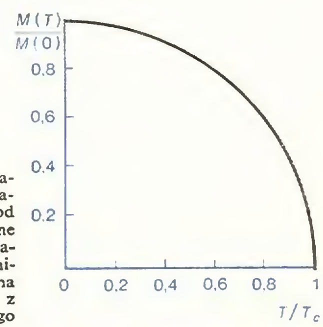
$$M = Ng\mu_B S B_s(x), \qquad (23)$$

gdzie

$$x = \frac{g\mu_B S H_e}{k_B T} \qquad (24)$$

Funkcja Brillouina B_I we wzorze (12) zależna jest od pola zewnętrznego H. Zakładamy teraz, że zewnętrzne pole magnetyczne równe jest zeru. Argument funkcji B_s we wzorze (23) zależy od pola molekularnego, a tym samym od namagnesowania M (wzór 22). Mówiąc inaczej, równanie (23) na namagnesowanie M jest uwikłane. Najprostszym sposobem znalezienia namagnesowania M w danej temperaturze jest graficzne rozwiązanie układu równań (21) i (23). Na rysunku 5 przedstawiony jest przykładowy przebieg

spontaniczne namagnesowanie ferromagnetyka



Rys. 5. Zależność spontanicznego namagnesowania ferromagnetyka od temperatury. Czarne punkty przedstawiają dane doświadczalne dla niklu; krzywa teoretyczna wynika przy $S = 1/2$ z teorii pola molekularnego

spontanicznego namagnesowania w funkcji temperatury otrzymany tą metodą. Temperatura T_C, w której znika spontaniczne namagnesowanie, nazywana jest temperaturą Curie lub temperaturą krytyczną. Wzór na temperaturę Curie otrzymany w teorii pola molekularnego ma postać:

temperatura Curie

$$T_C = \frac{2zJS(S+1)}{3k_B}. \qquad (25)$$

prawo Curie–Weissa

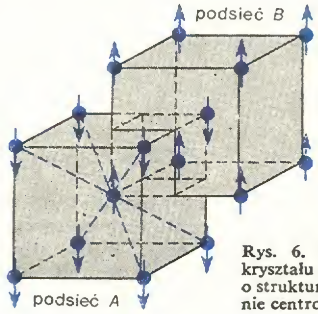
Dla temperatur $T > T_C$ ferromagnetyk zachowuje się w zewnętrznym polu magnetycznym tak jak paramagnetyk, przy czym podatność

$$\chi = \frac{\text{const}}{T - T_C}. \qquad (26)$$

Wzór (26) znany jest jako prawo Curie–Weissa.

spontaniczne namagnesowanie antyferromagnetyka

Metodę pola molekularnego można również stosować do opisu spontanicznego uporządkowania typu antyferromagnetycznego lub ferrimagnetycznego. W najprostszym modelu materiału antyferromagnetycznego sieć magnetycznych atomów można podzielić na dwie równoważne, wzajemnie przenikające się podsieci A i B w ten sposób, że najbliższymi sąsiadami atomów A są atomy B i odwrotnie. Dwupodsieciowy model antyferromagnetyka przedstawiony jest na rys. 6. Dla każdej podsieci oblicza się spontaniczne

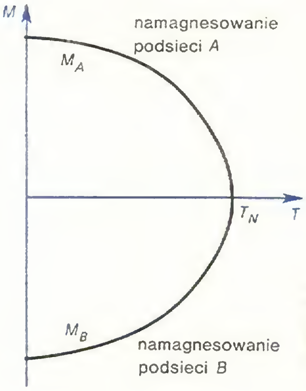


Rys. 6. Dwupodsieciowy model kryształu antyferromagnetycznego o strukturze regularnej, przestrzennie centrowanej

namagnesowanie oddzielnie. Ponieważ wartości bezwzględne wszystkich spinów są jednakowe, wypadkowe namagnesowanie podsieci A znosi się z namagnesowaniem podsieci B, a zatem moment magnetyczny kryształu jest równy zeru. Temperaturowa zależność namagnesowania podsieci dla uporządkowania antyferromagnetycznego przedstawiona jest schematycznie na rys. 7.

temperatura Néela

Temperatura T_N, w której znika jednocześnie uporządkowanie w obydwu podsieciach, nazywana jest temperaturą Néela. Typowym przykładem prostego antyferromagnetyka jest mangan.



Rys. 7. Jakościowa zależność spontanicznego namagnesowania podsieci kryształu antyferromagnetycznego od temperatury

spontaniczne namagnesowanie ferrimagnetyka

Podobny model stosuje się do wyznaczania temperaturowej zależności spontanicznego namagnesowania w przypadku uporządkowania ferrimagnetycznego. Dla prostych ferrimagnetyków wprowadza się dwie podsieci A i B jak na rys. 6, przy czym momenty magnetyczne atomów A i B nie są sobie równe. Wypadkowe namagnesowanie ferrimagnetyka równe jest sumie wektorowej momentów magnetycznych obydwu podsieci:

$$\vec{M} = \vec{M}_A + \vec{M}_B.$$

Klasycznym przykładem substancji ferrimagnetycznych są ferryty.

Metodą pola molekularnego osiągamy dość dobrą zgodność jakościową, a w wielu wypadkach również ilościową, wyników teoretycznych z danymi doświadczalnymi. Dotyczy to szczególnie prostych struktur magnetycznych w zakresie wyższych temperatur.

Metoda fal spinowych

W zakresie niezbyt wysokich temperatur w stosunku do temperatury T_C energia stanów wzbudzonych niewiele się różni od energii stanu podstawowego. Stany takie można opisać za pomocą tzw. fal spinowych, które przechodząc przez kryształ zaburzają jego uporządkowanie.

fale spinowe w kryształach ferromagnetycznych

Rozpatrzmy na wstępie uproszczoną sytuację, w której N spinów jest ustawionych w tym samym kierunku, przy czym odległość między sąsiednimi spinami jest stała i równa a (kryształ jednowymiarowy). Załóżmy ponadto, że spin w miejscu i-tym oddziałuje tylko z najbliższymi sąsiadami znajdującymi się w miejscach $i-1$ oraz $i+1$. Energię oddziaływania (16) dla takiego układu spinów można wtedy zapisać w postaci:

$$U = -\frac{1}{2}\sum_i \vec{\mu}_i \vec{H}_i, \qquad (27)$$

gdzie $\vec{\mu}_i$ jest momentem magnetycznym związanym ze spinem $\vec{S}_i$ zależnością $\vec{\mu}_i = -g\,\mu_B\,\vec{S}_i$. Oddziaływanie i-tego spinu z sąsiadami zastąpiliśmy przez oddziaływanie tego spinu z efektywnym polem magnetycznym

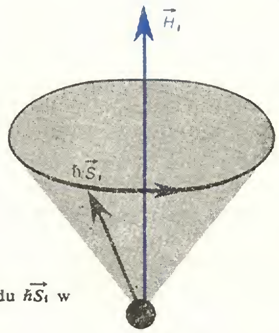
$$\vec{H}_i = -\frac{2I}{g\,\mu_B}(\vec{S}_{i-1}+\vec{S}_{i+1}). \qquad (28)$$

Zmianę momentu pędu $\hbar\vec{S}_i$ pod wpływem tego pola określa następujący związek:

$$\hbar\frac{d\vec{S}_i}{dt} = \vec{\mu}_i \times \vec{H}_i, \text{ czyli } \hbar\frac{d\vec{S}_i}{dt} = -g\mu_B \vec{S}_i \times \vec{H}_i. \qquad (29)$$

precesja momentu pędu

Moment skręcający wyrażony iloczynem wektorowym $\vec{\mu}_i \times \vec{H}_i$, powoduje zmianę jedynie kierunku wektora momentu pędu $\hbar\vec{S}_i$ (jego długość pozostaje stała, rys. 8). Rozpisując równanie (29) na składowe w kierunku osi x i y i zakładając, że amplitudy S_i^x



Rys. 8. Precesja momentu pędu $\hbar\vec{S}_i$ w polu magnetycznym $\vec{H}_i$

i S_i^y są małe, otrzymujemy następujący przybliżony układ równań:

$$\frac{dS_i^x}{dt} = \frac{2IS}{\hbar}(2S_i^y - S_{i-1}^y - S_{i+1}^y), \qquad (30)$$

$$\frac{dS_i^y}{dt} = -\frac{2IS}{\hbar}(2S_i^x - S_{i-1}^x - S_{i+1}^x).$$

Rozwiązania tego układu poszukujemy w postaci płaskiej fali bieżącej:

$$S_i^x = u \exp i(\omega t - kn_i a); \quad S_i^y = v \exp i(\omega t - kn_i a), \qquad (31)$$

którą nazywać będziemy falą spinową. Występujące tu wielkości u i v są stałymi, $i = \sqrt{-1}$, $k = 2\pi/\lambda$ jest wektorem falowym związanym z falą o długości λ. Wielkość $n_i a$, w której n_i jest liczbą całkowitą, opisuje odległość spinu S_i od spinu wybranego jako początek układu odniesienia.

zależność dyspersyjna

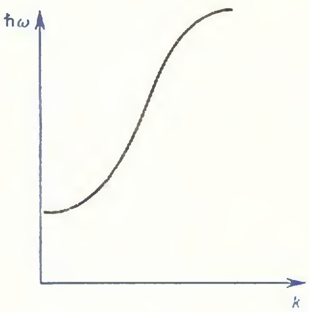
Po podstawieniu wielkości (31) do równań (30) dostajemy układ równań, który ma rozwiązania tylko wtedy, gdy spełniony jest następujący związek:

$$\hbar\omega = 4IS(1-\cos ka). \tag{32}$$

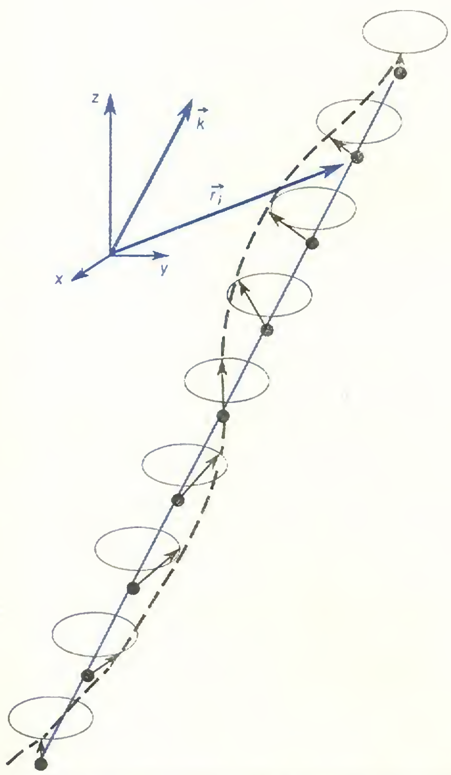
Jest to zależność dyspersyjna $\omega = \omega(k)$ dla fal spinowych w jednowymiarowym krysztale. Jeżeli uwzględnić w rozważaniach działanie przyłożonego z zewnątrz pola magnetycznego o natężeniu H, to zależność dyspersyjna przybiera postać:

$$\hbar\omega = 4IS(1-\cos ka)+g\mu_B H. \tag{33}$$

Zależność tę przedstawia rys. 9.



Rys. 9. Zależność dyspersyjna fali spinowej rozchodzącej się w krysztale jednowymiarowym



Rys. 10. Bieżąca fala spinowa w krysztale trójwymiarowym

bieżąca fala spinowa

Falę spinową w krysztale trójwymiarowym (rys. 10) opisują wzory:

$$S_i^x = u\exp i(\omega t-\vec{k}\vec{r_i}); \quad S_i^y = v\exp i(\omega t-\vec{k}\vec{r_i}), \tag{34}$$

gdzie wektor falowy $\vec{k}$ o długości $2\pi/\lambda$ skierowany jest wzdłuż kierunku rozchodzenia się fali, zaś wektor $\vec{r}$ opisuje położenie danego jonu w sieci krystalicznej. Zależność dyspersyjną dla trójwymiarowego kryształu o strukturze regularnej wyraża wzór

$$\hbar\omega = 2IS[z-\sum_m \cos(\vec{k}\vec{a}_m)]+g\mu_B H; \tag{35}$$

sumowanie rozciąga się na z wektorów $\vec{a}_m$ łączących wybrany atom z jego najbliższymi sąsiadami. Rozwijając $\cos(\vec{k}\vec{a}_m)$ w szereg dostajemy dla długich fal spinowych ($ka \ll 1$, gdzie a — stała sieci krystalicznej) zależność dyspersyjną w postaci przybliżonej:

$$\hbar\omega \approx 2ISa^2k^2+g\mu_B H. \tag{36}$$

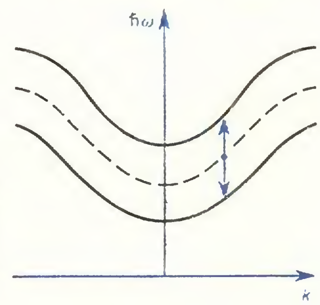
Ze wzoru tego wynika, że częstość fal spinowych w ferromagnetykach jest proporcjonalna do kwadratu wektora falowego.

fale spinowe w kryształach antyferromagnetycznych

Rozpatrzmy teraz kryształ antyferromagnetyczny, którego sieć krystaliczna przedstawiona jest na rys. 6. Rozumując podobnie jak przy ferromagnetykach dochodzimy do następującego wzoru określającego częstość fal spinowych w krysztale antyferromagnetycznym:

$$\hbar\omega = g\mu_B[(H_e+H_a)^2-H_e^2(\tfrac{1}{z}\sum_m \cos\vec{k}\vec{a}_m)^2]^{1/2} \pm g\mu_B H, \tag{37}$$

gdzie H_e, określone wzorem (22), jest efektywnym polem opisującym oddziaływanie wymienne pomiędzy najbliższymi sąsiadami z różnych podsieci. Uwzględniona została również za pośrednictwem pola H_a anizotropia magnetokrystaliczna ośrodka (→ Struktura domenowa i procesy magnesowania), która w antyferromagnetykach jest zazwyczaj bardzo silna. W wyniku silnej anizotropii częstość fal spinowych w antyferromagnetykach jest zwykle znacznie większa niż w ferro- czy w ferrimagnetykach. Z powyższego wzoru wynika, że przyłożenie zewnętrznego pola magnetycznego powoduje rozszczepienie krzywej dyspersyjnej fal spinowych na dwie gałęzie (rys. 11).



Rys. 11. Zależność dyspersyjna fal spinowych w krysztale antyferromagnetycznym w nieobecności zewnętrznego pola magnetycznego (linia niebieska), oraz po przyłożeniu pola H (linie czarne

Ze wzoru (37) dostajemy dla długich fal spinowych ($ka \ll 1$) następującą zależność dyspersyjną:

$$\hbar\omega \approx g\mu_B\left[(H_e+H_a)^2-H_e^2\left(1-\frac{2k^2a^2}{z}\right)\right]^{1/2} \pm g\mu_B H. \tag{38}$$

Jeśli pominiemy podobnie, jak przy ferromagnetyzmie, anizotropię magnetokrystaliczną kładąc $H_a = 0$, to otrzymamy proporcjonalność częstości ω do pierwszej potęgi k (w ferromagnetykach zależność ta jest kwadratowa).

fale spinowe w kryształach ferrimagnetycznych

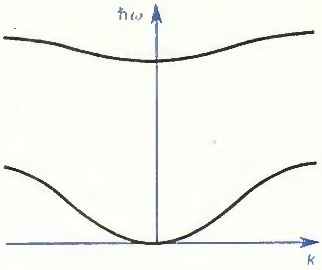
Na rys. 6 widzimy dwupodsieciowy kryształ ferrimagnetyczny. Różni się on od kryształu antyferromagnetycznego tym, że wartości bezwzględne spinów z obu podsieci są różne, w wyniku czego wypadkowy spin i moment magnetyczny są różne od zera. Taki ferrimagnetyk ma dwie gałęzie zależności dyspersyjnej (rys. 12). Gałęzie te dla stosunkowo długich fal spinowych opisane są przez następujące wzory:

$$\hbar\omega_1 = 4\sqrt{2}\left|\frac{IS^AS^B}{S^A-S^B}\right|a^2k^2+g\mu_B H, \tag{39}$$

$$\hbar\omega_2 = 16I|S^A-S^B|\left[1+\frac{S^AS^B}{4(S^A-S^B)}a^2k^2\right]^{1/2}+g\mu_B H, \tag{40}$$

gdzie S^A i S^B są spinami odpowiednio z jednej i drugiej podsieci. Występuje przy tym ścisła analogia pomiędzy gałęzią opisaną przez wzór (39) a zależnością dyspersyjną w ferromagnetykach ($\omega \sim k^2$). Druga

gałąź zależności dyspersyjnej występuje tylko w ferrimagnetykach. W przeciwieństwie do sytuacji odpowiadającej pierwszej gałęzi, precesujące z częstością ω_2 spiny z obu podsieci nie są teraz antyrównoległe, co prowadzi do nieznikania energii wymiennej nawet w granicy $k \to 0$.



Rys. 12. Dwie gałęzie zależności dyspersyjnej fal spinowych w krysztale ferrimagnetycznym w nieobecności pola magnetycznego H. Przyłożenie tego pola powoduje przesunięcie obu gałęzi w stronę wyższych częstości

Kwantowanie fal spinowych. Własności kryształów magnetycznych w niskich temperaturach

Energie fal spinowych są skwantowane. Oznacza to, że energia fali spinowej o częstości ω_k i wektorze falowym $\vec{k}$ nie może przyjmować dowolnych wartości, lecz tylko wartości:

$$E_k = (n_k + \tfrac{1}{2})\hbar\omega_k, \tag{41}$$

gdzie $n_k = 0, 1, 2, \ldots$ nosi nazwę kwantowej liczby fali spinowej, zaś wielkość $\hbar\omega_k$ określona jest zależnością dyspersyjną.

magnony

Każde elementarne wzbudzenie o energii $\hbar\omega_k$ nosi nazwę magnonu i może być traktowane jako poruszająca się w krysztale oddzielna cząstka. W tym ujęciu liczbę kwantową n_k można traktować jako liczbę magnonów o tej samej długości wektora falowego $\vec{k}$ i częstości ω_k.

Na podstawie powyższych stwierdzeń można podać następującą interpretację nisko położonych stanów wzbudzonych kryształu magnetycznego: każdy taki stan jest odchyleniem od konfiguracji idealnego uporządkowania w wyniku występowania wielu fal spinowych. W pierwszym przybliżeniu fale te można traktować jako niezależne od siebie. Dany stan wzbudzony określa się podając wartości liczb kwantowych $n_1, \ldots, n_n$ dla fal spinowych o wektorach $\vec{k}_1, \ldots \vec{k}_n$, przy czym dla skończonej próbki może występować tylko pewna skończona liczba wzbudzeń n_k.

Weźmy teraz pod uwagę jednostkę objętości ferromagnetyka, w skład której wchodzi N spinów, każdy o wartości $\vec{S}$. Rzuty spinu $\vec{S}$ na pewien wyróżniony kierunek określone są przez magnetyczną liczbę spinową, która może przyjmować jedną z wartości $S, S-1, \ldots, -S$. Wynikają stąd dwa wnioski. Po pierwsze, w ujęciu kwantowym fali spinowej nie można przedstawić tak, jak to pokazuje rys. 10. Po drugie, magnetyczna liczba całkowitego spinu układu składającego się z N spinów, m_S, może przyjmować jedynie wartości $NS, NS-1, \ldots, -NS$.

W stanie podstawowym, w którym wszystkie spiny skierowane są w jedną stronę, magnetyczna liczba kwantowa układu $m_S = NS$. Wzbudzenie jednego magnonu wiąże się z odchyleniem jednego ze spinów w wyniku czego liczba kwantowa m_S maleje do wartości $NS-1$. To odchylenie spinu na skutek oddziaływania wymiennego nie pozostaje zlokalizowane w jednym miejscu, lecz wędruje stale przez całą sieć krystaliczną. Takiemu wędrującemu odchyleniu spinowemu można, podobnie jak w przedstawieniu klasycznym, przypisać pewien wektor falowy $\vec{k}$ oraz związaną z nim częstość ω_k.

Jeśli w krysztale zostało wzbudzonych n magnonów, to liczba m_S wynosi

$$m_S = NS - n, \tag{42}$$

gdyż wzbudzenie każdego magnonu zmniejsza wartość rzutu całkowitego spinu na pewien wyróżniony kierunek o jedność. Z rzutem tym wiąże się namagnesowanie

$$M = g\,\mu_B m_S. \tag{43}$$

Tak więc jeśli liczba spinów N odnosi się do jednostki objętości rozpatrywanego kryształu, to wzbudzenie n magnonów powoduje obniżenie namagnesowania M od wartości maksymalnej $M_0 = Ng\mu_B S$ do wartości

$$M = M_0 - ng\mu_B. \tag{44}$$

Powyższy wynik można sformułować w sposób następujący: wzbudzenie jednego magnonu prowadzi do zmniejszenia momentu magnetycznego próbki o wartość równą $g\mu_B$ niezależnie od tego, jaki jest wektor falowy i częstości magnonu.

Wynika stąd ważny wniosek, że jeśli znamy całkowitą liczbę magnonów o różnych wektorach falowych $\vec{k}$ w jednostce objętości kryształu, $\sum_k n_k$, to potrafimy obliczyć też zmianę namagnesowania, która wynosi:

$$M_0 - M = g\mu_B \sum_k n_k. \tag{45}$$

fermiony i bozony

Jak już mówiliśmy, magnony można traktować podobnie jak elektrony czy fonony, tzn. jako cząstki obdarzone pewną energią i pędem. Średnia liczba cząstek o danej energii zależy od temperatury, przy czym zależność ta opisywana jest przez wzory statystyczne, które w zależności od rodzaju cząstek przyjmują jedną z poniższych postaci:

$$\bar{n}_k = \frac{1}{\exp(\hbar\omega_k/k_B T) + 1}, \tag{46}$$

lub

$$\bar{n}_k = \frac{1}{\exp(\hbar\omega_k/k_B T) - 1}, \tag{47}$$

gdzie k_B jest stałą Boltzmanna, T — temperaturą bezwzględną, zaś $\hbar\omega_k$ jest energią cząstki o wektorze falowym $\vec{k}$.

O cząstkach, do których stosuje się wzór (46) powiadamy, że podlegają statystyce Fermiego–Diraca. Cząstki te nazywamy fermionami. Przykładem fermionów są np. elektrony.

Cząstki, do których stosuje się wzór (47), podlegają statystyce Bosego–Einsteina i noszą nazwę bozonów. Do bozonów zaliczamy zarówno fonony, jak i magnony. Rozkład Bosego stosuje się wtedy, gdy wszystkie stany kwantowe układu można obsadzać danymi cząstkami bez ograniczeń. Natomiast rozkład Fermiego stosuje się do takich cząstek, które w danym układzie muszą znajdować się w różnych stanach kwantowych. Inaczej mówiąc do fermionów stosuje się zakaz Pauliego.

Korzystając ze wzoru (47) można obliczyć liczbę magnonów wzbudzonych w danej temperaturze

$$\sum_k n_k = N\frac{0{,}0587}{b}\left(\frac{k_B T}{2IS}\right)^{3/2}, \tag{48}$$

gdzie $b = 1, 2, 4$ odpowiednio dla sieci regularnej prymitywnej, regularnej centrowanej przestrzennie i regularnej płasko centrowanej. Ze związków (45) i (48), jak też na podstawie zależności $M_0 = M(0) = Ng\mu_B S$ dostajemy ostatecznie:

prawo 3/2 Blocha

$$M(T) = M(0)\left[1 - \frac{0{,}0587}{bS}\left(\frac{k_B T}{2IS}\right)^{3/2}\right]. \tag{49}$$

Jest to tzw. „prawo $^3/_2$ Blocha" dobrze opisujące zależność namagnesowania od temperatury w zakresie niskich temperatur.

W podobny sposób można znaleźć energię układu związaną ze wzbudzonymi falami spinowymi. Jak wynika z mechaniki kwantowej, energia ta jest wyrażona wzorem:

$$U_M = \sum_k \bar{n}_k \hbar\omega_k. \tag{50}$$

ciepło właściwe ferromagnetyków

Można stąd obliczyć ciepło właściwe przy stałym namagnesowaniu, C_M, które jest pochodną U_M po temperaturze:

$$C_M = \frac{0{,}113}{b} N k_B \left(\frac{k_B T}{2IS}\right)^{3/2} \tag{51}$$

Powyższe wzory odnoszą się do ferromagnetyków. Stosują się one jednak również do ferrimagnetyków.

ferrimagnetyki

Wprawdzie w ferrimagnetykach występują dwa rodzaje fal spinowych (rys. 12), jednakże fale spinowe o dużej energii (nie dążącej do zera przy $k \to 0$) nie są właściwie wzbudzane aż do temperatury około 10 K. W związku z tym do odchylenia namagnesowania od maksymalnej wartości M_0 przyczyniają się głównie fale spinowe o zależności dyspersyjnej analogicznej do zależności charakteryzującej ferromagnetyk, tj. fale spinowe o energii proporcjonalnej do k^2 (wzór 39).

antyferromagnetyki

Obliczenia dla antyferromagnetyków są znacznie bardziej skomplikowane z uwagi na duży zazwyczaj wkład anizotropii magnetokrystalicznej do wyrażenia określającego energię fal spinowych. Jeśli jednak pominąć anizotropię, to okazuje się, że zmiana namagnesowania wraz z temperaturą dana jest przez poniższe wyrażenie:

$$M(T) = M_0(1 - AT^2). \tag{52}$$

Prawo $^3/_2$ Blocha nie stosuje się zatem do antyferromagnetyków. W dodatku obliczenia ciepła właściwego prowadzą do wniosku, że zmienia się ono proporcjonalnie do T^3 (a nie do $T^{3/2}$, jak we wzorze 51). Jest to wynikiem proporcjonalności energii fal spinowych do pierwszej potęgi k, a nie do k^2, jak to miało miejsce w ferromagnetykach.

Zależność częstości fal spinowych od ich wektora falowego wyznacza się m.in. za pomocą niesprężystego rozpraszania neutronów. Neutrony zostają rozproszone przez strukturę magnetyczną wraz z jednoczesną kreacją lub anihilacją magnonu. Znając energię i pęd padających i rozproszonych neutronów można wyznaczyć na tej podstawie zależność dyspersyjną dla magnonów.

Poznanie właściwości fal spinowych pozwala zrozumieć wiele zjawisk obserwowanych doświadczalnie w ciałach magnetycznych. Teoria fal spinowych stanowi jednocześnie opartą na mechanice kwantowej metodę opisu podstawowych właściwości tych ciał.

stosowalność teorii fal spinowych

Stosuje się ją z dobrymi wynikami w zakresie stosunkowo niskich temperatur, rzędu $^1/_{10}$ temperatury Curie. W wyższych temperaturach liczba magnonów staje się na tyle duża, że założenie o niezależności fal spinowych przestaje być słuszne. Należy uwzględniać wtedy również oddziaływania pomiędzy falami spinowymi. Przy dostatecznie dużej liczbie wzbudzeń fale spinowe zaczynają rozpraszać się na falach spinowych, przy czym to rozpraszanie może znacznie zmieniać ich energię. W rezultacie wprowadzonych poprawek można przedłużyć zakres stosowalności metody fal spinowych do temperatur przewyższających $^1/_2$ temperatury Curie. W temperaturach jeszcze wyższych należy stosować już inne metody opisu.

Przedstawiony tu zarys teorii magnetyzmu zawiera jedynie jej podstawowe elementy. Prezentowane wyniki słuszne są dla stosunkowo prostych, modelowych przypadków. Bogactwo form, w jakich przejawia się przyroda, powoduje, że teoria dająca dobry opis rzeczywistych układów fizycznych jest znacznie bardziej złożona, szczególnie pod względem matematycznym. Prezentowane przykładowo, na rys. 4, typy uporządkowania magnetycznego są najprostszymi z możliwych. Wiele magnetyków wykazuje bardziej złożone uporządkowanie, wymagające np. wprowadzenia do modelowego przedstawienia więcej niż dwu podsieci. Omówione bezpośrednie oddziaływanie wymienne dwóch sąsiednich atomów magnetycznych również nie jest jedynym możliwym. W przypadku wielu rzeczywistych, niemetalicznych magnetyków, aby opisać prawidłowo oddziaływanie wymienne, należy uwzględnić wpływ atomów niemagnetycznych, zajmujących miejsca między atomami magnetycznymi w sieci krystalicznej.

rola atomów niemagnetycznych

Atomy niemagnetyczne uczestniczą w oddziaływaniach wymiennych atomów magnetycznych będąc pewnego rodzaju „przekaźnikami" tych oddziaływań. Kwantowy opis pośredniego oddziaływania wymiennego jest oczywiście bardziej złożony. Czytelnika głębiej interesującego się teorią magnetyzmu odsyłamy do specjalistycznej literatury podanej niżej.

Teoria magnetyzmu jest jedną z trudniejszych teorii fizycznych, ciągle się jeszcze rozwijającą.

C. Kittel *Wstęp do fizyki ciała stałego*, Warszawa 1974; D. H. Martin *Magnetism in solids*, London 1967; A. H. Morrish *Fizyczne podstawy magnetyzmu*, Warszawa 1970; J. S. Smart *Effective Field Theories of Magnetism*, Philadelphia 1966 (ros. Moskwa 1968).

Struktura domenowa i procesy magnesowania

Henryk Szymczak i Rita Szymczak

Zagadnienie domen należy do głównych zagadnień w fizyce materiałów magnetycznych. Poznanie własności statycznych i dynamicznych domen jest kluczem do wyjaśnienia mechanizmów określających podstawowe parametry techniczne materiału magnetycznego. Badanie domen jest pomostem pomiędzy badaniami technicznymi a badaniami o charakterze czysto poznawczym, gdyż struktura domenowa zdeterminowana jest z kolei oddziaływaniami o charakterze kwantowym. Do opisu własności domen stosuje się zarówno formalizm kwantowy jak i fenomenologiczny. Do swego rodzaju paradoksu urasta fakt, że własności domen cylindrycznych zostały opisane przez ich odkrywcę (A. H. Bobeck, 1967) za pomocą aparatu matematycznego jedynie z zakresu szkoły średniej. Tematyce domen cylindrycznych poświęca się obecnie najwięcej prac (spośród prac poświęconych fizyce magnetyków) i nic nie wskazuje na to, by zainteresowanie tym tematem miało maleć. A przecież problem domen jest jednym z najstarszych problemów badawczych. Hipotezę o istnieniu domen wysunął już w 1907 r. P. Weiss jednocześnie z hipotezą pola molekularnego (zwanego również polem Weissa).

pole molekularne (Weissa)

Hipoteza o istnieniu wewnętrznych, bardzo silnych (często rzędu 10^8 A/m) pól magnetycznych potrzebna była Weissowi do wyjaśnienia zjawiska spontanicznego uporządkowania. Weiss nie wyjaśnił natury tego pola. Według współczesnych poglądów pole molekularne jest efektywnym polem opisującym oddziaływania wymienne między jonami magnetycznymi (tzn. jonami o różnym od zera momencie magnetycznym). Ma więc ono naturę kwantową. Przyjęcie hi-

potezy pola molekularnego (→ Teoria magnetyzmu) wyjaśnia wiele podstawowych własności kryształów magnetycznych (zależność namagnesowania od temperatury, istnienie temperatury Curie itp.), nie wyjaśnia jednak faktu, że kryształy ferromagnetyczne są najczęściej obserwowane w stanie o zerowym wypadkowym namagnesowaniu. To właśnie doprowadziło Weissa do postawienia drugiej hipotezy — hipotezy o istnieniu domen. Według tej hipotezy domeny magnetyczne (zwane również obszarami Weissa) są mikroobszarami, w których namagnesowanie jest jednorodne (zarówno ze względu na kierunek jak i amplitudę). Tak więc w obszarze domeny mamy do czynienia z pełnym (maksymalnym) namagnesowaniem. Gdy kryształ znajduje się w stanie rozmagnesowania, kierunki namagnesowania w różnych domenach rozrzucone są przypadkowo tak, że wypadkowy moment magnetyczny w dowolnym kierunku jest równy zeru. Tu dochodzimy do najbardziej, jak się wydaje, frapującej i zagadkowej własności materiałów ferromagnetycznych. Wiadomo oczywiście, że materiały ferromagnetyczne mogą być trwale namagnesowane. O wiele bardziej intrygujący jest jednak fakt, że namagnesowanie ich może być zmieniane przy użyciu stosunkowo niewielkich pól magnetycznych. Pola rzędu 80 A/m są często wystarczające do nasycenia lub do zmiany kierunku namagnesowania w krysztale, chociaż jak już wspomniano, pole molekularne ma natężenie 10^7–10^8 A/m. Ten zadziwiający fakt hipoteza Weissa tłumaczy bardzo prosto w następujący sposób: efektem przyłożenia zewnętrznego pola magnetycznego nie są zmiany na poziomie atomowym (chyba, że natężenie tego pola jest porównywalne z natężeniem pola molekularnego), ale ustawienie się wektorów namagnesowania poszczególnych domen w kierunku pola bądź to w rezultacie obrotu tych wektorów, bądź też zmiany kształtu i wielkości poszczególnych domen. Weiss nie był w stanie wyjaśnić, w jaki sposób powstaje struktura domenowa. Dopiero w 1935 r. L. D. Landau i J. M. Lifszyc podali teoretyczne uzasadnienie hipotezy Weissa udowadniając, że struktura domenowa powstaje w wyniku dążenia układu do stanu o minimalnej energii wewnętrznej. Wskutek podziału kryształu na domeny zmniejsza się znacznie energia pola rozmagnesowującego. Pochodzenia tego pola należy szukać w oddziaływaniu dipolowym, które jest znacznie słabsze niż oddziaływania wymienne. Może się wydać rzeczą nienormalną, że faworyzowana jest raczej struktura domenowa niż stan jednorodnego namagnesowania. Przyczyną tego stanu rzeczy jest fakt, że oddziaływania dipolowe są oddziaływaniami dalekiego zasięgu, stosunkowo wolno malejącymi z odległością, podczas gdy oddziaływania wymienne mają charakter krótkozasięgowy i ograniczone są zazwyczaj do najbliższych sąsiadów. Dlatego nie ma sprzeczności w fakcie istnienia struktury domenowej w całym krysztale i istnienia jednorodnego lub prawie jednorodnego namagnesowania na małych odległościach, tzn. wewnątrz pojedynczej domeny w krysztale. Energia wewnętrzna określająca rozmiary i własności struktury domenowej składa się z trzech zasadniczych części: energii wymiany, energii anizotropii magnetokrystalicznej i energii magnetostatycznej pola rozmagnesowującego.

domeny magnetyczne — obszary Weissa

teoria Landaua i Lifszyca

energia wewnętrzna struktury domenowej

Energia wymiany

Jest to podstawowy rodzaj energii w krysztale magnetycznym, warunkujący powstanie uporządkowania magnetycznego. Oddziaływania wymienne określają rodzaj uporządkowania magnetycznego (ferro-, antyferro-, ferrimagnetyczne), zależność namagnesowania od temperatury, wpływają też istotnie na inne parametry magnetyczne materiału (np. na anizotropię). Energię oddziaływania wymiennego (→ Teoria magnetyzmu, wzór 16) dwóch jonów i-tego z j-tym, obdarzonych spinami S_i i S_j, można zapisać w postaci klasycznej:

$$E_{wym} = -2JS^2\cos\varphi, \tag{1}$$

gdzie φ jest kątem pomiędzy sąsiednimi spinami. Zakładając, że kąt ten jest mały, można skorzystać z przybliżonej zależności $\cos\varphi = 1-{}^1/_2\varphi^2$ i wzór (1) zapisać (po pominięciu stałej, niezależnej od φ) w postaci:

$$E_{wym} = JS^2\varphi^2. \tag{2}$$

Analiza powyższego wzoru prowadzi do wniosku o istnieniu obszarów przejściowych między domenami, zwanych inaczej ścianami domenowymi. Ścianą domenową nazywa się warstwę przejściową, która oddziela przylegające do siebie domeny namagnesowane w różnych kierunkach. Całkowita zmiana kierunku spinu (lub momentu magnetycznego) między domenami nie zachodzi w postaci nagłego skoku w obszarze jednej płaszczyzny atomowej, lecz — w sposób stopniowy — na przestrzeni wielu płaszczyzn atomowych. Energia wymiany zmniejsza się, gdy zmiana kierunku spinów rozkłada się na wiele spinów. Aby to zrozumieć załóżmy, że spiny w sąsiednich domenach skierowane są przeciwnie. Jeżeli całkowita zmiana o kąt π zachodzi stopniowo przez N różnych przejść, to kąt między sąsiednimi spinami jest równy π/N, a więc energia wymiany przypadająca na parę sąsiednich jonów jest równa

ściany domenowe

$$E_{wym} = JS^2(\pi/N)^2.$$

Całkowita energia wymiany odpowiadająca łańcuchowi złożonemu z $N+1$ jonów wynosi:

$$E_{wym} = JS^2\pi^2/N. \tag{3}$$

Tak więc stopniowa zmiana kierunków spinów w ściance domenowej znacznie zmniejsza energię wymiany. Ze wzoru (3) wynika, że ścianka domenowa powinna się rozszerzać bez ograniczenia (gdy $N \to \infty$, wtedy $E_{wym} \to 0$). Byłoby tak, gdyby nie występowały inne rodzaje energii, a przede wszystkim energia anizotropii magnetokrystalicznej.

Energia anizotropii magnetokrystalicznej

Energia anizotropii magnetokrystalicznej jest tą częścią energii kryształu, która zależy od kierunku wektora namagnesowania. Anizotropia magnetokrystaliczna jest przyczyną występowania kierunków łatwego (i trudnego) magnesowania. Przez kierunek łatwego magnesowania rozumie się kierunek, w którym kryształ magnesuje się do nasycenia przy najmniejszym natężeniu pola magnetycznego. W dalszych rozważaniach rozpatrzymy przykład, gdy kryształ ma jeden kierunek łatwego magnesowania. Z punktu widzenia anizotropii magnetokrystalicznej takie kryształy nazywa się jednoosiowymi. Charakter anizotropii magnetokrystalicznej, czyli istnienie jednego lub wielu kierunków łatwego magnesowania, zależy od symetrii sieci krystalicznej. Kryształy należące do układu heksagonalnego, tetragonalnego czy romboedrycznego są kryształami jednoosiowymi (w sensie magnetycznym). Nie zawsze jednak taka korelacja między strukturą krystaliczną i anizotropią istnieje. W 1970 r. odkryto, że w niektórych kryształach o strukturze granatu (a więc w kryształach regularnych) w procesie wzrostu może być indukowana jednoosiowa anizotropia. Jeszcze ciekawszy efekt zaobserwowano w 1973 r., kiedy to udało się wytworzyć jednoosiową anizotropię w amorficznych cienkich warstwach stopów Gd-Co i Gd-Fe. A przecież w materiałach amorficznych trudno wyróżnić jakiś kierunek. W obu wypadkach natura anizotropii jednoosiowej nie została jeszcze do końca wyjaśniona. Gęstość energii anizotropii magnetokrystalicznej (a więc energii na jednostkę objętości) dla magnetyków jednoosiowych zapisuje się najczęściej w postaci rozwinięcia typu:

kierunek łatwego magnesowania

$$E_{an} = K_1\sin^2\theta + K_2\sin^4\theta + \ldots, \tag{4}$$

gdzie K_1, K_2 są stałymi anizotropii magnetokrystalicznej, θ — kątem pomiędzy wektorem namagnesowania i wyróżnioną osią. Do opisu większości materiałów jednoosiowych wystarcza jedna stała K_1.

szerokość ścianki domenowej

Znając wyrażenie opisujące energię anizotropii magnetokrystalicznej można by już przystąpić do obliczenia szerokości ścianki domenowej i energii niezbędnej do jej utworzenia. Nie wiemy jednak, jak zmienia się kierunek spinu w ściance (zakładamy jednorodny charakter zmian). Dokładnie funkcję opisującą zmianę kierunku spinu w ściance można obliczyć metodami rachunku wariacyjnego. Przy szacunkowych obliczeniach przyjmujemy, że w ściance $\theta = \pi/2$. Wtedy na podstawie wzoru (4) energia anizotropii na jednostkę powierzchni ścianki wynosi

$$\sigma_{\text{an}} = K_1 Na, \tag{5}$$

gęstość energii ścianki domenowej

gdzie a jest stałą sieci krystalicznej. Całkowita gęstość energii ścianki domenowej σ_w zawiera oprócz σ_{an} jeszcze wyraz opisujący energię wymiany, E_{wym}/a^2, czyli

$$\sigma_w = JS^2\pi^2/Na^2 + K_1 Na. \tag{6}$$

Parametr N wyznacza się z warunku, że energia ścianki domenowej powinna osiągać wartość minimalną:

$$\frac{d\sigma_w}{dN} = 0,$$

czyli

$$N = \sqrt{\pi^2 JS^2/K_1 a^3}. \tag{7}$$

bieguny powierzchniowe

Gdy namagnesowanie jest jednorodne i skierowane prostopadle do powierzchni kryształu, wtedy nie pojawiają się oczywiście bieguny objętościowe, natomiast łatwo obliczyć gęstość biegunów na powierzchni. Wynosi ona $\varrho = +|\vec{M}|$, gdy namagnesowanie skierowane jest ku powierzchni i $\varrho = -|\vec{M}|$ — w przeciwnym przypadku. Energię magnetostatyczną oblicza się podobnie jak energię oddziaływań ładunków elektrycznych:

$$E_{\text{magn}} = \frac{1}{8\pi\mu_0}\int\limits_V\int\limits_V \frac{\varrho(r_1)\varrho(r_2)}{r_{12}}\, dV_1 dV_2. \tag{11}$$

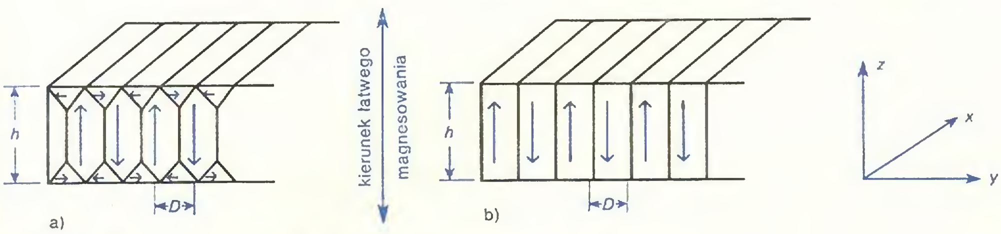
We wzorze (11) całkowanie prowadzi się po objętości kryształu V; r_{12} — odległość między punktem o współrzędnych r_1 i punktem o współrzędnych r_2. Z wyprowadzonych wzorów można korzystać przy analizie najprostszych struktur domenowych.

Strukturę domenową w jednoosiowych magnetykach opisuje się zazwyczaj za pomocą jednego z dwu modeli:

1) model struktury zamkniętej — zaproponowany w 1935 r. przez L. D. Landaua i J. M. Lifszyca (rys. 1a);

2) model struktury otwartej — zaproponowany w 1949 r. przez Ch. Kittela (rys. 1b).

Całkowanie we wzorze (11) znacznie się upraszcza jeśli założyć, że kryształ ma kształt płytki o grubości h, nieograniczonej w płaszczyźnie xy. Jak widać z rysunków kryształ rozbity jest na bloki prostopadłościenne o szerokości D (D — szerokość domeny),



Rys. 1. Modele struktury domenowej: a) struktura zamknięta, b) struktura otwarta

Szerokość ścianki domenowej $\delta = Na$, zatem

$$\delta = \pi\sqrt{JS^2/K_1 a}, \tag{8}$$

zaś energia ścianki

$$\sigma_w = 2\pi\sqrt{K_1 JS^2/a}. \tag{9}$$

Mimo wielu założeń upraszczających, wyrażenia (8) i (9) różnią się nieznacznie od wyrażeń dokładnych, wyznaczonych na podstawie rachunku wariacyjnego. Dla ilustracji oszacujmy wartości δ i σ_w w żelazie przyjmując $J = 2{,}16\cdot 10^{-21}$ J, $S = 1$, $K_1 = 4{,}2\cdot 10^4$ J/m^3, $a = 2{,}86\cdot 10^{-10}$ m:
$\delta = 4{,}2\cdot 10^{-8}$ m (czyli ok. 150 stałych sieci),
$\sigma_w = 1{,}1\cdot 10^{-3}$ J/m^2.
Jak widać z powyższych obliczeń ścianki domenowe są na ogół bardzo wąskie. Dlatego bada się je najczęściej metodami mikroskopii elektronowej.

Energia magnetostatyczna pola rozmagnesowującego

Jest to najtrudniejsza do obliczenia część energii kryształu zawierającego domeny magnetyczne. Zazwyczaj w fenomenologicznym przybliżeniu korzysta się z wynikającej z równań Maxwella formalnej analogii między statycznymi polami magnetycznymi i elektrycznymi. Wprowadza się pojęcie gęstości biegunów (lub gęstości ładunków magnetycznych) zdefiniowanej następująco:

gęstość biegunów magnetycznych

$$\varrho = -\operatorname{div}\vec{M}; \tag{10}$$

$\vec{M}$ — wektor namagnesowania.

struktura domenowa zamknięta (Landaua–Lifszyca)

wysokości h i długości nieograniczonej. Ścianki domenowe, wychodzące na powierzchnię płytki tworzą układ równoległych linii (stąd inna nazwa omawianych struktur — struktury paskowe). Struktura Landaua–Lifszyca różni się od struktury Kittela obecnością przy powierzchni tzw. domen zamykających. Domeny te powodują, że namagnesowanie nie ma składowej prostopadłej do powierzchni płytki, a więc $\varrho \equiv 0$. Można łatwo obliczyć energię całkowitą (liczoną na jednostkę powierzchni) dla modelu Landaua–Lifszyca. Energia całkowita składa się tu z 2 części:

$$E_{\text{LL}} = E_p + E_a. \tag{12}$$

Pierwsza część jest energią związaną z obecnością ścianek domenowych. Jeżeli rozmiary kryształu w kierunkach x i y oznaczymy chwilowo przez L_x i L_y (pamiętając, że $L_x, L_y \to \infty$), to całkowita powierzchnia ścian domenowych (bez domen zamykających, których wkład można pominąć) wynosi hL_xL_y/D. Stąd

$$E_p = \sigma_w h/D.$$

Druga część we wzorze (12) reprezentuje energię anizotropii magnetokrystalicznej związaną z istnieniem domen zamykających, w których $\theta = \pi/2$;

$$E_a = KD/2.$$

Szerokość domen D wyznacza się z warunku na minimum energii układu:

$$\frac{dE_{\text{LL}}}{dD} = 0,$$

stąd

$$D = \sqrt{2\sigma_w h/K}\,, \quad E_{LL} = \sqrt{2\sigma_w K h}\,. \tag{13}$$

struktura otwarta (Kittela)

W przypadku struktury Kittela całkowita energia również składa się z 2 części:

$$E_K = E_p + E_m\,, \tag{14}$$

gdzie E_p ma postać identyczną jak we wzorze (12), a E_m reprezentuje energię magnetostatyczną związaną z obecnością biegunów na powierzchni płytki. Korzystając z (11) można otrzymać następujące wyrażenie:

$$E_m = 1{,}08\cdot 10^5 M^2 D.$$

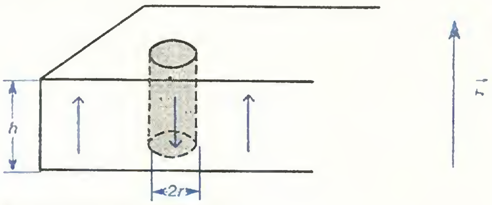
Z warunku na minimum energii układu otrzymuje się:

$$D = \sqrt{\sigma_w h/1{,}08\cdot 10^5 M^2}\,,$$
$$E_K = 2\sqrt{1{,}08\cdot 10^5 M^2 \sigma_w h}\,. \tag{15}$$

Porównując E_K i E_{LL} można ocenić, która z omawianych struktur jest energetycznie korzystniejsza. Łatwo pokazać, że gdy

$$K/M^2 > 2{,}16\ 10^5, \tag{16}$$

wtedy $E_{LL} > E_K$. Spełnienie nierówności (16) jest warunkiem koniecznym (chociaż nie zawsze dostatecznym) istnienia struktury otwartej.



Rys. 2. Domeny cylindryczne

domeny cylindryczne

W obecności zewnętrznego pola magnetycznego, struktura otwarta może przejść niekiedy w strukturę jak na rys. 2. Na rysunku tym widać charakterystyczne domeny o kształcie cylindrycznym. Są to wspomniane wyżej domeny cylindryczne, odkryte i zbadane przez A. H. Bobecka w 1967 r. Zainteresowanie tymi domenami wiąże się z ich małymi rozmiarami i możliwością kontrolowanego ich przesuwania, z czego wynika, że mogą być one wykorzystane do budowy pamięci dla maszyn cyfrowych.

Początkowo domeny cylindryczne obserwowano w ortoferrytach (kryształach tlenkowych o strukturze rombowej, zawierających jony żelaza i lantanowców). Średnica domen cylindrycznych w ortoferrytach wynosi ok. 100 μm. W 1970 r. odkryto domeny cylindryczne w granatach (są to kryształy tlenkowe o strukturze regularnej, zawierające jony żelaza i lantanowców). W granatach średnica domen cylindrycznych zawarta jest w granicach 1–10 μm. Wreszcie w 1973 r. odkryto domeny cylindryczne w cienkich warstwach amorficznych, takich jak np. stopy Gd-Co czy Gd-Fe. Średnica domen cylindrycznych w tych warstwach może być znacznie mniejsza od 1 μm.

energia kryształu z domenami cylindrycznymi

Przeanalizujmy, od czego zależy energia kryształu zawierającego domeny cylindryczne. Sposób postępowania nie różni się tu niczym od analizy rozpatrywanych już modeli. Oprócz rozpatrywanych poprzednio rodzajów energii należy jeszcze uwzględnić energię oddziaływania domen z zewnętrznym polem magnetycznym o natężeniu H:

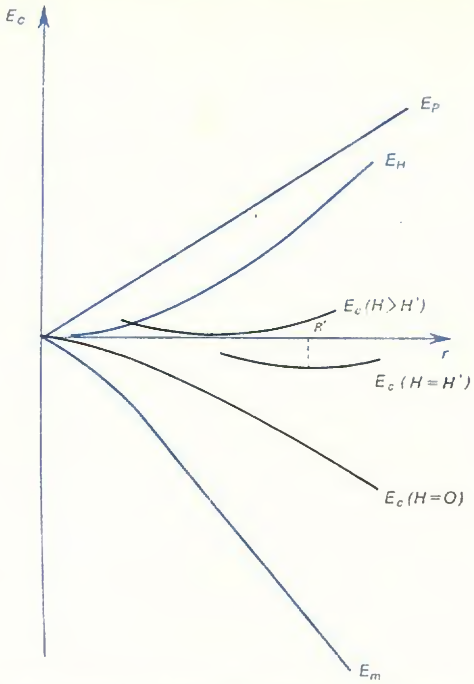
$$E_H = -\int_V (\vec{M}\vec{H})\,dV. \tag{17}$$

W wypadku domen cylindrycznych energia magnetostatyczna ma dość skomplikowaną postać (pojawiają się tu całki eliptyczne). W związku z tym analizę układu przeprowadzimy jakościowo. Energia kryształu zawierającego pojedynczą domenę cylindryczną składa się z 3 części:

$$E_c = E_p + E_m + E_H. \tag{18}$$

zależność energii od promienia domeny

Ponieważ E_H jest proporcjonalne do objętości domeny (o promieniu r) a E_p do jej powierzchni, więc $E_H \sim r^2$, a $E_p \sim r$. Okazuje się, że $E_m \sim r^2$ dla $r \ll h$, natomiast $E_m \sim r$ dla $r \gg h$. Rysunek 3 przedstawia zależność całkowitej energii E_c od promienia domeny r (w dowolnych jednostkach). Jak wynika z wykresu, gdy $H = 0$, to E_c nie ma minimum (energetycznie wygodniejsza jest tu struktura paskowa). Przy pewnym natężeniu pola magnetycznego $H = H'$ energia całkowita ma minimum i pojawiają się domeny o promieniu R'. Ze wzrostem natężenia pola H zmniejsza się promień domeny aż do chwili, gdy



Rys. 3. Zależność energii kryształu zawierającego jedną domenę od promienia domeny r

$$\min E_c(R) = 0.$$

znikanie domeny cylindrycznej

Wtedy domena cylindryczna znika. Ze względu na analogię z zapadaniem grawitacyjnym mówi się, że domena zapada się. Tak więc domeny cylindryczne mogą istnieć tylko w określonym przedziale natężeń pól magnesujących. Również promień domeny cylindrycznej zmienia się przy zmianie H w określonym przedziale. Przedział ten zależy od stosunku σ_w/M^2. W konkretnym materiale magnetycznym nie można zatem wytworzyć dowolnie małej domeny. Stąd ustawiczne poszukiwania nowych materiałów, w których można by wytwarzać domeny mające mniejsze rozmiary.

W urządzeniach wykorzystujących domeny cylindryczne (np. w → Pamięciach magnetycznych) buduje się specjalne układy do detekcji domen. Działanie ich oparte jest na zjawisku zmiany pola rozmagnesowującego w pobliżu domeny cylindrycznej. Na tej zasadzie buduje się hallotronowe lub magnetooporowe detektory domen. W badaniach własności fizycznych domen korzysta się z innych metod.

Metody obserwacji struktury domenowej

Istnieje wiele metod obserwacji struktury domenowej, najczęściej jednak stosuje się metodę figur proszkowych lub metodę magnetooptyczną. Metoda figur proszkowych polega na pokrywaniu powierzchni kryształu specjalnym koloidem zawierającym drobny

metoda figur proszkowych

proszek magnetyczny (np. magnetyt). Proszek ten gromadzi się na granicach domen, w miejscach, gdzie ścianki domenowe wychodzą na powierzchnię. Metoda figur proszkowych daje więc informację jedynie o strukturze domenowej na powierzchni kryształu. Więcej informacji można uzyskać za pomocą metod magnetooptycznych. W metodach tych wykorzystuje się fakt, że stan polaryzacji światła odbitego od powierzchni kryształu lub przechodzącego przez kryształ zależy od kierunku wektora namagnesowania. Uzyskuje się więc informację nie tylko o rozkładzie i kształcie domen, ale również w wielu wypadkach o kierunku wektora namagnesowania w poszczególnych domenach. Występowanie określonego zjawiska magnetooptycznego w odbiciu — efektu Kerra biegunowego, podłużnego i poprzecznego, w transmisji — efektu Faradaya i efektu Cottona–Moutona) zależy od geometrii układu, a więc od wzajemnego położenia wektora falowego światła, wektora polaryzacji światła i wektora namagnesowania (→ Magnetooptyczne zjawiska). Na il. 89–92 (tabl. 23) przedstawiono kilka zdjęć struktury domenowej, uzyskanych za pomocą różnych technik badawczych na kryształach o jednoosiowej anizotropii i to takich, w których występują domeny cylindryczne. We wszystkich przypadkach kierunek łatwego magnesowania jest prostopadły do płaszczyzny płytki. Ilustracja 90a przedstawia zdjęcie struktury domenowej ferrytu barowego $BaFe_{12}O_{19}$ uzyskane metodą figur proszkowych. Widać, że powierzchniowa struktura domenowa jest bardzo skomplikowana. Zdjęcie 90b przedstawia ten sam obszar kryształu badany metodą Faradaya. W przeciwieństwie do struktury powierzchniowej, struktura wewnętrzna jest stosunkowo prosta. Można również zauważyć istnienie korelacji między strukturą powierzchniową i strukturą wewnętrzną. Nie jest to jednak regułą i często struktura powierzchniowa nie wykazuje żadnego podobieństwa do struktury wewnętrznej. Wracając do przedstawionych zdjęć należy podkreślić, że są one unikalne i to z dwóch powodów. Po pierwsze, na zdjęciu 90b widać rzadkie zjawisko istnienia obok siebie domen cylindrycznych o różnych kierunkach wektora namagnesowania (na zdjęciu domeny czarne i białe). Istnieją one bez zewnętrznego pola magnetycznego, stabilizowane polem rozmagnesowującym. A także dlatego, że wykonane zostało na stosunkowo grubym krysztale (ok. 1 mm). Kryształy o takiej grubości są całkowicie nieprzezroczyste dla światła widzialnego. Dlatego do obserwacji domen metodą Faradaya wykorzystano tu podczerwoną część widma. Technika badania struktury domenowej ferrytów heksagonalnych (takich jak $BaFe_{12}O_{19}$) rozwinięta została w Instytucie Fizyki PAN i pozwoliła na obserwację struktury domenowej w płytkach o rekordowej grubości dochodzącej do ok. 0,5 cm.

metody magnetooptyczne

struktura powierzchniowa a wewnętrzna

struktura domenowa granatu

Seria zdjęć 89a, b, c i d przedstawia strukturę domenową (obserwowaną techniką Faradaya) w cienkich warstwach granatu $Y_{1.6}Gd_{0.6}Bi_{0.8}Fe_{3.8}Ga_{1.2}O_{12}$. Zdjęcie 89a przedstawia strukturę w zerowym polu magnetycznym. Zdjęcia 89b, c i d pokazują ten sam odcinek kryształu w polu magnetycznym (skierowanym prostopadle do warstwy) o natężeniu odpowiednio 2170 A/m, 2900 A/m i 3790 A/m). Widać wyraźnie, jak jedne domeny (na zdjęciach — białe) rosną kosztem domen mających przeciwną polaryzację i jak struktura paskowa przechodzi w strukturę cylindryczną.

domeny cylindryczne w granacie

Ilustracja 91 (tabl. 23) przedstawia zdjęcie siatki domen cylindrycznych w granacie $Y_{2.5}Gd_{0.3}Fe_4GaO_{12}$. Siatkę taką wytwarza się przykładając prostopadle do próbki pulsujące pole magnetyczne o określonym natężeniu. Domeny cylindryczne tworzą tu dwuwymiarową heksagonalną siatkę.

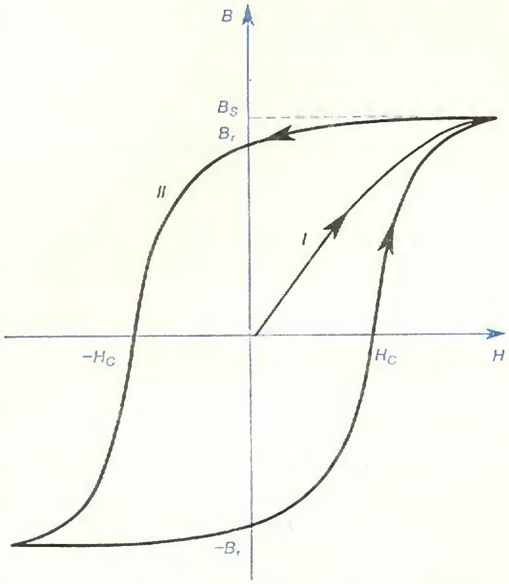
Ilustracja 92 (tabl. 23) przedstawia zdjęcie struktury domenowej materiału najbardziej interesującego z punktu widzenia wykorzystania domen cylindrycznych w technice — cienkich warstw amorficznych (stop Gd-Co). Ponieważ warstwa jest całkowicie nieprzezroczysta dla światła, do obserwacji struktury wykorzystano tu efekt Kerra (biegunowy).

struktura domenowa stopu Gd-Co

Procesy magnesowania

Obecność struktury domenowej powoduje, że bez zewnętrznego pola magnetycznego kryształ jest w zasadzie rozmagnesowany, tzn. wypadkowy moment magnetyczny jest równy zeru. Przyłożenie zewnętrznego pola magnetycznego powoduje powstanie niezerowego wypadkowego namagnesowania, które rośnie w miarę zwiększania pola magnetycznego, aż osiąga wartość M_0 równą namagnesowaniu jednorodnego, bezdomenowego kryształu. Indukcja magnetyczna osiąga przy tym wartość B_s (rys. 4) — zwaną indukcją nasycenia. Cykliczna zmiana pola magnetycznego wywołuje zmianę namagnesowania i indukcji po zamkniętej krzywej (krzywa *II* z rys. 4) zwanej pętlą histerezy. Pętlę histerezy charakteryzują dwa podstawowe parametry: indukcja szczątkowa (lub pozostałość magnetyczna) B_r określona jako wartość B w punkcie $H = 0$; siła koercji H_c (lub pole koercji, natężenie powściągające) określona jako wartość pola H, przy którym $B = 0$. Z wymienionych powyżej parametrów B_s, B_r i H_c tylko B_s jest parametrem całkowicie zależnym od materiału. Pozostałe parametry zależą od obróbki technologicznej materiału, od rodzaju i rozkładu defektów, od parametrów struktury domenowej itd. Można więc je zmieniać. Stąd też podstawowe zadanie teorii procesów magnesowania polega na zrozumieniu mechanizmów determinujących parametry B_r i H_c — parametry, które mają istotne znaczenie przy praktycznym wykorzystaniu materiału magnetycznego.

indukcja nasycenia



Rys. 4. Pętla histerezy indukcji magnetycznej; B_s indukcja nasycenia, B_r indukcja szczątkowa; H_c pole koercji

indukcja szczątkowa i pole koercji

Istnieją dwa zasadnicze procesy wpływające na kształt krzywej magnesowania: procesy przesuwania ścian domenowych i procesy rotacji wektora namagnesowania w domenach. W pierwszym procesie ruch ścian domenowych prowadzi do powiększenia objętości domen, w których namagnesowanie skierowane jest zgodnie z kierunkiem zewnętrznego pola magnetycznego. W drugim procesie nie zachodzi w zasadzie zmiana objętości domen, natomiast wektor namagnesowania w domenach obraca się w kierunku zewnętrznego pola magnetycznego. Oba procesy mogą zachodzić w sposób odwracalny lub nieodwracalny.

kształt krzywej namagnesowania

materiały magnetyczne miękkie

Ze względu na charakter procesów magnesowania materiały magnetyczne dzielą się na dwie duże grupy: materiały magnetycznie miękkie i materiały magnetycznie twarde. Do grupy materiałów magnetycznie miękkich zaliczamy ferro- i ferrimagnetyki, których magnesowanie lub przemagnesowanie zachodzi w stosunkowo słabych polach magnetycznych. Materiały te mają dużą wartość indukcji nasycenia, małe pole koercji, dużą przenikalność magnetyczną i małe straty na przemagnesowanie. Do podstawowych materiałów tej grupy należą: żelazo, stale krzemowe, stopy żelaza z niklem, niektóre ferryty. Do grupy materiałów magnetycznie twardych zalicza się ferro- i ferrimagnetyki, w których magnesowanie zachodzi w stosunkowo silnych polach magnetycznych. Materiały te wykorzystywane są najczęściej do wytwarzania magnesów trwałych. Muszą one mieć dużą wartość indukcji szczątkowej oraz duże pole koercji. W ostatnich latach i w tej grupie materiałów dokonano ważnych odkryć. Odkryto nowe, bardzo obiecujące materiały do wytwarzania magnesów trwałych. Są to związki międzymetaliczne typu MCo_5 (M — lantanowiec, najczęściej Sm, Pr, La, Nd). Związki te swoimi parametrami znacznie przewyższają stosowane dotychczas materiały do wytwarzania magnesów trwałych. Np. dla $SmCo_5$ pole koercji $H_c = 800$ kA/m a $B_s = 1,0$ Vs/m². Tak duże wartości H_c związane są z gigantyczną anizotropią magnetokrystaliczną (i ewentualnie magnetosprężystą) tych materiałów. Natura anizotropii w tych związkach nie została dotychczas w pełni wyjaśniona. Dużą wartość H_c uzyskuje się po odpowiedniej obróbce technologicznej materiału (prasowanie proszków w silnych polach magnetycznych). Omawianych związków używa się do wyrobu miniaturowych magnesów trwałych, koniecznych do radykalnego zminiaturyzowania sprzętu radiotechnicznego, elektrotechnicznego itp.

materiały magnetyczne twarde

procesy magnesowania w MCo_5

Większość prac naukowych prowadzonych obecnie w zakresie fizyki domen magnetycznych dotyczy, obok wspomnianych zagadnień domen cylindrycznych, właśnie problemu procesów magnesowania w związkach MCo_5. Badania te obejmują m.in. obserwacje struktur domenowych w tych materiałach i poszukiwanie korelacji między strukturą domenową, parametrami technicznymi materiału (głównie polem koercji) i obróbką technologiczną. Chodzi tu również o wyjaśnienie mechanizmów hamowania rotacji wektora namagnesowania w związkach MCo_5.

domeny w ferroelektrykach

Na zakończenie należy jeszcze raz podkreślić, że domeny magnetyczne występują zawsze tam, gdzie istnieje możliwość obniżenia energii układu przy przechodzeniu od konfiguracji jednorodnego nasycenia z wysoką energią do konfiguracji domenowej z niższą energią. A zatem można oczekiwać, że struktura domenowa występuje również i w innych materiałach. I rzeczywiście, podobne własności do domen w magnetykach mają domeny w ferroelektrykach. Strukturę domenową obserwuje się w nadprzewodnikach I rodzaju i w stanie przejściowym przy przejściach fazowych I rodzaju. Mechanizm powstawania domen w wymienionych przypadkach jest bardzo podobny (w ferroelektrykach — identyczny) do opisanego mechanizmu powstawania domen w magnetykach. Widzimy więc, że pojęcia domeny nie należy wiązać jedynie z materiałami magnetycznymi. Ma ono znaczenie o wiele bardziej uniwersalne.

Domeny cylindryczne, I Szkoła Zimowa „Nowe Materiały Magnetyczne", Warszawa 1976; C. Kittel *Wstęp do fizyki ciała stałego*, Warszawa 1974; A. H. Morrish *Fizyczne podstawy magnetyzmu*, Warszawa 1970; M. Nałęcz (red.) *Cylindryczne domeny magnetyczne w technice cyfrowej*, Warszawa 1973; R. Wadas *Ferrimagnetyzm*, Warszawa 1968.

Magnetooptyka

Wiesław Wardzyński

Magnetooptyka jest działem nauki o zjawiskach optycznych, zajmującym się wpływem pola magnetycznego na rozchodzenie się promieniowania elektromagnetycznego w ośrodku. Zjawiska magnetooptyczne wiążą się zatem ze wzajemnym oddziaływaniem promieniowania elektromagnetycznego i materii. Zjawiska te stanowią istotne źródło informacji o oddziaływaniach promieniowania elektromagnetycznego z materią, a więc o specyficznych właściwościach tych ciał, w których promieniowanie się rozchodzi.

zakresy promieniowania

Z punktu widzenia charakteru promieniowania elektromagnetycznego rozróżniamy zjawiska magnetooptyczne w zakresie mikrofal, promieniowania podczerwonego oraz w zakresie światła widzialnego i promieniowania pozafioletowego. Różnice w zjawiskach magnetooptycznych w tych obszarach promieniowania wiążą się ze specyficznymi cechami tego promieniowania, wynikającymi z częstości drgań, a co za tym idzie — energii fotonów.

rodzaje ośrodka

Z punktu widzenia ośrodka, w którym rozchodzi się światło, możemy rozróżniać zjawiska magnetooptyczne w gazach, cieczach i ciałach stałych. Różnice w zjawiskach magnetooptycznych w tych ośrodkach wynikają z ich charakterystycznych cech. Najbardziej interesujące i różnorodne zjawiska magnetooptyczne obserwuje się w ciałach stałych ze względu na ich regularną, krystaliczną budowę oraz różnice we właściwościach elektrycznych i magnetycznych, które decydują o oddziaływaniu z promieniowaniem elektromagnetycznym. Mówiąc o zjawiskach magnetooptycznych bardzo często mamy właśnie na myśli zjawiska zachodzące w ciałach stałych.

Fale elektromagnetyczne o określonych częstościach przechodząc przez ośrodek mogą być tłumione (pochłaniane), przy czym stopień tłumienia w ciałach krystalicznych zależy od polaryzacji fali. Fale o różnej polaryzacji rozchodząc się w kryształach biegną z różnymi prędkościami w określonych kierunkach krystalograficznych. Zjawiskami związanymi z rozchodzeniem się światła w kryształach zajmuje się odrębny dział optyki — zwany optyką kryształów; znajomość tych zjawisk jest niezbędna dla zrozumienia zjawisk magnetooptycznych.

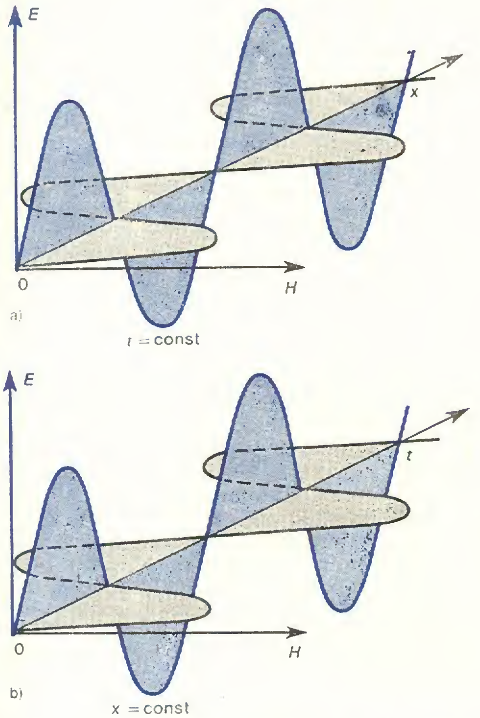
zjawiska magnetooptyczne

Znamy wiele zjawisk magnetooptycznych wiążących się z wpływem zewnętrznego pola magnetycznego tak na amplitudę drgań fali biegnącej w ośrodku umieszczonym w tym polu, jak i na prędkość rozchodzenia się fali o określonej polaryzacji w takim ośrodku. Pierwsze zjawisko magnetooptyczne odkrył w r. 1845 M. Faraday, obserwując skręcenie płaszczyzny polaryzacji światła biegnącego w szkle w kierunku równoległym do przyłożonego zewnętrznego pola magnetycznego. Doświadczenia podobne powtarzane następnie z innymi materiałami, m.in. z gazami, doprowadziły do odkrycia w r. 1896 zjawiska Zeemana. Polega ono na zmianach pochłaniania lub emisji promieniowania w badanym materiale pod wpływem zewnętrznego pola magnetycznego (rozszczepienie pod wpływem pola magnetycznego linii emisyjnych lub linii absorpcyjnych gazu). Te zjawiska magnetooptyczne mogły być opisane i wyjaśnione na podstawie klasycznej dynamiki elektronu w polu magnetycznym, przy pomocy teorii zjawisk elektromagnetycznych sformułowanej przez Maxwella oraz teorii dyspersji rozwiniętej przez H. Lorentza, P. Drudego, W. Voigta i in. w początkach XX w. Dalszy istotny postęp w badaniach magnetooptycznych nastąpił dopiero w latach pięćdziesiątych XX w., kiedy

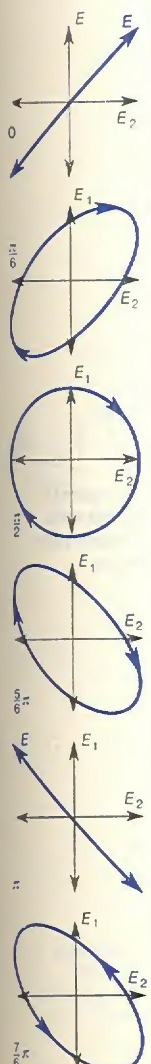
to badania objęły materiały o właściwościach przede wszystkim półprzewodnikowych, a następnie magnetycznych, do interpretacji zaś obserwowanych zjawisk użyto kwantowej teorii i wynikającego z niej opisu struktury pasmowej materiałów. Od tego też czasu magnetooptyka stała się jedną z podstawowych metod badawczych struktury elektronowej ciał stałych.

polaryzacja światła

Ponieważ w zjawiskach magnetooptycznych istotną rolę odgrywa polaryzacja światła, przypomnijmy pokrótce najbardziej istotne wiadomości na ten temat. Fala elektromagnetyczna polega na rozchodzeniu się periodycznych w czasie i przestrzeni zaburzeń elektrycznych i magnetycznych w postaci periodycznych zmian natężenia pola elektrycznego $\vec{E}$ i magnetycznego $\vec{H}$ (wektory $\vec{E}$ i $\vec{H}$ są względem siebie prostopadłe).



Rys. 1. Zmiany natężenia pola elektrycznego $\vec{E}$ i magnetycznego $\vec{H}$: a) w określonej chwili wzdłuż drogi, po której rozchodzi się fala i b) w czasie w określonym miejscu przestrzeni



Rys. 2.

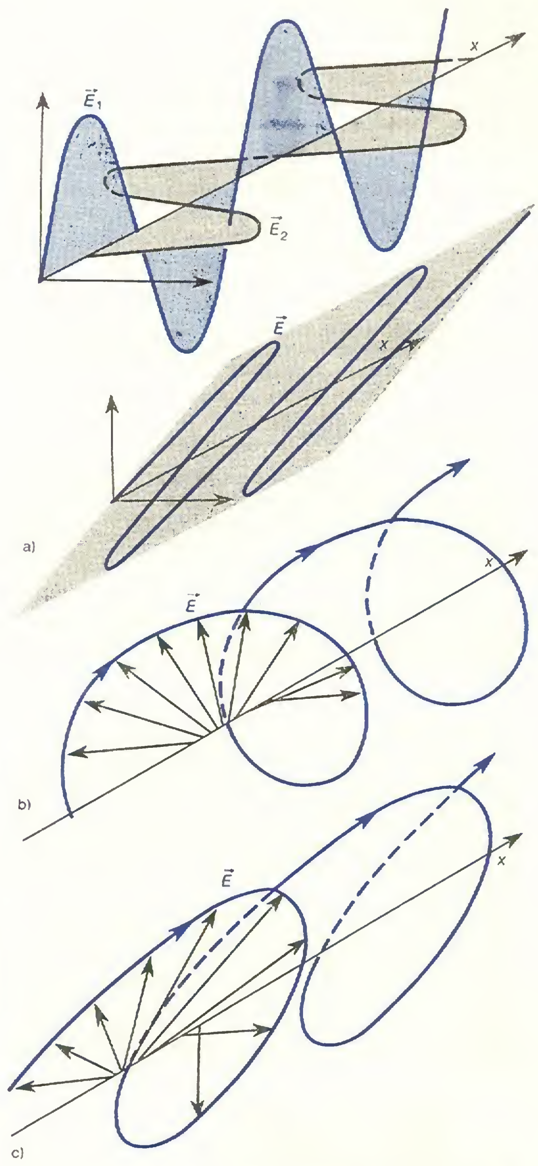
Obrazowo falę elektromagnetyczną można przedstawić posługując się wykresami natężenia pola elektrycznego lub magnetycznego bądź w ustalonym czasie dla różnych punktów drogi, wzdłuż której fala się rozchodzi, bądź też w ustalonym punkcie przestrzeni w funkcji czasu. Wykresy takie w wypadku fali sprężystej przedstawiają wychylenie z położenia równowagi drgających cząstek widziane w określonej chwili wzdłuż linii rozchodzenia się fali (periodyczność fali w przestrzeni — przestrzenny obraz fali), bądź też wychylenie wybranej cząstki w czasie (periodyczność fali w czasie — czasowy obraz fali). Aby sobie wyobrazić zmiany zachodzące przy rozchodzeniu się fali, należy nałożyć w wyobraźni na siebie obraz przedstawiający zmiany w czasie i obraz zmiany w przestrzeni. Analogia z falą sprężystą może być tu pomocna — łatwiej wyobrazić sobie wychylenia cząstek niż zmiany pola — należy jednak pamiętać, że jest to tylko analogia.

Fala spolaryzowana to taka fala, w której drgania zachodzą w sposób uporządkowany. Jeśli wektor natężenia pola elektrycznego w wiązce światła drga w określonej płaszczyźnie (analogia z drganiami cząstek w określonej płaszczyźnie), to taką falę świetlną nazywamy falą spolaryzowaną liniowo (rys. 1a, b).

W wyniku nałożenia się dwóch fal o jednakowych amplitudach, spolaryzowanych liniowo, których wektory elektryczne drgają w płaszczyznach wzajemnie prostopadłych, otrzymamy falę spolaryzowaną liniowo, kołowo lub eliptycznie, w zależności od różnicy faz δ obydwóch fal. Jeśli różnica faz wynosi 0, π, 2π, ... (wielokrotność π), wówczas fala wypadkowa jest falą spolaryzowaną liniowo, przy czym drgania odbywają się w płaszczyźnie tworzącej kąt $+45°$ lub $-45°$ z płaszczyznami drgania fal składowych. Jeśli różnica faz wynosi $^1/_2\pi$, $^3/_2\pi$, ... (nieparzysta wielokrotność $\pi/2$), wówczas fala jest spolaryzowana kołowo. Przy pozostałych różnicach faz mamy falę spolaryzowaną eliptycznie. Sytuację przedstawia poglądowo rys. 2, a obraz fali wypadkowej spolaryzowanej liniowo, kołowo i eliptycznie — rys. 3abc (tylko drgania wektorów $\vec{E}$). W fali spolaryzowanej kołowo (rys. 3b) wektor natężenia pola elektrycznego zmienia kierunek wzdłuż drogi, po której się rozchodzi fala, ale jest zawsze prostopadły do kierunku rozchodzenia się fali; koniec wektora porusza się po linii śrubowej. Patrząc wzdłuż kierunku rozchodzenia się fali stwierdzamy, że koniec wektora natężenia pola może zachodzić po obwodzie koła. Ruch ten może zachodzić bądź w prawo (zgodnie z ruchem wskazówek zegara), bądź w lewo (przeciwnie do ruchu wskazówek zegara). Mówimy o fali spolaryzowanej kołowo, w prawo lub w lewo.

nałożenie fal spolaryzowanych liniowo

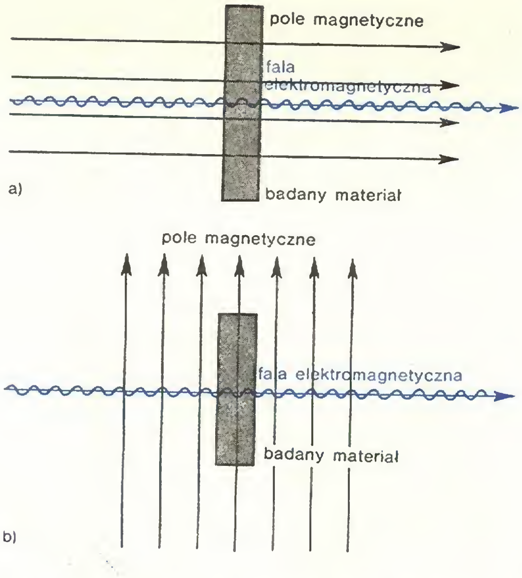
Każdą falę spolaryzowaną liniowo można zastąpić przez dwie fale spolaryzowane kołowo w prawo i lewo, przy czym fale te mają takie same fazy.



Rys. 3. Fale wypadkowe powstałe z nałożenia dwóch fal (o wektorach $\vec{E}_1$ i $\vec{E}_2$) spolaryzowanych liniowo: a) fala spolaryzowana liniowo, b) fala spolaryzowana kołowo, c) fala spolaryzowana eliptycznie

konfiguracja Faradaya i Voigta

Zjawiska magnetooptyczne obserwowane są przy określonym kierunku rozchodzenia się energii rozpatrywanej fali w stosunku do kierunku pola magnetycznego. Rozróżniamy dwie zasadnicze konfiguracje: kierunek rozchodzenia się energii i kierunek pola magnetycznego są do siebie równoległe (konfiguracja Faradaya, rys. 4a) albo kierunek rozchodzenia się energii i kierunek pola magnetycznego są do siebie wzajemnie prostopadłe (konfiguracja Voigta, rys. 4b).



Rys. 4. Schemat układu do obserwacji zjawisk magnetooptycznych: a) konfiguracja Faradaya, b) konfiguracja Voigta

Zjawisko Faradaya

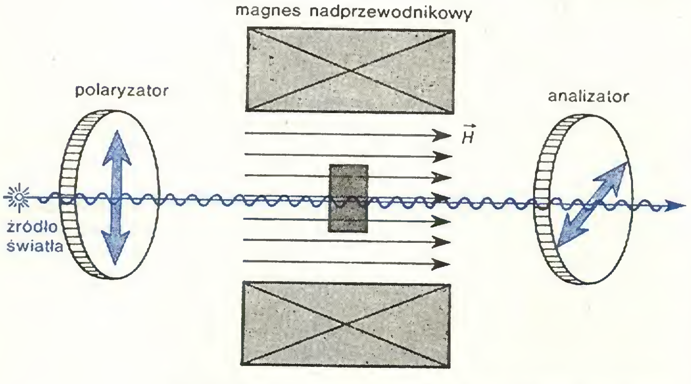
skręcenie płaszczyzny polaryzacji

Zjawisko Faradaya polega na skręceniu płaszczyzny polaryzacji światła spolaryzowanego liniowo, biegnącego w substancji umieszczonej w silnym polu magnetycznym, przy czym światło biegnie wzdłuż linii sił pola magnetycznego (rys. 5). Kąt skręcenia płaszczyzny polaryzacji θ jest proporcjonalny do długości drogi l, którą światło przebiega w ośrodku skręcającym, znajdującym się w polu magnetycznym, oraz do natężenia pola magnetycznego H:

$$\theta = klH.$$

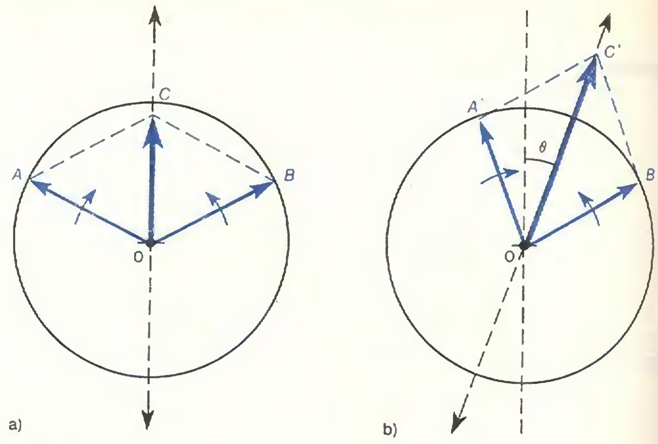
Współczynnik k (stała Verdeta) charakteryzuje zdolność danej substancji do skręcania płaszczyzny polaryzacji w polu magnetycznym. W materiałach magnetycznych występują silne pola wewnętrzne, które są tylko porządkowane przez pola zewnętrzne — kąt skręcenia zależy w tym wypadku głównie od stopnia namagnesowania ciała badanego a nie od pola zewnętrznego.

Skręcenie płaszczyzny polaryzacji wyjaśniamy w sposób następujący: kołowo spolaryzowane fale (na które można rozłożyć falę spolaryzowaną liniowo) biegną w ośrodku z różnymi prędkościami, wskutek czego po wyjściu z ośrodka pojawia się między nimi pewna różnica faz, zależna od drogi, jaką światło przebyło w rozpatrywanym ośrodku, oraz od różnicy współczynników załamania n_l i n_p fal spolaryzowanych kołowo w lewo i w prawo.



Rys. 5. Schemat układu pozwalającego obserwować zjawisko Faradaya

Sytuację w punkcie wejścia i wyjścia fali z ośrodka przedstawia rys. 6. W tym wypadku drganie prawoskrętne ma większą prędkość. Po dodaniu obu fal



Rys. 6.

spolaryzowanych kołowo po wyjściu z ośrodka otrzymujemy drgania liniowe OC'. Fala spolaryzowana liniowo przy wejściu do ośrodka, w której wektor elektryczny drgał wzdłuż OC doznała skręcenia płaszczyzny polaryzacji o kąt θ. Jeśli skręcenie płaszczyzny polaryzacji (widziane od strony obserwatora, do którego oka wpada światło) jest zgodne z ruchem wskazówek zegara, to przypadek taki nazywamy skręceniem prawoskrętnym lub dodatnim. Kierunek skręcenia płaszczyzny polaryzacji jest taki sam jak kierunek obrotów szybszej z dwóch kołowo spolaryzowanych składowych. Różnica faz δ obu fal równa jest 2θ. Można ją wyrazić przez różnicę współczynników załamania obu fal:

$$\delta = \frac{2\pi l}{\lambda_0}(n_l - n_p),$$

gdzie λ_0 jest długością fali w próżni. Tak więc

$$\theta = \frac{\pi l}{\lambda_0}(n_l - n_p).$$

dyspersja skręcenia magnetycznego

Zjawisko Faradaya tłumaczy się wpływem pola magnetycznego na prędkość rozchodzenia się fali spolaryzowanej kołowo w prawo i w lewo w danym ośrodku. Wielkość skręcenia magnetycznego zależy od długości fali, a więc wykazuje dyspersję.

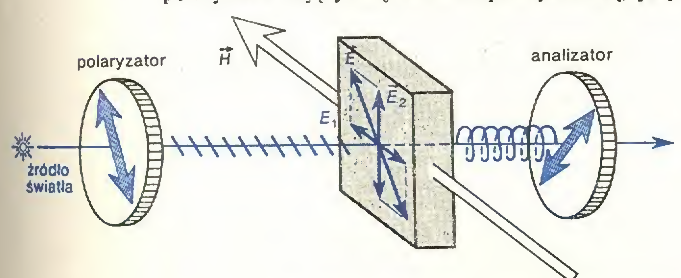
Zjawisko magnetycznego skręcenia płaszczyzny polaryzacji wywołane jest precesją, jaką wykonują elektrony swobodne oraz elektrony wchodzące w skład atomów i cząsteczek substancji pod wpływem zewnętrznego pola magnetycznego. W rezultacie takiej precesji współczynniki załamania (prędkości) promieni światła spolaryzowanych kołowo w prawo i w lewo przyjmują różną wartość. Opis teoretyczny zjawiska Faradaya zależy od tego, czy mamy do czynienia ze swobodnymi nośnikami prądu (np. w półprzewodnikach), czy też z nośnikami związanymi. W pierwszym wypadku mówimy o wewnątrzpasmowym zjawisku Faradaya, w wypadku drugim — o zjawisku Faradaya międzypasmowym.

Magnetyczne skręcenie płaszczyzny polaryzacji pojawia się w bardzo krótkim czasie po wytworzeniu pola magnetycznego (rzędu 10^{-9} s) i równie szybko zanika po usunięciu pola magnetycznego. Obserwacje tego zjawiska prowadzi się w konfiguracji Faradaya, w obszarach widmowych, dla których ośrodek jest przezroczysty. W zależności od materiału obserwuje się je głównie w podczerwieni lub obszarze widzialnym widma promieniowania elektromagnetycznego.

Zjawisko Voigta (Cottona–Moutona)

magnetyczna dwójłomność

Światło, przechodząc przez ośrodek umieszczony w zewnętrznym polu magnetycznym o kierunku linii sił prostopadłym do kierunku rozchodzenia się światła, doznaje podwójnego załamania. Takie magnetyczne podwójne załamanie (magnetyczna dwójłomność), obserwowane w parach i gazach zostało nazwane zjawiskiem Voigta, a identyczne zjawisko obserwowane w cieczach — zjawiskiem Cottona–Moutona. Obecnie największe znaczenie ma badanie tego zjawiska w ciałach stałych, zwłaszcza w półprzewodnikach i magnetykach. Najczęściej nazywamy je w tych materiałach zjawiskiem Voigta. Zjawisko magnetycznego podwójnego załamania w parach i gazach ujawnia się dla światła o częstości bliskiej linii pochłaniania. Opisowe wyjaśnienie tego zjawiska jest następujące. Współczynniki załamania fal liniowo spolaryzowanych o polaryzacji równoległej do kierunku pola magnetycznego i prostopadłej do tego kierunku stają się pod wpływem pola magnetycznego różne. (Fala spolaryzowana prostopadle i równolegle do kierunku pola magnetycznego rozchodzi się z różną prędkością.) W gazach różnica współczynników załamania $n_{\parallel}-n_{\perp}$ jest bardzo mała. Po przejściu w badanym ośrodku drogi l obie fale zyskują pewną różnicę faz, co prowadzi do tego, że fala początkowo liniowo spolaryzowana, po wyjściu z ośrodka staje się falą spolaryzowaną na ogół eliptycznie. Podobnie ciecze pierwotnie optycznie izotropowe stają się pod wpływem pola magnetycznego dwójłomne, przy czym zachowują się jak kryształy jednoosiowe o osi optycznej równoległej do pola magnetycznego. Różnice współczynników załamania fali spolaryzowanej równolegle i prostopadle do pola magnetycznego są znacznie większe, a fala po przejściu drogi w omawianym ośrodku zyskuje znacznie większą różnicę faz aniżeli w gazach. Tak więc po wyjściu z ośrodka może być ona spolaryzowana eliptycznie, kołowo lub liniowo w zależności od uzyskanej przez obie fale różnicy faz. Schemat układu do obserwacji zjawiska Voigta przedstawia rys. 7. Światło przechodzi przez polaryzator dający falę liniowo spolaryzowaną, przy



Rys. 7. Schemat układu pozwalającego obserwować zjawisko Voigta

czym drgania zachodzą w płaszczyźnie tworzącej kąt 45° z kierunkiem zewnętrznego pola magnetycznego $\vec{H}$. Następnie światło pada na badany ośrodek umieszczony w polu magnetycznym o kierunku prostopadłym do kierunku rozchodzenia się światła. W ośrodku falę rozłożyć można na dwie, z których jedna jest spolaryzowana równolegle, a druga — prostopadle do kierunku pola magnetycznego. Obie te fale rozchodzą się w badanym ośrodku z różną prędkością, co prowadzi do powstania różnicy faz i fala po wyjściu z ośrodka ma polaryzację, którą badać można za pomocą analizatora.

zjawisko Voigta w kryształach

W kryształach sytuacja jest bardziej skomplikowana, ponieważ charakter magnetycznego podwójnego załamania zależy od tego, jak przyłożone jest pole magnetyczne w stosunku do osi krystalograficznych kryształu. W kryształach regularnych, gdy pole magnetyczne przyłożone jest równolegle do głównej osi krystalograficznej, obserwujemy podwójne załamanie, jak w kryształach jednoosiowych. Przy innych orientacjach pola magnetycznego kryształ może zachowywać się jak kryształ dwuosiowy. Jeśli kryształ jest anizotropowy i wykazuje już naturalną dwójłomność, zjawisko magnetycznej dwójłomności nakłada się na tę dwójłomność naturalną.

dyspersja zjawiska Voigta

Różnica faz obserwowana w zjawisku Voigta jest proporcjonalna do długości drogi, którą przebiega światło w ośrodku i do kwadratu natężenia pola magnetycznego. Zależy również od długości fali promieniowania, efekt ten bowiem, podobnie jak efekt Faradaya, wykazuje dyspersję. Obserwuje się to zjawisko w obszarach widmowych, dla których ośrodek badany jest przezroczysty.

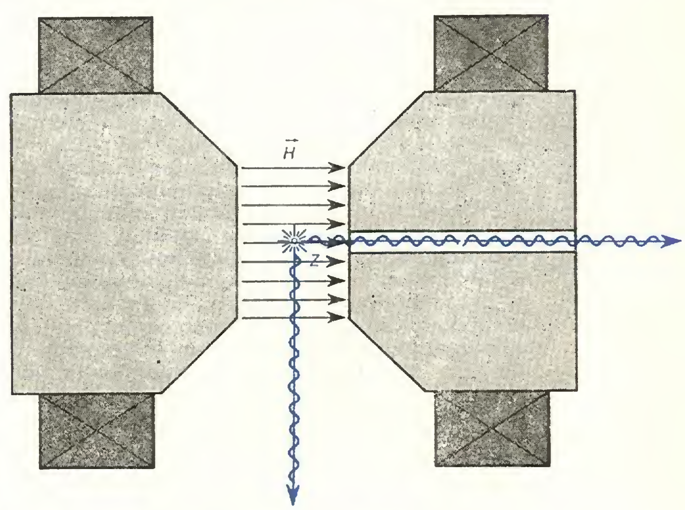
Podkreślić należy, że działanie pola magnetycznego wywołuje jednocześnie pojawienie się zjawiska skręcenia płaszczyzny polaryzacji, gdy światło biegnie w kierunku pola magnetycznego (efekt Faradaya) i dwójłomności magnetycznej, gdy światło biegnie prostopadle do pola. Czyniąc obserwacje w kierunku pola magnetycznego stwierdzamy pojawienie się efektu Faradaya; gdy kierunek obserwacji staje się coraz bardziej zbliżony do prostopadłego do pola magnetycznego, coraz większą rolę odgrywa dwójłomność wywołana polem magnetycznym. W specjalnych sytuacjach możemy pominąć jeden z efektów i brać pod uwagę tylko drugi. Ogólnie rzecz biorąc, gdy nie można pominąć jednego ze zjawisk, efektywna różnica faz wynika z nałożenia się obu zjawisk.

tensor przenikalności magnetycznej

Właściwości optyczne materiału charakteryzuje tensor przenikalności elektrycznej. Przyłożenie pola magnetycznego modyfikuje ten tensor. Tensor przenikalności elektrycznej przedstawić można za pomocą tensora symetrycznego i parzystego ze względu na natężenie pola magnetycznego — odpowiedzialnego za dwójłomność magnetyczną, oraz tensora antysymetrycznego i nieparzystego ze względu na natężenie pola magnetycznego — odpowiedzialnego za skręcenie płaszczyzny polaryzacji. Stąd wynika różna zależność obu efektów od natężenia pola magnetycznego.

Zjawisko Zeemana

Zjawisko Zeemana polega na rozszczepieniu linii widmowych w polu magnetycznym. W r. 1896 P. Zeeman zauważył, że gdy świecące pary sodu (płomień sodowy) umieści się w silnym polu magnetycznym, wówczas żółte linie widmowe, charakterystyczne dla świecących par sodu, ulegają znacznemu poszerzeniu. Dalsze badania prowadzone przy użyciu przyrządów o dużej zdolności rozszczepiającej wykazały, że linie rozszczepiają się w polu magnetycznym na szereg składowych. Schemat urządzenia do badania efektu



Rys. 8. Schemat układu do badania zjawiska Zeemana

Zeemana przedstawia rys. 8. Źródło światła umieszczone jest w polu magnetycznym wytworzonym za pomocą elektromagnesu. Obserwujemy widmo promieniowania wysyłanego bądź w kierunku prostopadłym do linii sił pola magnetycznego, bądź w kierunku równoległym do linii sił. Liczba składowych i rodzaj ich polaryzacji zależy od kierunku obserwacji. W kierunku prostopadłym do linii sił pola magnetycznego obserwujemy rozszczepienie na trzy linie, przy czym są one spolaryzowane liniowo. W kierunku równoległym do pola magnetycznego obserwujemy rozszczepienie na dwie linie, przy czym składowe są spolaryzowane kołowo w kierunkach przeciwnych. Opisane rozszczepienie nazywamy normalnym. Zarówno rodzaj rozszczepienia, jak i rodzaj polaryzacji można wyjaśnić na gruncie klasycznej teorii elektronów.

zjawisko Zeemana normalne

Często jednak występuje rozszczepienie na większą liczbę składowych. Tego rodzaju zjawisko nosi nazwę anomalnego zjawiska Zeemana. Wytłumaczenie anomalnego zjawiska Zeemana możliwe jest tylko na gruncie teorii kwantowej. Zjawisko rozszczepienia linii widmowych pod wpływem pola magnetycznego obserwujemy również w widmie pochłaniania, gdy ośrodek pochłaniający światło umieszczony jest w polu magnetycznym. Czasami zjawisko to nazywamy odwróconym zjawiskiem Zeemana.

zjawisko Zeemana anomalne

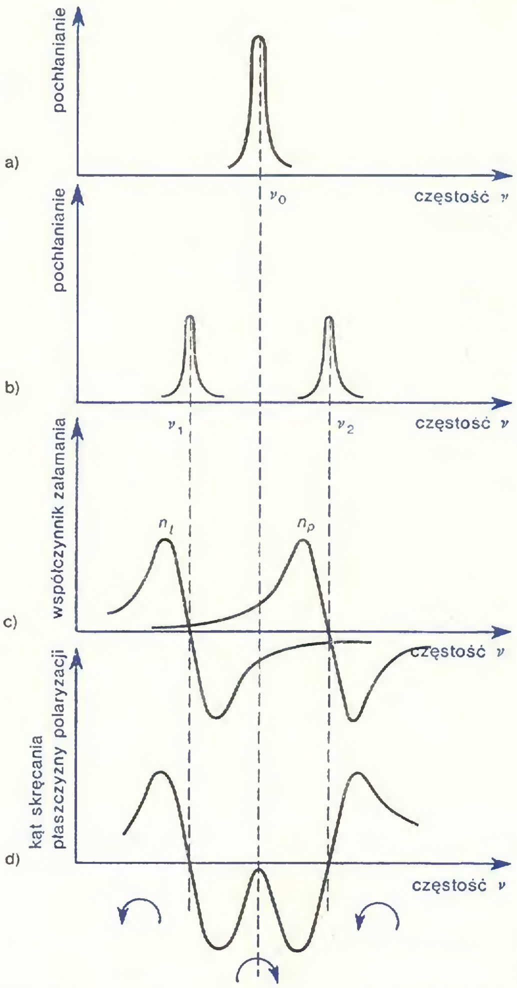
W ciele stałym występuje zjawisko analogiczne do rozszczepienia Zeemana — rozszczepienie poziomów energetycznych pasm (→ Struktura elektronowa ciał stałych) w zewnętrznym polu magnetycznym (tzw. poziomy Landaua). Przejścia pomiędzy poziomami Landaua pasm walencyjnych i pasma przewodnictwa odpowiedzialne są za obserwowane maksima pochłaniania w pobliżu krawędzi absorpcji.

W ciałach stałych zarówno w półprzewodnikach jak i magnetykach istotne znaczenie mają domieszki. Przejścia pomiędzy poziomami domieszek, a także pomiędzy poziomami domieszek i poziomami pasm prowadzą do pojawienia się określonych linii lub pasm pochłaniania, szczególnie w niskich temperaturach. Obserwujemy również w niskich temperaturach bogate widma luminescencji związane z przejściami pomiędzy takimi poziomami. W polu magnetycznym pojawia się rozszczepienie tych linii, co również można uznać za odmianę efektu Zeemana w ciałach stałych. W celu wyjaśnienia wszystkich tych zjawisk konieczny jest opis kwantowy.

związek między zjawiskiem Faradaya i zjawiskiem Zeemana

Pomiędzy zjawiskiem Faradaya i Voigta a zjawiskiem Zeemana istnieje ścisły związek, wynikający z relacji dyspersyjnych, wiążących ze sobą zmiany współczynnika załamania w funkcji długości fali ze zmianami pochłaniania wraz z długością fali. Innymi słowy, znając widmo pochłaniania w całym zakresie widmowym, a więc w zasadzie dla częstości od zera do nieskończoności, możemy (korzystając z zależności dyspersyjnych) obliczyć wartość współczynnika załamania dla dowolnych częstości promieniowania. W praktyce procedurę opisaną stosować możemy przy określonych ograniczeniach dla znacznie węższego obszaru widmowego. Rozpatrzmy najprostszą sytuację — pojedynczą linię widmową. Bez pola magnetycznego występuje pojedyncza linia pochłaniania przy częstości ν_0 związana z przejściem pomiędzy określonymi poziomami w atomie (rys. 9a). W polu magnetycznym, przy obserwacji w konfiguracji Faradaya, linia ta rozszczepia się na dwie o częstościach ν_1 i ν_2, przy czym jedna spolaryzowana jest kołowo w prawo, a druga — w lewo (rys. 9b). Z każdą linią pochłaniania zgodnie z relacją dyspersyjną związane są zmiany współczynnika załamania n_l dla linii ν_1 i n_p dla linii ν_2 (rys. 9c). Z różnicą współczynników załamania $n_l - n_p$ wiąże się, jak wiemy, kąt skręcenia płaszczyzny polaryzacji θ (rys. 9d). Przykład ten wyraźnie wskazuje, że rozszczepienie linii widmowych w polu magnetycznym (efekt Zeemana) prowadzi jednocześnie do pojawienia się — dla światła o długości fali w pobliżu linii pochłaniania — skręcenia płaszczyzny polaryzacji, a więc do efektu Faradaya. W zupełnie analogiczny sposób znajdziemy związek między zjawiskiem Zeemana a zjawiskiem Voigta.

W ciałach stałych, szczególnie w półprzewodnikach, obserwuje się, poza omówionymi, wiele zjawisk magnetooptycznych charakterystycznych dla tych materiałów. Zwróćmy uwagę na dwa spośród nich.



Rys. 9. Zależność od częstości: a) pochłaniania bez pola magnetycznego, b) pochłaniania w polu magnetycznym (obserwacja w konfiguracji Faradaya), c) współczynników załamania fali spolaryzowanej kołowo w prawo i w lewo, d) kąta skręcenia płaszczyzny polaryzacji θ (lub różnicy współczynników $n_l - n_p$)

rezonans cyklotronowy

Rezonans cyklotronowy polega na absorpcji promieniowania elektromagnetycznego (z obszaru mikrofal lub podczerwieni) przez elektrony lub dziury znajdujące się w półprzewodniku, gdy jest on umieszczony w polu magnetycznym.

Według wyobrażeń teorii elektronowej, poruszające się ładunki przebiegają określoną drogę między aktami kolejnych zderzeń. W polu magnetycznym ładunki zakreślają tory będące odcinkami obwodów kół leżących w płaszczyźnie prostopadłej do kierunku pola magnetycznego, przy czym tym większą część obwodu koła zakreśla ładunek, im rzadziej doznaje rozproszenia. Częstość kołowa, z jaką poruszają się w polu magnetycznym ładunki — tzw. częstość cyklotronowa ω_c zależy od natężenia pola magnetycznego, od masy, jaką należy przypisać ładunkom, oraz od wielkości ładunku (dla dziur i elektronów będzie to ładunek elementarny e odpowiednio ze znakiem + lub −):

$$\omega_c = He/cm^*$$

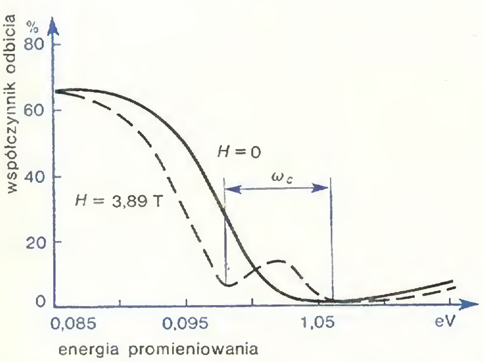
(m^* — masa efektywna).

wyznaczanie masy efektywnej

W półprzewodnikach zarówno dziurom jak i elektronom przypisać należy masę efektywną m_h^* lub m_e^* różną od masy swobodnego elektronu i charakterystyczną dla danego półprzewodnika (→ Dynamika elektronu w ciałach stałych). Kierunek ruchu po obwodzie koła dziur i elektronów jest oczywiście przeciwny. Jeśli na półprzewodnik znajdujący się w polu magnetycznym pada promieniowanie elektromagnetyczne spolaryzowane kołowo (fala mikrofalowa lub podczerwona), to gdy częstość tego promieniowania zbliża się do częstości cyklotronowej, pochłanianie fali elektromagnetycznej gwałtownie rośnie. Maksymalne pochłanianie zachodzi w warunkach rezonansu, gdy częstość fali elektromagnetycznej jest równa częstości cyklotronowej. Tak więc mierząc pochłanianie przy różnych częstościach padającego promieniowania możemy wyznaczyć częstość cyklotronową, a następnie obliczyć (znając natężenie pola magnetycznego) masę efektywną nośników. Co więcej, wiedząc, jakie promieniowanie jest pochłaniane: o polaryzacji prawoskrętnej czy lewoskrętnej, możemy ustalić, czy za pochłanianie odpowiedzialne są nośniki dodatnie (dziury) czy ujemne (elektrony). Efekt cyklotronowy obserwować można jednak tylko w stosunkowo czystych (odznaczających się małą koncentracją nośników) półprzewodnikach i w dostatecznie niskich temperaturach.

odbicie magneto-plazmowe

W półprzewodnikach o dużej koncentracji nośników obserwujemy zjawisko odbicia magnetoplazmowego. Promieniowanie elektromagnetyczne z obszaru podczerwieni jest wówczas pochłaniane przez nośniki swobodne. Współczynnik odbicia przechodzi przez wyraźne minimum, po czym szybko wzrasta tworząc tzw. krawędź plazmową. Położenie minimum odbicia zależy od koncentracji nośników, a także takich parametrów półprzewodnika jak masa efektywna. W po-



Rys. 10. Rozszczepienie krawędzi plazmowej arsenku indu w polu magnetycznym

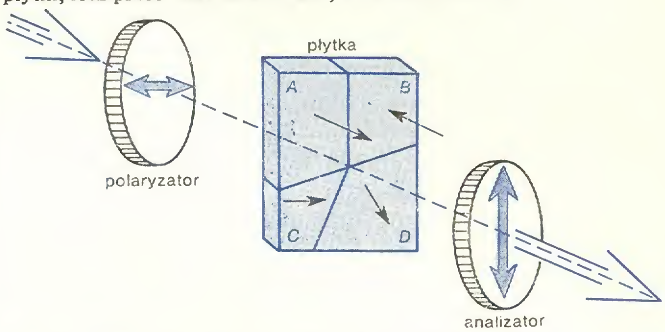
lu magnetycznym minimum to ulega rozszczepieniu na dwa (rys. 10). Odległość między minimami równa jest częstości cyklotronowej. Mierząc tę odległość natychmiast wyznaczamy masę efektywną. Jest to bardzo cenna metoda wyznaczania masy efektywnej w silnie domieszkowanym materiale.

Zastosowanie zjawisk magnetooptycznych w materiałach magnetycznych

W materiałach magnetycznych zjawiska magnetooptyczne pozwalają na obserwacje i badanie rozkładu przestrzennego namagnesowania w kryształach. Jak wiadomo w kryształach takich występują domeny, przy czym w obszarze domeny namagnesowanie ma określoną wartość i kierunek. Postaramy się obecnie pokrótce wyjaśnić, jak wykorzystując omawiane powyżej zjawiska magnetooptyczne, można wyodrębnić poszczególne domeny w krysztale.

badanie domen

Wyobraźmy sobie płytkę krystaliczną zawierającą szereg obszarów, w których wektor namagnesowania jest różnie zorientowany w stosunku do powierzchni płytki. Rozważamy tylko najprostsze wypadki, gdy wektor ten jest bądź prostopadły do powierzchni płytki bądź do niej równoległy. Umieśćmy taką płytkę między skrzyżowanymi polaryzatorem i analizatorem, tak jak na rys. 11. W obszarach A i B wektory namagnesowania są prostopadłe do powierzchni płytki, lecz przeciwnie skierowane, w obszarach C i D

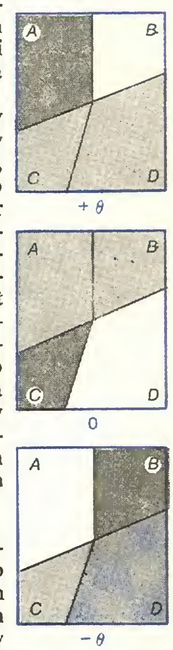


Rys. 11. Schemat doświadczenia umożliwiającego badanie domen w krysztale magnetycznym

wektory te leżą w płaszczyźnie płytki, przy czym w obszarze C wektor namagnesowania jest równoległy do kierunku drgań wektora elektrycznego fali świetlnej przechodzącej przez polaryzator. W obszarze A wektor namagnesowania skierowany jest zgodnie z kierunkiem rozchodzenia się światła i w takim razie płaszczyzna polaryzacji światła przechodzącego przez ten obszar płytki ulegnie skręceniu o kąt $+\theta$. W obszarze B wektor namagnesowania skierowany jest przeciwnie do kierunku rozchodzenia się światła i w tym obszarze płaszczyzna polaryzacji ulegnie skręceniu o kąt $-\theta$. Światło rozchodzące się w obszarze C drga zgodnie z kierunkiem namagnesowania, nie obserwujemy tu zatem ani skręcenia płaszczyzny polaryzacji jak w obszarach A i B (efekt Faradaya) ani magnetycznej dwójłomności. W obszarze D, aczkolwiek wektor namagnesowania jest prostopadły do kierunku rozchodzenia się światła i nie powoduje skręcenia płaszczyzny polaryzacji, to jednak w odróżnieniu od obszaru C tworzy pewien kąt z kierunkiem drgań wektora elektrycznego fali świetlnej i dzięki temu w obszarze tym będziemy mieli do czynienia z dwójłomnością magnetyczną.

Jeśli więc obserwujemy płytkę przez skrzyżowany z polaryzatorem analizator, gdy analizator skręcony jest od pozycji skrzyżowanej dodatkowo o kąt $+\theta$, 0, $-\theta$, to obszary A, B, C i D uwidocznią się jako obszary o różnej jasności (rys. 12). Gdy analizator skręcony jest o kąt $+\theta$, obszar A jest ciemny — skręca on bowiem o taki kąt płaszczyznę polaryzacji, obszar B jest jasny, a obszary C i D wykazują oświetlenie pośrednie. Gdy analizator skręcony jest o kąt $-\theta$, analogicznie obszar B jest ciemny, a obszar A — jasny. Gdy analizator ustawiony jest pod kątem 0 — obszar C jest ciemny, obszary A i B są jednakowo rozjaśnione, natomiast w obszarze D (ze względu na dwójłomność magnetyczną) stopień rozjaśnienia (przy obserwacji w świetle białym także zabarwienie) zależy od grubości płytki i od kąta między wektorem namagnesowania w tym obszarze, a kierunkiem drgań światła wychodzącego z polaryzatora.

Obserwacje omówione powyżej łatwo przeprowadzić za pomocą mikroskopu polaryzacyjnego. Pamiętać jednak należy, że także naprężenia prowadzą do pojawienia się dwójłomności w kryształach, a zatem naprężenia występujące ewentualnie w kryształach mogą w istotny sposób zniekształcać obserwowany obraz.

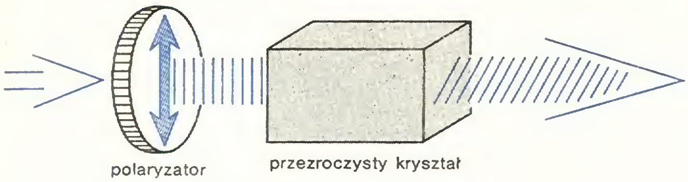


Rys. 12.

Zjawiska magnetooptyczne w materiałach magnetycznych wykorzystywane są szeroko przy budowie różnego typu urządzeń. Omówimy kilka z nich.

rotator

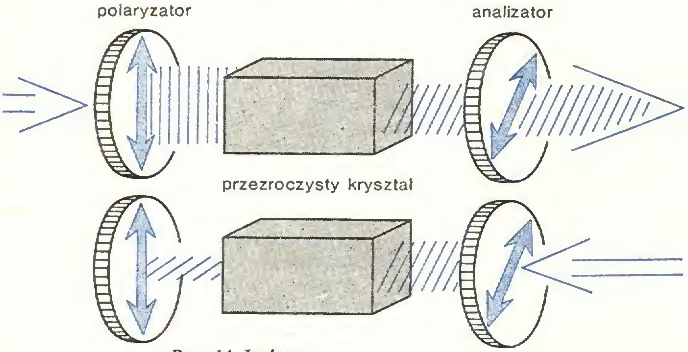
Rotator (rys. 13) — przyrząd obracający płaszczyznę polaryzacji światła spolaryzowanego liniowo o określony kąt. Wektor namagnesowania kryształu jest równoległy do kierunku rozchodzenia się światła (lub ma różną od zera składową namagnesowania w kierunku rozchodzenia się światła). Długość kryształu dobiera się tak, aby przy określonej długości fali płaszczyzna polaryzacji skręcana była o wymagany kąt.



Rys. 13. Rotator

izolator

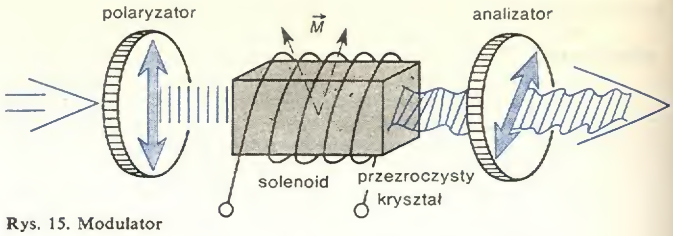
Izolator (lub wentyl, rys. 14) — urządzenie, w którym światło może przechodzić w jednym kierunku. Kryształ skręca płaszczyznę polaryzacji o kąt 45°; analizator jest skręcony względem polaryzatora o kąt 45°. Światło przechodzi z lewa na prawo, nie przechodzi natomiast z prawa na lewo.



Rys. 14. Izolator

modulator

Modulator (rys. 15) — przyrząd zmieniający amplitudę przechodzącego światła za pomocą sterującego go sygnału. Zewnętrzne pole magnetyczne solenoidu, zasilanego sygnałem sterującym modulator, może

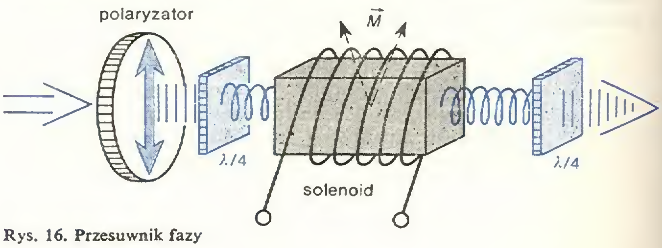


Rys. 15. Modulator

zmieniać wektor namagnesowania $\vec{M}$ kryształu. Zmiana namagnesowania powoduje zmianę skręcenia płaszczyzny polaryzacji, analizator przepuszcza więc falę o zmiennym — modulowanym natężeniu.

przesuwnik fazy

Przesuwnik fazy (rys. 16) — przyrząd zmieniający fazę fali wychodzącej. Spolaryzowane liniowo światło wychodzące z polaryzatora jest w ćwierćfalówce (płytce krystalicznej dającej różnicę faz między promieniem zwyczajnym i nadzwyczajnym, odpowiadającą $\lambda/4$; λ — długość fali) przekształcane na światło spolaryzowane kołowo. Wchodzi ono następnie do kryształu. Zmiana wektora namagnesowania kryształu powoduje zmianę współczynnika załamania fali spolaryzowanej kołowo, a zatem — długości drogi optycznej światła w krysztale. Po wyjściu z kryształu światło spolaryzowane kołowo może być zamienione na światło spolaryzowane liniowo przez drugą ćwierć-



Rys. 16. Przesuwnik fazy

falówkę. Po wyjściu z ćwierćfalówki faza liniowo spolaryzowanej fali zależy od wartości wektora namagnesowania kryształu i może być zmieniana sygnałem doprowadzonym do solenoidu.

Domeny cylindryczne, I Szkoła Zimowa „Nowe Materiały Magnetyczne", Warszawa 1976; J. F. Dillon *Magnetooptical properties of magnetic cristals* w *Magnetic Properties of Materials*, New York 1971; L. Landau, E. Lifszic *Elektrodynamika ośrodków ciągłych*, Warszawa 1960; J. F. Nye *Własności fizyczne kryształów w ujęciu tensorowym i macierzowym*, Warszawa 1962.

Pamięć magnetyczna

Henryk Lachowicz

Pamięć magnetyczna, podobnie jak i pamięci, w których wykorzystuje się inne niż magnetyczne zjawiska fizyczne, np. nadprzewodnictwo lub zjawisko piezoelektryczne, jest urządzeniem zdolnym do przyjmowania informacji, ich przechowywania oraz udostępniania i odtwarzania w postaci nie zmienionej. Pamięć magnetyczna jest najstarszą formą pamięci na Ziemi (starszą niż pamięć gatunku *homo sapiens*), choć świadomość tego człowiek posiadł bardzo niedawno, bo dopiero w dwudziestym stuleciu. Wynikła ona z badań geofizycznych, zwłaszcza zaś z badań zachowania się magnetycznego pola ziemskiego w pradziejach naszego globu. Informacje o tym polu uzyskuje się na podstawie pomiarów paleomagnetycznych. Okazuje się bowiem, że stygnące skały magmowe jak również skały osadowe ze względu na to, że zawierają w sobie minerały ferromagnetyczne, magnesowały się trwale w czasie stygnięcia lub osadzania w taki sposób, że ich namagnesowanie przyjmowało kierunek zgodny z aktualnie istniejącym magnetycznym polem ziemskim

pomiary paleomagnetyczne

oraz wartość proporcjonalną do wartości tego pola. Dzięki temu zjawisku, pomiary namagnesowania próbek skalnych pozwalają wnioskować o wartości i kierunku ziemskiego pola magnetycznego w dalekiej przeszłości geologicznej. Właśnie na tej podstawie odkryto doniosłe, o dużych konsekwencjach, zjawisko zmiany biegunowości tego pola, które wystąpiło po raz ostatni 700 tys. lat temu. Ta liczba doskonale unaocznia wielką trwałość pamięci magnetycznej.

Pamięć dźwięku i obrazu

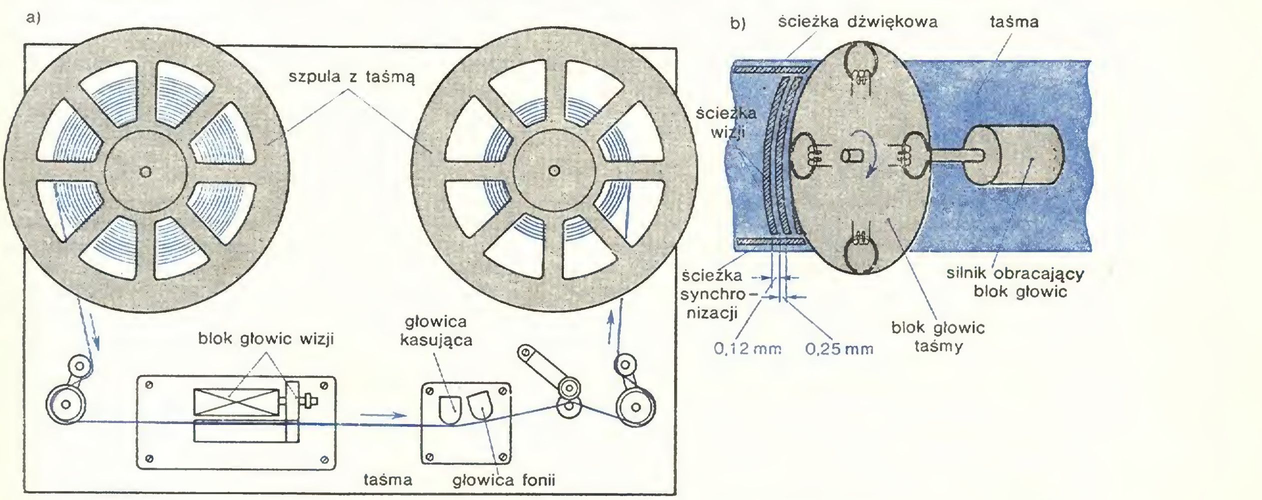
Historia współczesnych pamięci magnetycznych zaczyna się dopiero w końcu ubiegłego wieku, kiedy to O. Smith (1880 r.) wykorzystał do zapisu magnetycznego stalowy drut, co znalazło powszechniejsze zastosowanie przeszło pięćdziesiąt lat później, w radiofonii.

W pamięciach tych wykorzystywane jest w zasadzie to samo zjawisko, które od miliardów lat utrwala nam

zapis sygnałów akustycznych

informacje o zachowaniu się ziemskiego pola magnetycznego. W początkowej fazie rozwój magnetycznych urządzeń pamięciowych koncentrował się na zapisie dźwięku i doprowadził do tak doskonałych już konstrukcji, jak np. magnetofon stereofoniczny. Zapis sygnałów akustycznych, polega w nim na ich odpowiednim odwzorowaniu w postaci dwuwymiarowego (w płaszczyźnie taśmy) rozkładu namagnesowania szczątkowego warstwy magnetycznej, którą pokryta jest taśma magnetofonowa. Wymaga to odpowiedniego przetworzenia tych sygnałów na zmienne pole magnetyczne, proporcjonalne do nich zarówno w częstości jak i amplitudzie. Pole to magnesuje warstwę w taśmie przesuwającej się ze stałą prędkością pod wytwarzającą je głowicą. Przy odczycie rozproszone pole magnetyczne, występujące nad powierzchnią warstwy i odwzorowujące istniejący w niej rozkład namagnesowania, indukuje w uzwojeniu głowicy (w wyniku ruchu taśmy względem głowicy), siłę elektromotoryczną proporcjonalną do szybkości zmian tego pola, a więc i do zapisanego sygnału akustycznego. Charakterystyka namagnesowania warstwy nie jest liniowa (ma postać pętli histerezy), co prowadzi do zniekształceń harmonicznych w czasie zapisu dźwięku. W celu wyeliminowania tego efektu jest zwykle stosowane pole podmagnesowania, wytwarzane przez prąd zmienny o dużej częstości płynący przez uzwojenie głowicy w trakcie zapisu. Częstość tego prądu musi być na tyle duża (40–80 kHz), aby przesuwający się pod głowicą odcinek warstwy zdążył się kilkakrotnie przemagnesować. W latach pięćdziesiątych naszego stulecia pamięć magnetyczną taśmową wykorzystano również do zapisu obrazu zarówno czarno-białego jak i kolorowego. Pierwsze urządzenia potocznie zwane ampeksami (od nazwy firmy AMPEX), pojawiły się w 1956 r. (rys. 1). Sygnał wizyjny niesie znacznie większą ilość informacji w jednostce czasu aniżeli sygnał

zapis obrazu



Rys. 1. Urządzenie do magnetycznego zapisu obrazu (system ampeks): a) mechanizm, b) wirujący blok głowic

foniczny. Odpowiadające temu widmo częstości zawiera się w granicach 50 Hz – 6 MHz (przy zapisie dźwięku przenoszone pasmo zwykle nie przekracza 20 kHz). Największa częstość, jaką urządzenie jest zdolne przenieść, jest m.in. uwarunkowana prędkością przesuwu taśmy. Stosując podobne rozwiązanie jak przy zapisie fonicznym, prędkość przesuwu taśmy musiałaby wynosić ok. 100 m/s, podczas gdy dla dźwięku jest poniżej 1 m/s (np. zapis 10 min audycji telewizyjnej wymagałby aż 60 km taśmy). Trudność tę pokonano, stosując zapis poprzeczny (ścieżki dla sygnału wizji prostopadłe do długości taśmy), w którym dzięki umieszczeniu czterech głowic zamocowanych na obracającym się bębnie (rys. 1b), było możliwe zwiększenie względnej prędkości nośnika magnetycznego w stosunku do głowic. Prędkość obrotowa bębna wynosi 200–240 obr/min (odpowiada to prędkości liniowej głowicy względem taśmy do 30 m/s), prędkość przesuwu taśmy 0,4 m/s. Zapis wizji jest dokonywany z zachowaniem pewnego marginesu na obydwu brzegach taśmy, co jest niezbędne do wyeliminowania zakłóceń powstających w momentach stykania się głowicy z taśmą. Margines ten pozwala jednocześnie na zapisywanie dźwięku oraz impulsów synchronizacji na ścieżkach położonych wzdłuż brzegów taśmy. Przy zapisie wizji, obraz jest zwykle analizowany wg standardowego systemu telewizyjnego, opartego na zasadzie rastru (jedna linia rastru zajmuje w opisanym systemie odcinek o długości ok. 0,4 mm ścieżki na taśmie). Przy rejestracji obrazu kolorowego zapis składa się z 3 ścieżek podstawowych składowych świetlnych sygnału telewizyjnego. Magnetyczny zapis obrazu znajduje najszersze zastosowanie w telewizji, gdzie ułatwia realizację programu (możliwość montażu, wznowienia programu, wymiany programów czy też otrzymywania wielu kopii przez jednoczesny zapis na kilku urządzeniach), bywa również stosowany do celów specjalnych np. na sztucznych satelitach, przy badaniu powierzchni kuli ziemskiej.

Pamięci cyfrowe: ferrytowe i elektromechaniczne

Niezwykle dynamiczny rozwój elektronicznych maszyn cyfrowych — EMC (zwanych powszechnie komputerami), który nastąpił po II wojnie światowej, pobudzał m.in. również i rozwój pamięci magnetycznych, oczywiście przede wszystkim pod kątem ich zastosowania w EMC. Historycznie pierwsza EMC o nazwie ENIAC (z ang. Electronic Numerical Integrator and Calculator) została uruchomiona w 1947 r. w USA i działała z szybkością 5 tys. dodawań na 1 s.

ENIAC — pierwszy komputer

Zasadnicza różnica w działaniu magnetycznej, jak zresztą i każdej innej pamięci cyfrowej, w stosunku do urządzeń przeznaczonych do zapisywania dźwięku lub obrazu, wynika z faktu, że wszelkie operacje w maszynie cyfrowej są dokonywane na liczbach. Posługiwanie się systemem dziesiętnym wymagałoby umiejętności rozróżniania przez każdy układ maszyny dziesięciu stanów odpowiednio przyporządkowanych dziesięciu cyfrom tego systemu (1,2, ... 9,0). Wykorzystując różne zjawiska fizyczne łatwiej jest budować układy wykazujące dwa stabilne stany równowagi (np. prąd płynie — prąd nie płynie, świeci — nie świeci itp.). Z tych też względów w EMC jest powszechnie stosowany system dwójkowy (binarny), w którym każdą cyfrę, a tym samym i liczbę w układzie dziesiętnym, można przedstawić jako ciąg jedynek

system dwójkowy

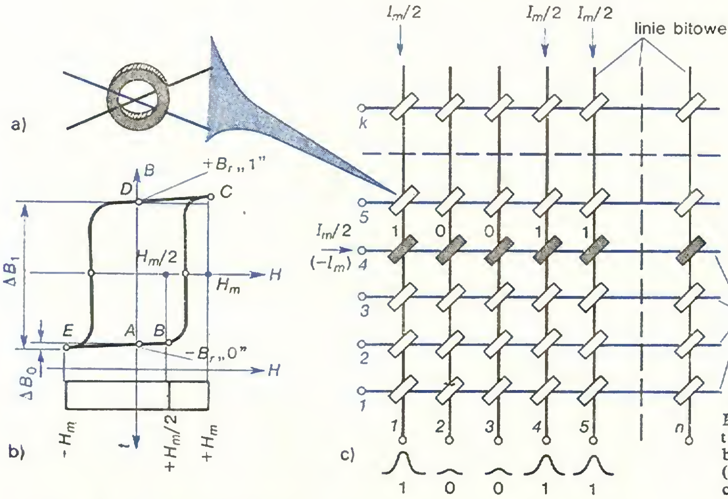
i zer. System lub kod dwójkowy jest utworzony za pomocą dwóch cyfr 0 i 1 umieszczanych na pozycjach, którym są przyporządkowane kolejne potęgi liczby „dwa", np. ciąg $1 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0$, odpowiadający liczbie 19 w systemie dziesiętnym, jest zapisany w systemie dwójkowym jako 10011. Wymaga to od układów pamięci zdolności do rozróżniania tylko dwóch informacji elementarnych 0 i 1 (elementarną jednostkę informacji 0 lub 1 przyjęto nazywać w technice cyfrowej bitem — z ang. *binary digit* — cyfra dwójkowa). W pamięciach magnetycznych informacjom tym są zwykle przyporządkowane dwa stabilne stany namagnesowania, różniące się zwrotem.

pamięć ferrytowa rdzeniowa

Z zasadą działania magnetycznej pamięci cyfrowej zapoznamy się na przykładzie pamięci ferrytowej, w której elementarnymi komórkami pamięciowymi są rdzenie ferrytowe w kształcie pierścionków (toroidów) o średnicach zewnętrznych wielkości dziesiętnych części mm (rys. 2a). Charakterystyka magnesowania takiego rdzenia ma postać prostokątnej pętli histerezy (rys. 2b). Rdzeń jest tym lepszy do użycia w pamięci, im bliższe jedności są jego współczynniki prostokątności i kwadratowości, zdefiniowane stosunkami B_r/B_m i $B(-\frac{1}{2}H_m)/B_m$ (w praktyce osiągane są wartości w granicach 0,9–0,98). W rdzeniu łatwo uzyskać dwa różniące się zwrotem stabilne stany namagnesowania, odpowiadające dodatniej i ujemnej pozostałości magnetycznej ($+B_r$ i $-B_r$). Stanom tym są przyporządkowane odpowiednio cyfry 1 i 0 systemu dwójkowego. Zapis, jak również i odczyt informacji, są dokonywane za pomocą impulsów prądu, płynących przez uzwojenia i nazywane liniami słów i bitów (słowo jest podstawową jednostką obliczeniową danej maszyny i może zawierać np. 256 bitów), które przechodzą przez rdzeń, tworząc płat pamięciowy (rys. 2c). Zakładając, że stan wszystkich rdzeni w wybieranej linii słowa (np. w linii *4*) odpowiada cyfrze 0 (jak przekonamy się stan taki jest rzeczywiście ustalony po każdym odczycie) określoną informację (np. liczbę $19 \equiv 10011$) wpisujemy, pobudzając wszystkie rdzenie w tej linii impulsem pola o amplitudzie równej połowie amplitudy maksymalnej ($\frac{1}{2}H_m$) oraz jednocześnie impulsem pola o tej samej amplitudzie i zwrocie w tych liniach bitowych, w których ma być wpisana cyfra 1. Rdzenie pobudzane tylko polem połówkowym, wytworzonym przez impuls prądu $\frac{1}{2}I_m$, przepływający przez linię słowa, są magnesowane zgodnie z obiegiem $A \rightarrow B \rightarrow A$ na pętli histerezy (rys. 2b) i w chwili zaniku impulsu tego pola w rdzeniu ustala się ponownie stan 0. Rdzenie pobudzane dwoma, sumującymi się impulsami pola, są magnesowane zgodnie z obiegiem $A \rightarrow B \rightarrow C \rightarrow D$ — i ustala się w nich stan 1. Pobudzając teraz rdzenie impulsem pola o pełnej amplitudzie, lecz przeciwnym zwrocie wytworzonym przez impuls prądu ($-I_m$) płynący przez linię słowa na wyjściach linii bitowych (1, 2, ... *n*) pojawią się napięcia odpowiadające zmianom indukcji magnetycznej w poszczególnych rdzeniach zgodnie ze wzorem $u(t) = -A\, dB/dt$ (A — stała). Zmiana indukcji w rdzeniach, które zawierały informację 0, jest niewielka (ΔB_0 — obieg $A \rightarrow E \rightarrow A$) i odpowiadające jej napięcie również małe. W rdzeniach z informacją 1, napięcie jest duże wskutek dużej zmiany indukcji (ΔB_1 — obieg $D \rightarrow E \rightarrow A$). Jak wspominaliśmy, po odczycie wszystkie rdzenie wybranego słowa znajdują się w stanie 0, co umożliwia wpisanie nowej informacji, lub ponowne wpisanie dopiero co odczytanej. Pamięci ferrytowe są powszechnie stosowane w EMC jako pamięci operacyjne (lub wewnętrzne), stanowiące podstawową część jednostki centralnej maszyny cyfrowej. Pojemność pamięci ferrytowych wyrażająca maksymalną liczbę bitów, jaką można jednocześnie przechowywać, zawiera się w granicach od kilkuset bitów do kilku Mbitów. Uzyskiwany w nich czas cyklu, określany czasem, który jest potrzebny do dokonania zapisu słowa, jego odczytu i zwykle ponownego wpisania odczytanej informacji wynosi od 100 ns do kilku µs. Czas ten decyduje o szybkości działania EMC.

wpisywanie informacji

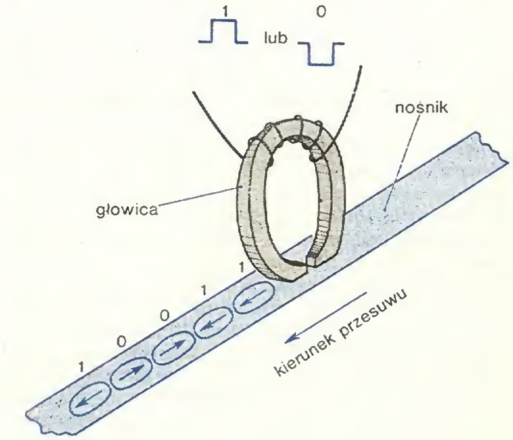
pamięć operacyjna (wewnętrzna)



Rys. 2. Pamięć cyfrowa na rdzeniach ferrytowych: a) toroidalny rdzeń ferrytowy, b) charakterystyka magnesowania rdzenia (prostokątna pętla histerezy), c) płat pamięciowy

urządzenia elektromechaniczne

Dużą grupę pamięci magnetycznych stanowią urządzenia elektromechaniczne, których zasada działania jest podobna do urządzeń do zapisu dźwięku lub obrazu. Informacja cyfrowa jest zapisywana w warstwie magnetycznej, stanowiącej powierzchnię



Rys. 3. Zasada zapisu informacji binarnej na ruchomym nośniku magnetycznym w pamięciach elektromechanicznych

taśmy, bębna lub dysku. Zapis, zwykle wielościeżkowy, jest dokonywany za pomocą głowicy, przez której uzwojenie przepływają impulsy prądu o polaryzacji dodatniej lub ujemnej, zależnie od rodzaju zapisywanej informacji 0 lub 1 (rys. 3). Pole magnetyczne, wytwarzane w szczelinie głowicy magnesuje kolejno obszary przesuwającej się warstwy w kierunku zgodnym lub przeciwnym do kierunku jej ruchu (na rys. 3 w sposób poglądowy przedstawiono zapis liczby 19 ≡ ≡ 10011). W czasie odczytu rozproszone pole namagnesowanych obszarów, zamykając się w obwodzie magnetycznym głowicy, indukuje w jej uzwojeniu (w trakcie ich przesuwania się pod szczeliną roboczą) siłę elektromotoryczną w postaci ciągu impulsów, których polaryzacja jest zależna od zwrotu namagnesowania, a tym samym i od rodzaju informacji dwójkowej.

pamięć zewnętrzna

Pamięci elektromechaniczne, taśmowe, bębnowe i dyskowe są dzisiaj powszechnie stosowane w EMC jako urządzenia zewnętrzne, służące do zapamiętywania tzw. masowej informacji, której zapis wymaga bardzo dużej pojemności, ok. 10^{10} bitów. Pamięci dyskowe i taśmowe mają praktycznie nieograniczoną pojemność ze względu na możliwość wymiany pakietów dysków lub krążków taśmy. Poważnym brakiem tych urządzeń jest natomiast ich stosunkowo mała szybkość działania. Wadą też jest istnienie w tych konstrukcjach dość skomplikowanych zespołów mechanicznych, stwarzających spore kłopoty w czasie eksploatacji. Pamięci ferrytowe i elektromechaniczne, mimo prób wprowadzenia innych pamięci magnetycznych, są jednak nadal bezkonkurencyjne i stosowane w najnowszych generacjach EMC.

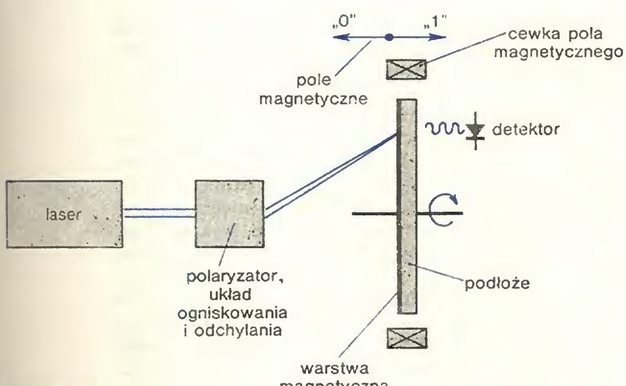
Najnowsze konstrukcje pamięci cyfrowych

pamięć termomagnetooptyczna

Znaczny postęp uzyskany w technice laserowej, a przede wszystkim w konstrukcji laserów półprzewodnikowych, ponownie skupił uwagę na pamięci termomagnetooptycznej, której pomysł i pierwszy model zrodził się już w 1958 r. Nieco później (1967 r.) zwrócono uwagę na możliwość wykorzystania cylindrycznych domen magnetycznych (→ Struktura domenowa i procesy magnesowania) w układach pamięci. Oczekuje się, że pamięci te wyeliminują urządzenia elektromechaniczne głównie ze względu na potencjalną możliwość osiągnięcia w nich bardzo dużych pojemności przy małych rozmiarach geometrycznych (gęstość upakowania informacji rzędu 10^8 bitów/cm² powierzchni nośnika; dla porównania — gęstość tę w mózgu ludzkim ocenia się na 10^{10} bitów/cm³). W obydwu rodzajach pamięci jako nośnik informacji jest stosowana cienka warstwa (grubości rzędu kilku μm) wykazująca anizotropię magnetyczną, przy czym kierunek łatwego magnesowania leży w płaszczyźnie warstwy lub prostopadle do jej powierzchni.

stosowane materiały

Stosowanymi materiałami są bizmutek manganu (MnBi), związki europu (np. EuS lub EuSe) dla pamięci termomagnetooptycznych oraz proste lub mieszane granaty magnetyczne (np. $Sm_{0,5}Y_{2,5}Ga_{1,2}Fe_{3,8}O_{12}$) i warstwy amorficzne stopów metali grupy przejściowej z lantanowcami (np. GdCo) dla obydwu rodzajów pamięci.



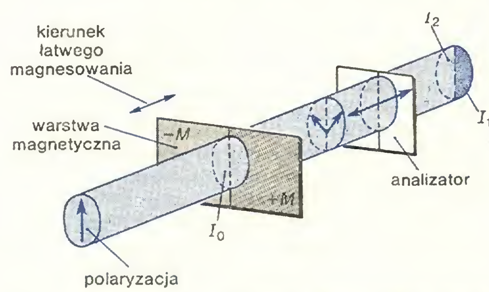
Rys. 4. Schemat blokowy pamięci termomagnetooptycznej z nieruchomym nośnikiem informacji

zapis informacji

W pamięci termomagnetooptycznej zapis jest dokonywany za pomocą silnej wiązki światła laserowego (rys. 4), współdziałającej z polem magnetycznym. Impuls światła — zwykle o czasie trwania mniejszym od 1 μs — pada na wybraną, za pomocą układu odchylenia wiązki, elementarną komórkę pamięciową w ciągłej warstwie nośnika i powoduje jej lokalne nagrzanie do temperatury większej od temperatury Curie materiału (→ Teoria magnetyzmu), w wyniku czego komórka jest termicznie rozmagnesowana. (Powierzchnia tej komórki jest określona przez przekrój, w płaszczyźnie warstwy, zogniskowanej wiązki światła, a uzyskiwane w praktyce wiązki mają średnicę ok. 2 μm). Do otrzymania takich temperatur (rzędu kilkuset °C) używa się zwykle lasera gazowego He-Ne o mocy ok. 100 mW. Współdziałające pole magnetyczne, synchroniczne z impulsami światła, wymusza w czasie stygnięcia materiału w obrębie komórki wymagany zwrot namagnesowania. Pole to ma kierunek prostopadły do powierzchni warstwy (gdy kierunek łatwego magnesowania jest prostopadły do powierzchni warstwy) lub równoległy (gdy kierunek łatwego magnesowania leży w płaszczyźnie warstwy). Zwrot pola jest zależny od rodzaju zapisywanej informacji 0 lub 1, dzięki czemu, po zaniku impulsów światła i pola, namagnesowanie w elementarnej komórce pamięci ma zwrot odpowiadający wpisanej cyfrze dwójkowej.

odczyt informacji

Odczyt zapisanej w ten sposób informacji jest dokonywany za pomocą tej samej wiązki laserowej o świetle liniowo spolaryzowanym, lecz tym razem o wielokrotnie mniejszej energii. Zależnie od rodzaju warstwy magnetycznej, w celu rozróżnienia odczytywanej informacji, wykorzystuje się zjawisko Faradaya lub Kerra (→ Magnetooptyka). Zasadę odczytu



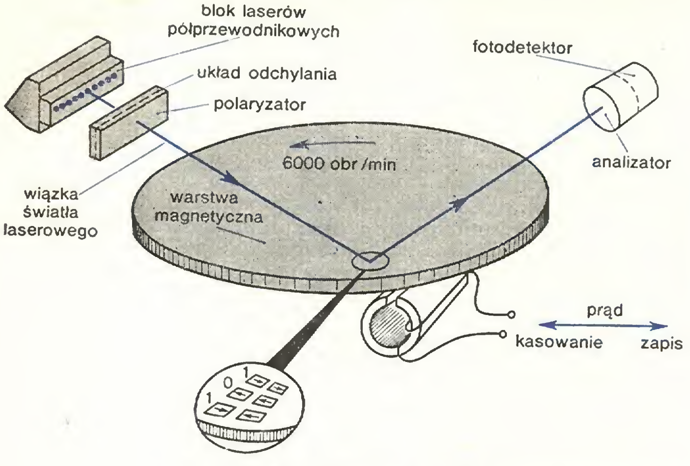
Rys. 5. Zasada odczytu informacji przy wykorzystaniu efektu Faradaya

w wypadku przezroczystej warstwy o prostopadłym kierunku łatwego magnesowania pokazuje rys. 5. Po przejściu przez analizator otrzymuje się wiązki światła o różnym natężeniu I_1 i I_2. Różnica natężeń wiązek jest spowodowana skręceniem płaszczyzny polaryzacji światła (zjawisko Faradaya) przechodzącego przez obszary warstwy o przeciwnych zwrotach namagnesowania, co odpowiada informacjom 0 i 1.

Padając na fotodetektor (np. fotorezystor lub fotodiodę), wiązki I_1 i I_2, niosące odpowiednio informację 0 i 1, generują na wyjściu detektora impulsy elektryczne o zróżnicowanej amplitudzie odpowiednio do różnicy ich natężeń ΔI. W opisanym rozwiązaniu pamięci termomagnetooptycznej największą trudność sprawia właściwe ogniskowanie wiązki laserowej oraz jej odchylanie. Z tych też względów rozwiązanie to, w zasadzie analogiczne do pamięci holograficznej, nie znalazło dotąd szerszego zastosowania.

Większe nadzieje wiąże się z rozwiązaniem (rys. 6), w którym nośnik informacji naniesiony na podłoże w kształcie dysku obraca się, a informacja jest magazynowana wzdłuż koncentrycznych ścieżek (podobnie

jak w pamięciach dyskowych). W tej konstrukcji stosowane są lasery półprzewodnikowe (GaAs), uszeregowane w bloku w ten sposób, że jeden laser przypada na jedną ścieżkę. Przy odczycie wykorzystuje się wiązkę odbitą (efekt Kerra), w związku z czym w tym



Rys. 6. Schemat konstrukcji pamięci termomagnetooptycznej z nośnikiem w formie wirującego dysku

rozwiązaniu stosuje się warstwy o kierunku łatwego magnesowania leżącym w płaszczyźnie warstwy. Pojemność tej pamięci, uwarunkowana głównie precyzją zogniskowania światła laserowego (ok. 2 µm) i odległością między sąsiednimi źródłami światła w bloku (ok. 50 µm), może być rzędu kilkuset Mbitów/dysk, co odpowiada gęstości ok. 200 ścieżek/cm.

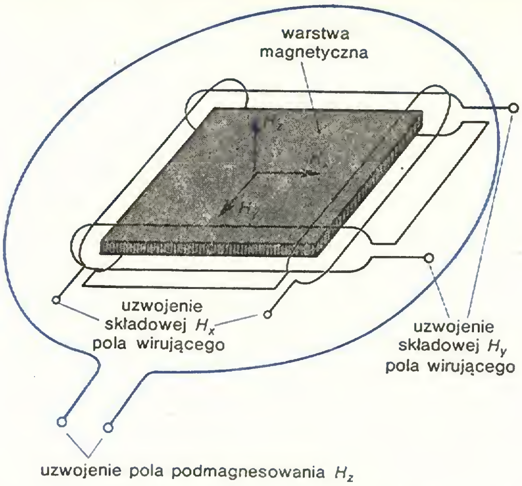
domeny cylindryczne jako nośnik informacji

Najnowszym rozwiązaniem pamięci magnetycznej jest pamięć cyfrowa, w której nośnikiem informacji są magnetyczne domeny cylindryczne wytwarzane w epitaksjalnych (→ Mikroelektronika) warstwach granatów lub warstwach amorficznych stopów typu Gd–Co. Domenie cylindrycznej występującej w określonym miejscu warstwy jest przyporządkowana cyfra 1, jej brakowi w tym miejscu cyfra 0 (wakansja). Odpowiednie ciągi składające się z domen i ich wakansji tworzą liczby (słowa) w kodzie dwójkowym. Domeny cylindryczne mogą się przemieszczać między określonymi punktami na powierzchni warstwy, co jest równoznaczne z przekazywaniem informacji.

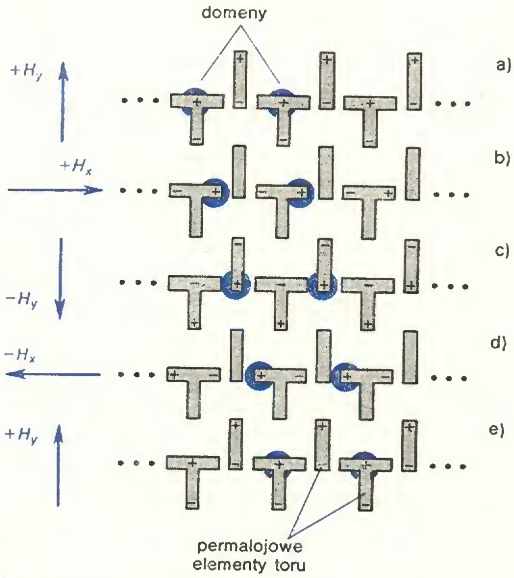
Najmniejsze średnice domen ok. 0,5 µm uzyskano dotychczas w warstwach amorficznych. Ze względu na możliwości współczesnych technik fotolitograficznych ocenia się, że średnica ta nie może być jednak mniejsza niż 0,1 µm. W pierwszych stosowanych w praktyce magnetycznych pamięciach cyfrowych średnica domeny wynosiła 6 um, w nowszych — 3 µm. W rozwiązaniach tych z reguły stosuje się epitaksjalne warstwy granatów. Warstwy amorficzne nie znalazły jednak jeszcze powszechniejszego zastosowania.

pętle pamięciowe

Podstawowym elementem w pamięci magnetycznej jest zamknięta pętla pamięciowa — stanowiąca rejestr szeregowy — do której informacja jest wpisywana, przechowywana i następnie z niej pobierana. Przesuwanie ciągów informacyjnych (składających się z domen i wakansji) w obrębie pętli, niezbędne do przeprowadzania wymienionych operacji, jest realizowane za pomocą torów propagacyjnych wykonywanych zwykle w postaci cienkowarstwowych struktur permalojowych (stop Ni–Fe), naparowywanych przez odpowiednią maskę na warstwę nośnika oraz współpracującego z nimi wirującego w płaszczyźnie warstwy pola magnetycznego. Ruch poszczególnych domen w ciągu następuje w rezultacie oddziaływań między domeną i biegunami magnetycznymi wytwarzanymi w elementach permalojowych toru przez wirujące pole. Źródłem tego pola są sinusoidalne prądy elektryczne płynące przez dwa uzwojenia o wzajemnie prostopadłych osiach, przesunięte w fazie o ćwierć okresu (rys. 7). Elementy toru mają najczęściej kształt liter T, I, X, Y lub krokiewek. Sposób przesuwania domen w torze typu T–I, najczęściej stosowany w praktyce, przedstawia rys. 8, na którym są widoczne kolejne fazy ruchu odpowiadające czterem położe-



Rys. 7. Zasada wytwarzania pola wirującego w płaszczyźnie warstwy i pola podmagnesowania prostopadłego do tej płaszczyzny

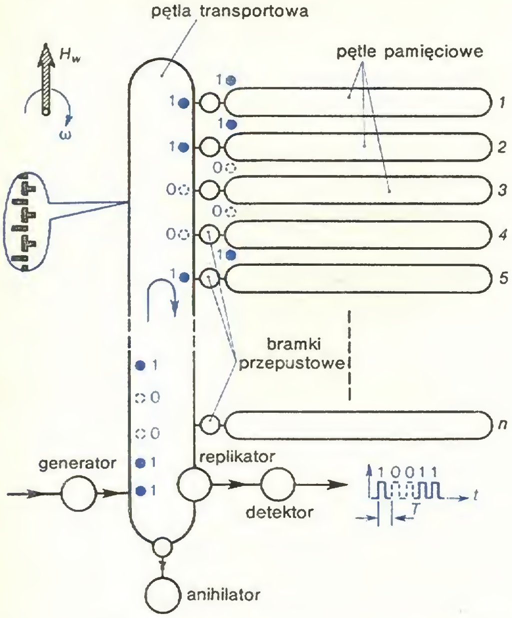


Rys. 8. Sposób przesuwania domen cylindrycznych w torze propagacyjnym typu T–I

niom pola wirującego (rys. 8a, b, c i d). Pełny obrót pola odpowiada przesunięciu o jedno położenie w periodycznej strukturze toru (rys. 8e). Podstawowy układ pamięci (rys. 9), zawiera szereg pętli pamięciowych (do kilkudziesięciu, każda o pojemności rzędu kilkuset bitów), sprzężonych przez bramki przepustowe z pętlą transportową, która służy do wprowadzania lub wyprowadzania informacji z pętli pamięciowych. Bramki przepustowe, kierujące ciągi domen w określone odgałęzienia torów, są zwykle wykonywane również w postaci odpowiednich struktur permalojowych i wbudowanych w nie ścieżek przewodzących, w których kierunek płynącego prądu decyduje o kierunku otwarcia bramki. W czasie zapisu poszczególne bity słowa, wytworzonego przez generator w postaci odpowiedniego ciągu domen (np. 10011 jak to pokazano dla przykładu na rys. 9), są wprowadzane do pętli transportowej i następnie przesunięte w niej w taki sposób, że mogą być jednocześnie wprowadzone do

wprowadzanie słów do pętli

pętli pamięciowych, których liczba odpowiada w zasadzie długości słowa mierzonej w bitach. Poszczególne bity wpisanego słowa są następnie synchronicznie przesuwane w pętlach pamięciowych, krążąc w nich do momentu, w którym przechowywana informacja



Rys. 9. Schemat podstawowego układu cyfrowej pamięci na domenach cylindrycznych

pobieranie informacji

ma być odczytana. Pobranie danego słowa przebiega w odwrotnej kolejności. Odczyt następuje w detektorze (wykorzystującym zjawisko Halla, zjawisko magnetooptyczne lub magnetooporowe), na którego wyjściu pojawia się sekwencja impulsów napięcia odpowiadająca odczytywanej informacji. Umieszczony w torze replikator pozwala na odtworzenie kierowanego do detektora ciągu informacyjnego, co umożliwia dalsze przechowywanie odczytywanej informacji. Sprzężony z pętlą transportową anihilator służy z kolei do niszczenia domen w tych ciągach informacyjnych, które nie są przeznaczone do dalszego pamiętania. Zasada działania anihilatora jest oparta na zjawisku zanikania domeny.

Pojedyncza płytka pamięciowa zawiera jeden lub kilka podstawowych układów. Szereg płytek pamięciowych (do kilkudziesięciu) umieszczonych we wnętrzu układu wytwarzającego pole wirujące i podmagnesowania tworzy podstawowy moduł pamięci, z których można składać układy o wymaganej pojemności. Na il. 98 (tabl. 24) przedstawiono moduł pamięci na domenach cylindrycznych. W istniejących rozwiązaniach tych pamięci osiągana jest gęstość zapisu 10^5–10^6 bitów/cm², szybkość przekazywania informacji 10^6–10^7 bitów/s oraz pojemności do kilku Mbitów.

wykorzystanie siatki domen

W najnowszym rozwiązaniu pamięci na domenach cylindrycznych jest wykorzystywana regularna (heksagonalna) siatka domen, powstająca w sposób naturalny w epitaksjalnej warstwie granatu magnetycznego. Położenia komórek pamięciowych wyznaczają same domeny, gęsto upakowane w warstwie; upakowanie w siatce domen jest 16-krotnie większe w porównaniu z poprzednio omawianym rozwiązaniem, w którym ze względu na magnetostatyczne oddziaływania odległość między sąsiednimi domenami nie może być mniejsza niż cztery średnice domeny. Cyfrom dwójkowym 0 i 1 są przyporządkowane odpowiednio dwa rodzaje domen, różniące się strukturą magnetyczną otaczającej je ścianki. W czasie odczytu jedne z nich poruszają się pod pewnym kątem w stosunku do gradientu wymuszającego ten ruch pola magnetycznego, drugie — równolegle. Pozwala to kierować domeny odpowiadające informacjom 0 i 1 do dwóch, niezależnych torów odczytu. Zapis, polegający na tworzeniu domen o kontrolowanej strukturze ścianki, jest dokonywany za pomocą dwóch koincydencyjnych pól skierowanych prostopadle i równolegle do płaszczyzny warstwy. Najistotniejszą zaletą tego rozwiązania jest bardzo duża stabilność siatki na zmiany warunków zewnętrznych.

Wobec niezwykle dynamicznego rozwoju badań materiałów magnetycznych w ostatnich latach na świecie, nie można wykluczyć i tego, że zanim publikacja ta dotrze do rąk Czytelników, pojawią się pamięci magnetyczne wykorzystujące materiały i zjawiska dziś jeszcze nie znane.

A. J. Meyerhoff (red.) *Cyfrowe zastosowania układów magnetycznych*, Warszawa 1964; M. Nałęcz (red.) *Cylindryczne domeny magnetyczne w technice cyfrowej*, Warszawa 1973; E. Nowak, Z. Sawicki *Pamięci maszyn cyfrowych: konstrukcja i technologia*, Warszawa 1977; B. Urbański *Magnetyczny zapis dźwięków i obrazów*, Warszawa 1965.

Najsilniejsze pola magnetyczne

Czesław Bazan

Pole magnetyczne wpływa w zasadzie na wszystkie zjawiska fizyczne, jest więc jednym z istotnych czynników fizycznego poznania świata. Ponieważ na ogół działanie pola magnetycznego jest tym większe im większe jest jego natężenie bądź też można je zaobserwować dopiero po przekroczeniu pewnych granicznych wartości, fizycy starają się stosować pola coraz silniejsze.

1 T ≙ 10^6 A/m

Wielkość pola określamy podając jego natężenie lub indukcję. Międzynarodowa jednostka natężenia pola: amper na metr (A/m) jest zbyt mała dla naszych celów. Bardziej odpowiednią jest wielkość 10^5 A/m = 1 kA/cm. Jednostką indukcji magnetycznej jest tesla (T). W próżni 1 T odpowiada ok. 10 kA/cm. Dla uproszczenia będziemy się posługiwać jednostką indukcji, pisząc krótko: pole 1 T — zamiast pole o natężeniu 10 kA/cm w powietrzu.

Człowiek bezpośrednio nie odczuwa obecności, a tym bardziej wielkości pola magnetycznego. Dlatego przypomnimy dla porównania wartości odpowiednich pól magnetycznych:

pole między biegunami magnesu podkowiastego — około 0,1 T;

pole magnetyczne Ziemi przy jej powierzchni — poniżej 10^{-4} T;

pola magnetyczne gwiazd — poniżej 5 T.

co to znaczy „silne"?

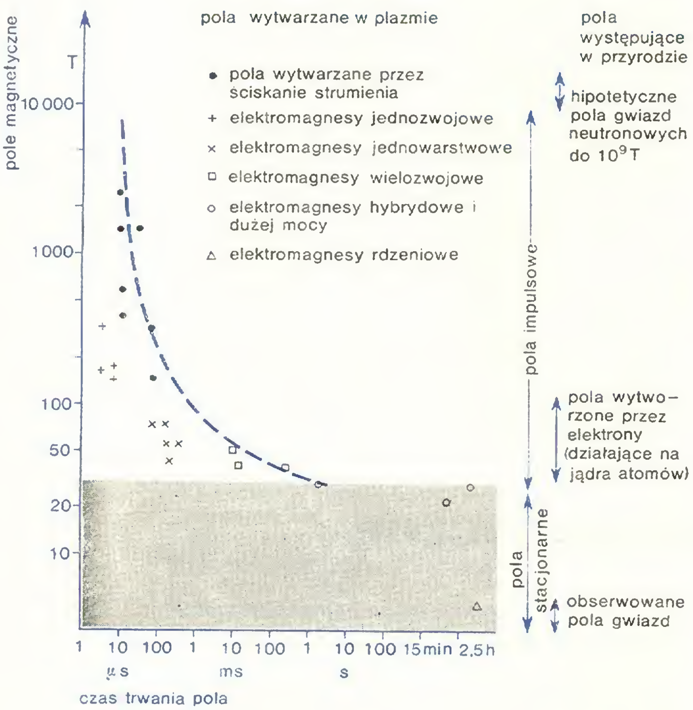
Pola magnetyczne powyżej 5 T uważane są za silne. Wartości silnych pól magnetycznych spotykanych w przyrodzie podano na rys. 1.

sposoby wytwarzania pola magnetycznego

Pola magnetyczne możemy wytwarzać trzema sposobami: a) używając magnesów trwałych; b) wyzyskując przepływ prądu elektrycznego np. przez przewodnik; c) łącząc metody a i b. Ponieważ w metodzie a i c stosuje się ferromagnetyki, wartość otrzymywanego pola jest ograniczona wartością namagnesowania nasycenia. Dalszy wzrost pola można uzyskać przez zwiększanie masy magnesu lub rdzenia, przy czym wzrost ten jest wykładniczy. Na przykład zwiększeniu masy rdzenia z 10 do 100 t odpowiada wzrost pola z 3,5 T do 7 T. Największy zbudowany elektromagnes rdzeniowy miał 120 t i wytwarzał pole 7 T. Dalszy rozwój tego typu źródeł silnych pól magnetycznych

Ważniejsze osiągnięcia w dziedzinie silnych pól magnetycznych

Rok	Autor	Kraj	Osiągnięcie
1820	H. Ch. Oersted	Dania	odkrycie zjawiska oddziaływania prądu na pole magnetyczne
1914	G. Deslandres, A. Perrot	Francja	zbudowanie pierwszego elektromagnesu bezrdzeniowego, chłodzonego wodą, wytwarzającego pole 5 T
1923	P. L. Kapica	W. Brytania	wytworzenie po raz pierwszy pola impulsowego 20 T
1936	F. Bitter	USA	zastosowanie w badaniach fizycznych elektromagnesów bezrdzeniowych, chłodzonych wodą, wytwarzających pola do 10 T
1956	A. D. Sacharow	ZSRR	otrzymanie pola impulsowego 2500 T metodą ściskania strumienia magnetycznego za pomocą eksplozji
1961	S. H. Autler, R. R. Hake, K. M. Olsen i in.	USA	wykonanie pierwszych elektromagnesów nadprzewodnikowych, wytwarzających pola powyżej 5 T
1977	M. J. Leupold, R. J. Weggel, Y. Iwasa	USA	uruchomienie elektromagnesu hybrydowego wytwarzającego pole stacjonarne 30 T (średnica wewnętrzna 4,0 cm)



Rys. 1. Pola magnetyczne osiągane różnymi metodami i spotykane w przyrodzie. Poszczególne punkty przedstawiają pola uzyskiwane doświadczalnie. Linia kreskowana oddziela obszar pól dotychczas dostępnych (poniżej krzywej) od obszaru pól nieosiągalnych. Na osi poziomej podano czas narastania impulsu lub czas możliwej pracy elektromagnesu

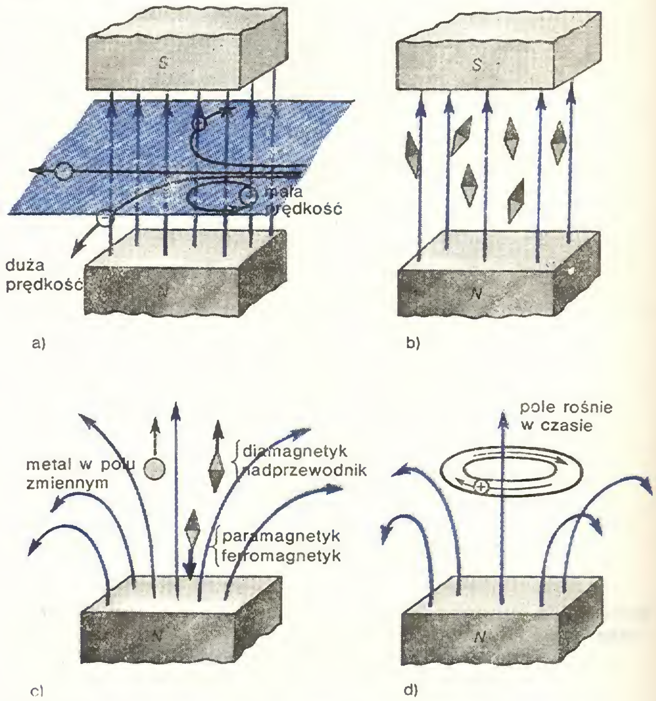
nie miałby już sensu i dlatego do uzyskiwania pól powyżej 5 T stosuje się elektromagnesy bezrdzeniowe.

Do czego mogą służyć silne pola magnetyczne

Zastosowanie silnych pól magnetycznych związane jest z podstawowymi przejawami działania pola magnetycznego przedstawionymi schematycznie na rys. 2.

dziedziny zastosowań

Silne pola magnetyczne stosuje się do formowania wiązek cząstek (naładowanych), do zmiany toru pojedynczych cząstek naładowanych o dużej energii (a i d), do utrzymywania plazmy w ograniczonym obszarze i rozdzielania ładunków w generatorach magnetohydrodynamicznych (a), do badania ogólnych własności ciała stałego i konkretnych własności materiałów, a także do obróbki metali (b i c). Zilustrujemy rolę silnych pól najpierw w badaniach ciała stałego.



Rys. 2. Działanie pola magnetycznego: a) Zakrzywianie torów cząstek naładowanych w płaszczyźnie prostopadłej do pola (na ruch w kierunku równoległym pole nie działa). Promień krzywizny $r = mv/qB$, gdzie v prędkość, q ładunek elektryczny, m masa, B indukcja magnetyczna. b) Porządkowanie kierunku dipoli magnetycznych wzdłuż linii pola. c) Działanie sił wciągających i wypychających w polu niejednorodnym. d) Przyspieszanie ładunków w polu zmiennym rosnącym w czasie

w nadprzewodnikach

Nadprzewodniki stosuje się m.in. do budowy elektromagnesów wytwarzających silne pola magnetyczne. Z drugiej strony silne pola są niezbędne do określenia własności nadprzewodników, przy czym muszą one być większe niż pola krytyczne nadprzewodników (sięgające 60 T). Fizyków interesują badania takich własności nadprzewodników, jak zależność pola krytycznego od temperatury, czy prądu krytycznego od pola i temperatury (rys. 3) bądź przebiegi namagnesowania, z których można wyciągnąć wnioski o mechanizmie powstawania nadprzewodnictwa, ruchu linii strumienia magnetycznego przenikających przez nadprzewodnik itp.

w metalach

Źródłem informacji o strukturze elektronowej metali są badania zmian oporu elektrycznego w polu magnetycznym. Opór metali rośnie w polu magnetycznym i to tym silniej, im niższy jest opór właściwy metalu. Dlatego w badaniach tych stosuje się niskie temperatury i czyste próbki. Niektóre metale wykazują silny (liniowy lub kwadratowy) wzrost oporu ze wzrostem pola, a inne — po początkowym wzroście — wykazują nasycenie. W monokryształach opór w jednych kierunkach może zmieniać się wg jednej zależności, a w innych — wg drugiej (anizotropia magnetooporu, rys. 4). Zjawiska te mają ścisły związek z własnościami gazu elektronowego (z kształtem powierzchni Fermiego), a bardziej poglądowo — z możliwościami ruchu elektronów w sieci krystalicznej w polu magnetycznym. Wiemy, że pod działaniem

pola magnetycznego elektrony powinny poruszać się po okręgach. W sieci krystalicznej natrafiają jednak na oddziaływania pól elektrycznych, które pozwalają im na ruch po okręgach lub — ogólniej — krzywych zamkniętych, albo po krzywych otwartych (np. po linii falistej). W pierwszym wypadku opór elektryczny w silnym polu nie zmienia się, a w drugim rośnie kwadratowo z polem. Czasem w dostatecznie silnym polu następuje przejście z jednego typu zależności w drugą, tzw. przebicie magnetyczne. Podobne zjawiska zachodzą w półprzewodnikach. Zmiany własności nośników prądu w metalach i półprzewodnikach rzutują również na własności cieplne i optyczne tych substancji.

przebicie magnetyczne

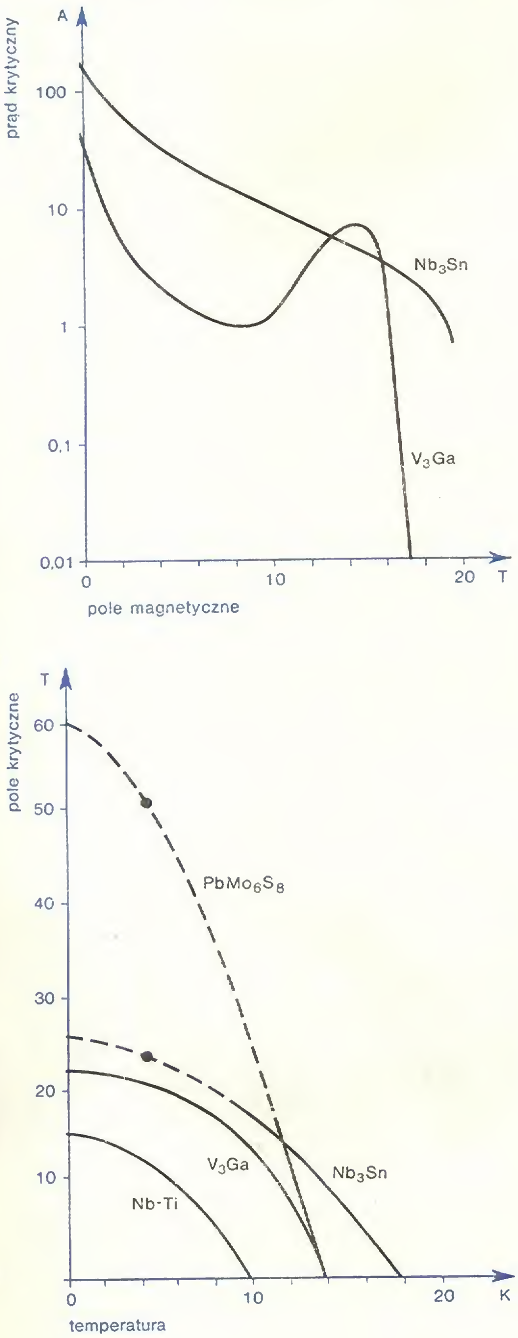
w paramagnetykach

W ciałach paramagnetycznych występuje zjawisko ustawiania się dipoli magnetycznych w kierunku pola. Obserwuje się to jako wzrost namagnesowania ciała (tj. paramagnetyk w polu staje się magnesem). Stopień uporządkowania zależy od temperatury, gdyż ruchy cieplne niszczą uporządkowanie. Aby w danej temperaturze uporządkować większość dipoli, trzeba stosować pole ok. 2 T na każdy stopień temperatury w paramagnetykach elektronowych i ok. 3000 T w paramagnetykach jądrowych. W celu namagnesowania tych ostatnich trzeba stosować temperatury poniżej 10^{-2} K.

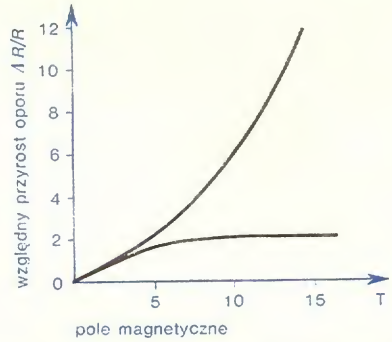
nadprzewodniki „twarde"

a)

b)



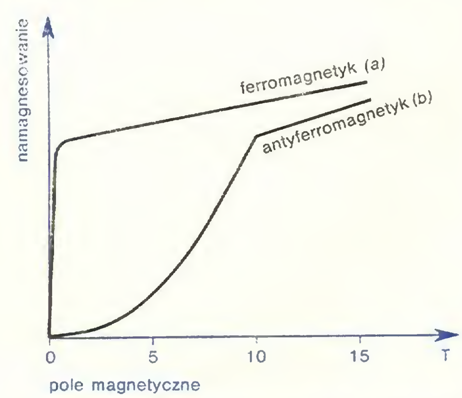
Rys. 3. Charakterystyki nadprzewodników „twardych": a) Zależność natężenia prądu krytycznego od natężenia pola magnetycznego w temperaturze 4,2 K; średnica drutów 0,5 mm. b) Zależność natężenia pola krytycznego od temperatury. Linie ciągłe przedstawiają pomiary w polach stałych, kropki — dane z pomiarów w polach impulsowych, linie przerywane — ekstrapolacje hipotetyczne. Nadprzewodnik $PbMo_6S_8$, o najwyższym polu krytycznym, nie jest jeszcze stosowany w praktyce



Rys. 4. Typowa zależność oporu elektrycznego od pola magnetycznego (monokryształ rutenu; pole magnetyczne skierowane prostopadle do dwu różnych osi krystalograficznych; temperatura 4,2 K)

w ferromagnetykach

W ferromagnetykach, dipole magnetyczne ustawione są równolegle względem siebie i zwrócone w pewnym wyróżnionym kierunku narzuconym przez strukturę krystaliczną, przy czym mniej więcej w połowie objętości materiału zwrot dipoli jest przeciwny niż w pozostałej jego części. Wypadkowy moment magnetyczny jest równy zeru. Gdy przyłożymy pole magnetyczne w tym właśnie kierunku krystalograficznym, już w stosunkowo słabym polu następuje zmiana zwrotu dipoli skierowanych przeciwnie (o 180°) i próbka zostaje namagnesowana do nasycenia. Dlatego ten kierunek krystalograficzny nazwano kierunkiem łatwego magnesowania. Aby próbkę namagnesować w kierunku odmiennym od łatwego, trzeba dipole obrócić o określony kąt względem osi łatwego magnesowania, a to wymaga włożenia odpowiedniej pracy w pokonanie energii pola krystalicznego. Całkowite namagnesowanie próbki uzyskuje się w polach magnetycznych dużo większych, dochodzących np. do 100 T. Wielkość pola magne-



Rys. 5. Krzywe namagnesowania ferromagnetyka (polikryształ USbSe) i antyferromagnetyka (monokryształ $PrSn_3$) w temperaturze 4,2 K. Do całkowitego namagnesowania trzeba by zastosować pola rzędu 50 T (*a*) i 30 T (*b*)

tycznego niezbędnego do całkowitego namagnesowania próbki jest więc miarą wielkości działających pól anizotropii magnetokrystalicznej i charakteryzuje dany materiał (rys. 5, krzywa *a*).

w antyferromagnetykach

Jeszcze ciekawiej wygląda przebieg namagnesowania antyferromagnetyków (krzywa *b*). Pola molekularne powodują w nich ustawienie sąsiednich dipoli w kierunkach przeciwnych, a anizotropia magnetokrystaliczna — wzdłuż określonej osi w krysztale. Pole magnetyczne stara się ustawić dipole w swoim kierunku i proces namagnesowania zależy od wielkości pola w stosunku do wielkości tych innych oddziaływań i od jego kierunku. Na ogół obserwuje się, że w określonym polu namagnesowanie zmienia się skokowo. Oznacza to pokonanie bądź pola anizotropii, bądź pola molekularnego (wówczas antyferromagnetyk zmienia się w ferromagnetyk). Badania w silnych polach pozwalają na określenie wielkości tych oddziaływań.

W miarę zwiększania natężenia pola elektrony w metalu krążą (o ile mogą) po okręgach o coraz mniejszych promieniach. W polach rzędu 10 tys. T (a więc jeszcze nieosiągalnych) rozmiary tych okręgów powinny być zbliżone do rozmiarów orbit atomowych. Również same orbity atomowe ulegną deformacji (ściskaniu), co powinno prowadzić do radykalnych zmian własności ciał i atomów.

w fizyce jądrowej

Silne pola magnetyczne stosowane są w urządzeniach fizyki jądrowej do przyspieszania cząstek, wyprowadzania ich z obszaru akceleratorów, separacji i identyfikacji. Do lat sześćdziesiątych stosowano wszędzie w tych wypadkach przestrzennie olbrzymie elektromagnesy rdzeniowe, wytwarzające pola tylko do ok. 2 T, zasilane mocami rzędu dziesiątków megawatów. Właśnie z tego powodu różnego typu akceleratory (np. synchrotrony) rozrastały się do olbrzymich, kilometrowych rozmiarów. Zwiększenie pola w tych urządzeniach np. pięciokrotnie pozwoliłoby na porównywalne zmniejszenie wymiarów lub zwiększenie uzyskiwanej energii. W wielu krajach przygotowuje się projekty zastosowania do tego celu elektromagnesów nadprzewodnikowych o indukcji ok. 5 T. Pola impulsowe do 30 T w elektromagnesach o stosunkowo dużych średnicach stosuje się do identyfikacji cząstek w komorach pęcherzykowych.

do syntezy termojądrowej

Podobna sytuacja istnieje w urządzeniach do syntezy termojądrowej. Do utrzymywania plazmy w komorze badawczej stosuje się „butelki" lub „zwierciadła" magnetyczne, wytwarzające pola do 20 T. W urządzeniach pierścieniowych typu Tokamak energia cząstek plazmy zależy od wielkości pola impulsowego działającego na cząstki, a z kolei do ich utrzymywania w rurze trzeba odpowiednio dużego pola skierowanego wzdłuż pierścienia. W urządzeniu Alcator zbudowanym ostatnio w Stanach Zjednoczonych zastosowano oba pola po 10 T.

W magnetohydrodynamicznych generatorach energii elektrycznej pole magnetyczne służy do rozdzielania ładunków w plazmie. Zwiększenie pola przy stosowaniu elektromagnesów nadprzewodnikowych prowadzi tu — tak jak i w akceleratorach — do zmniejszenia rozmiarów lub zwiększenia uzyskiwanej energii. Obecnie próbuje się stosować elektromagnesy nadprzewodnikowe wytwarzające pola 5 T zamiast elektromagnesów rdzeniowych. W Japonii zbudowano np. elektromagnes nadprzewodnikowy wytwarzający w komorze o poprzecznych wymiarach 1,3 × 0,4 m i długości 2,8 m pole do 4,5 T. Masa jego przekracza 50 ton.

do kształtowania metali

Pole magnetyczne może służyć nie tylko do przyspieszania cząstek elementarnych. Jeżeli w polu silnego elektromagnesu impulsowego umieścimy metalową kulkę, to prądy indukowane w metalu wytworzą własne pola skierowane przeciwnie do pola głównego i kulka będzie wypychana z pola na zewnątrz (rys. 2c). W doświadczeniach z ciałami o masie ok. 2 g uzyskano prędkości powyżej 100 km/s. Podobny efekt — wypychanie metalu w polu impulsowym — stosuje się w technice do walcowania prętów i prasowania nawet skomplikowanych kształtek. W ten sposób wytwarza się również wysokie ciśnienia stosowane do badań fizycznych, a także wysokie ciśnienia w przemyśle do kształtowania metali.

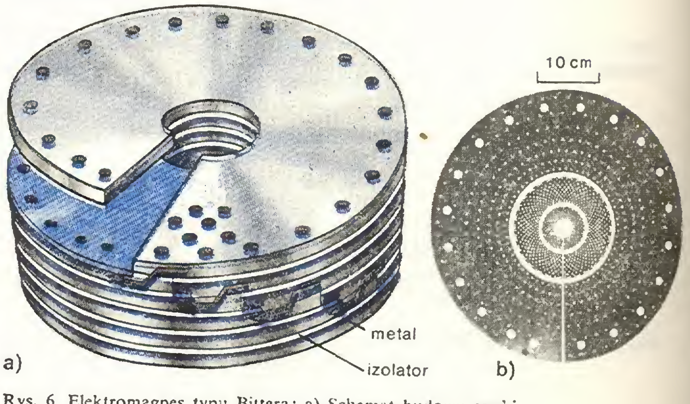
Wytwarzanie silnych pól magnetycznych

elektromagnesy bezrdzeniowe

Pole magnetyczne wytworzone wokół prostego przewodu z przepływającym prądem jest niedogodne do badań, gdyż zmienia się silnie w przestrzeni. Do doświadczeń fizycznych potrzebne są pola jednorodne w obszarach o rozciągłości od jednego do kilku centymetrów i większych. Uzyskuje się je najprościej w solenoidzie. Jednostką natężenia pola jest A/cm właśnie dlatego, że dla wytworzenia pola w solenoidzie trzeba przez uzwojenie przepuścić tyle amperów na 1 cm długości solenoidu, ile ma wynieść natężenie pola; np. dla wytworzenia pola o indukcji około 10 T, tzn. o natężeniu 10^5 A/cm, przez każdy cm długości solenoidu trzeba przepuścić prąd o natężeniu 100 tys. amperów (może to być 10^5 A w jednym przewodzie lub po 1000 A w stu przewodach). Pole można zwiększyć nakładając solenoidy na siebie, gdyż pola poszczególnych solenoidów dodają się. W ten sposób otrzymujemy cewkę cylindryczną z wewnętrznym otworem pomiarowym. Natężenie pola w danej cewce jest zawsze proporcjonalne do natężenia prądu i liczby zwojów, a ponadto zależy od kształtu i wielkości cewki oraz rozkładu prądu wewnątrz (w różnych częściach cewki można przepuszczać prądy o różnych natężeniach). Zwoje dalej położone od środka cewki wytwarzają pole słabsze, ich działanie jest więc mniej skuteczne.

kształty elektromagnesów

Elektromagnesy do wytwarzania silnych pól magnetycznych nie zawsze mają kształty cylindryczne (jak np. solenoid). W wielu doświadczeniach z fizyki jądrowej, w urządzeniach termojądrowych czy w generatorach magnetohydrodynamicznych stosuje się elektromagnesy o nieraz bardzo skomplikowanych kształtach, dostosowanych do żądanego przebiegu linii sił pola magnetycznego (rys. 6 i il. 61, tabl. 16). Pola magnetyczne uzyskiwane różnymi metodami mają zawsze ograniczoną wartość, która zależy od sposobu uzyskiwania pola.



Rys. 6. Elektromagnes typu Bittera: a) Schemat budowy cewki (strzałki oznaczają drogę przepływu prądu). Narysowano tylko część otworków tworzących w cewce kanaliki dla przepływu wody chłodzącej. b) Wygląd kompletu blach uzwojenia trzycewkowego elektromagnesu wytwarzającego pole 20 T (Międzynarodowe Laboratorium Silnych Pól Magnetycznych i Niskich Temperatur, Wrocław). Widoczne otworki — to kanaliki chłodzące

Energia elektryczna doprowadzana do elektromagnesu jest potrzebna do wytwarzania pola w czasie jego narastania oraz do pokonania oporu elektrycznego stawianego przez przewodnik przy przepływie prądu. Zużycie energii na wytworzenie pola odgrywa istotną rolę w elektromagnesach impulsowych. Wy-

dzielanie ciepła wskutek przepływu prądu zachodzi natomiast zarówno w elektromagnesach impulsowych, jak i elektromagnesach wytwarzających pola stałe. Dlatego najekonomiczniejsze są elektromagnesy nadprzewodnikowe.

elektromagnesy nadprzewodnikowe

Ze względu na brak oporu elektrycznego prąd może przepływać przez nadprzewodniki bez strat energii. Wydawałoby się więc, że wystarczy nawinąć cewkę z nadprzewodnika, zanurzyć w ciekłym helu, przepuścić określony prąd i otrzymamy odpowiednie pole. Okazuje się jednak, że istnieją tu dwa ograniczenia. Zasadnicze polega na tym, że każdy nadprzewodnik traci własności nadprzewodzące w polu powyżej pola krytycznego (rys. 3). Te szczytowe wartości praktycznie też nie są osiągalne, gdyż w danym polu przez nadprzewodnik może płynąć tylko ograniczony prąd. Można to skompensować znacznym wzrostem objętości cewki, co jest nieekonomiczne. Dlatego najwyższe pola magnetyczne uzyskiwane w elektromagnesach nadprzewodnikowych wynoszą 17,5 T, a projektuje się magnesy na 20 T (mimo że maksymalne pola krytyczne stosowanych materiałów wynoszą ok. 25 T). Tak więc dla uzyskania pól wyższych trzeba wydatkować znaczną energię.

straty energii

Aby zdać sobie sprawę z wielkości strat energii, zrobimy przykładowe oszacowanie. Wyobraźmy sobie, że z pręta miedzianego o grubości 1 cm wykonamy solenoid o średnicy 4 cm i długości 20 cm. Dla uzyskania 10 T musimy przepuścić prąd o natężeniu 10^5 A. Znając długość pręta (ok. 3 m) i opór właściwy miedzi możemy obliczyć opór pręta i wydzieloną moc. Wynosi ona aż 6 MW.

Rzeczywiste elektromagnesy wytwarzające takie pola pobierają moc mniejszą, ok. 2 MW. Ponieważ moc rośnie proporcjonalnie do kwadratu natężenia prądu — rośnie ona z kwadratem pola. Taki sam solenoid na 20 T zużyje już teoretycznie 8 MW. Dla porównania: trzydziestotysięczne miasto zużywa dla potrzeb gospodarstwa domowego wieczorem ok. 4 MW. Ta sama moc może ogrzać w ciągu sekundy 10 kg miedzi do temperatury topnienia.

Ostatni przykład wskazuje, że gdyby nie przedsięwziąć żadnych środków ostrożności, uzwojenie elektromagnesu stopiłoby się w ciągu kilkudziesięciu sekund. Dlatego elektromagnesy takie muszą być chłodzone, najczęściej wodą o wysokiej czystości (dla uniknięcia korozji i elektrolizy).

elektromagnesy dużej mocy

W zakresie mniej więcej do 20 T moc pobierana przez cewki rośnie proporcjonalnie do kwadratu pola, potem jednak wzrost staje się coraz szybszy. Tak np. dla uzyskania pola 40 T zamiast $2^2 \times 8 = 32$ MW trzeba by dostarczyć około 200 MW. Takie elektromagnesy są już projektowane, ale na budowę ich mogą sobie pozwolić tylko najbogatsze kraje. Przyczyną silnego wzrostu mocy są naprężenia mechaniczne w cewkach.

Wiadomo, że pole magnetyczne działa na przewodnik z prądem siłą proporcjonalną do iloczynu natężenia prądu i indukcji pola. Ponieważ zaś samo pole jest proporcjonalne do natężenia prądu, siły działające na uzwojenie rosną z kwadratem pola. Są to siły ściskające uzwojenie wzdłuż osi i rozrywające je promieniście na zewnątrz. W polach ok. 25 T osiąga się granicę wytrzymałości miedzi. Należy więc zastąpić ją materiałem bardziej wytrzymałym, ale o większym oporze właściwym (co daje dodatkową stratę energii), albo zmniejszyć natężenie prądu w ośrodku magnesu, a to oznacza powiększanie rozmiarów magnesu. Ponadto ilość kanałów chłodzących nie może być dowolnie duża, np. nie mogą się one z sobą stykać, gdyż uniemożliwiłyby przepływ prądu. Wszystko to prowadzi do wzrostu wielkości cewki, przy niezmiennej średnicy wewnętrznej, i do lawinowego wzrostu traconej mocy.

Największe pola o mocy do 23 T w magnesach chłodzonych wodą otrzymywane są w Laboratorium Silnych Pól CNRS w Grenoble i w Narodowym Laboratorium Magnetycznym MIT w Cabmridge (USA), poza tym istnieje ok. 15 ośrodków dysponujących polami słabszymi.

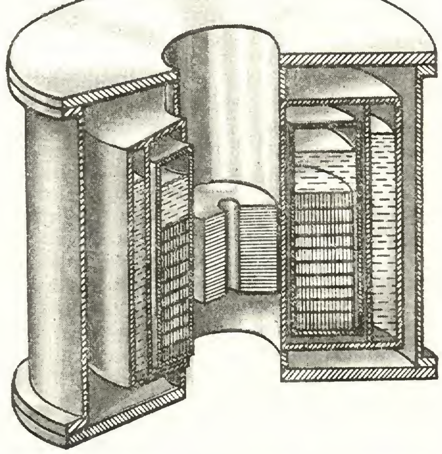
elektromagnesy kriogeniczne

W poszukiwaniu sposobów zmniejszenia mocy niezbędnych do zasilania elektromagnesów zwrócono uwagę na możliwość zmniejszenia oporu właściwego metalu przez obniżenie temperatury. Opór właściwy miedzi w ciekłym azocie maleje ok. 6 razy, w ciekłym wodorze lub ciekłym helu — kilkaset razy, a opór bardzo czystego aluminium — kilka tysięcy razy. Z megawatów przechodzimy więc już do kilowatów. Jakkolwiek pobór mocy bezpośrednio przez elektromagnes odpowiednio maleje, to, jeśli się wliczy koszty skraplania, okazuje się, że opłacalna jest tylko praca elektromagnesu wykonanego z aluminium chłodzonego ciekłym helem.

W latach sześćdziesiątych wykonano, głównie w Stanach Zjednoczonych, kilka elektromagnesów do pracy ciągłej. Nie znalazły one jednak szerszego zastosowania ze względu na znaczne ilości zużywanego chłodziwa. Dla przykładu: wykonany w Boulder elektromagnes na 10 T, o mocy zasilania 5 kW (zamiast normalnie ok. 3,5 MW), zużywał ok. 600 l ciekłego wodoru na godzinę. Toteż obecnie elektromagnesy kriogeniczne stosuje się na ogół do pracy impulsowej.

elektromagnesy hybrydowe

Stosunkowo niedawno (w 1966 r.) wysunięto ideę połączenia cech elektromagnesu dużej mocy (wytwarzanie silnych pól) i nadprzewodnikowego (moc zasilania bardzo mała). Nakładając na elektromagnes dużej mocy (np. 20 T) elektromagnes nadprzewodnikowy na 10 T otrzymamy układ wytwarzający pole 30 T przy nie zmienionej mocy zasilania. W 1974 r. ukończono budowę takich elektromagnesów w dwu ośrodkach (IEA w Moskwie i MIT w USA; rys. 7 i il. 60, tabl. 16). Obecnie na świecie pracują już 4 magnesy hybrydowe, a najsilniejsze pole wytworzone w nich wynosi 30 T.



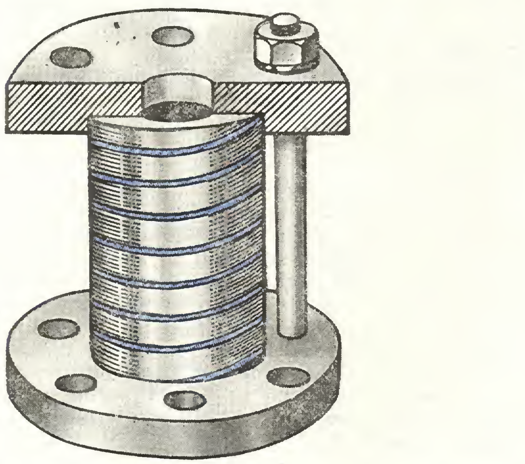
Rys. 7. Schemat elektromagnesu hybrydowego wytwarzającego pole 25 T, wykonanego w Instytucie Energii Atomowej w Moskwie. Część wewnętrzna chłodzona wodą wytwarza pole 18,4 T przy mocy zasilania 5,6 MW, część zewnętrzna, nadprzewodząca, wytwarza pole 6,3 T. Elektromagnes nadprzewodnikowy (zanurzony w zbiorniku z ciekłym helem) ma masę ok. 1,5 t i średnicę zewnętrzną 70 cm

Opisane dotychczas elektromagnesy mogą utrzymywać zadaną wartość pola przez czas dostatecznie długi, np. kilkadziesiąt minut. Wykorzystuje się je do badań wymagających dłuższego czasu lub takich, w których stałość pola odgrywa istotną rolę. Prowadzi się więc w nich badania fizyczne ciał stałych, zwłaszcza metali, opisane w poprzednim rozdziale. Stosuje się je również w niektórych urządzeniach fizyki jądrowej.

Pola silniejsze, powyżej 30 T, uzyskuje się metodą bardziej oszczędną. Ponieważ pobieranie przez dłuższy czas prądu o mocy dziesiątków lub setek mega-

elektromagnesy impulsowe

watów jest zbyt kosztowne, doprowadzamy taki prąd do elektromagnesów odpowiednio krótkimi impulsami, gromadząc energię elektryczną w urządzeniach zasilających przez czas stosunkowo długi, w przerwach między impulsami. Wówczas średnia moc pobierana może być dostatecznie niska. Przykładem może być największy elektromagnes chłodzony wodą w Canberrze (Australia), zasilany mocą szczytową 30 MW w ciągu paru sekund, a wytwarzający pole 30 T. Ogólnie ma tu zastosowanie zasada: silniejsze pole — krótszy czas. Dalszym sposobem zmniejszania mocy pobieranej jest zmniejszenie średnicy wewnętrznej cewki, np. do 1 cm. Cewki są wówczas nawinięte z drutu, bądź stanowią jednowarstwowe solenoidy (rys. 8), czy wreszcie składają się tylko z jednego zwoju. Im mniej zwojów, tym krótszy czas impulsu.



Rys. 8. Schemat jednowarstwowego elektromagnesu impulsowego. Pola impulsowe do ok. 50 T wytwarzane są tą metodą w Polsce w różnych instytutach Warszawy, Poznania i Wrocławia

Do zasilania typowych cewek stosuje się na ogół baterie kondensatorów. Tak np. w jednym z największych laboratoriów tego typu w Los Alamos bateria może zmagazynować energię 10 mln J. Minimalny czas rozładowania wynosi 4 μs. Dzieląc energię przez czas otrzymujemy moc rzędu 10^{12} W (milion megawatów); otrzymywane pola sięgają 400 T.

wpływ samoindukcji kondensatorów na moc elektromagnesów

Wydawałoby się, że można zwiększać liczbę kondensatorów i skracać czas rozładowania. Jednak jest to niemożliwe: czas rozładowania ogranicza bowiem samoindukcja kondensatorów, a zwiększenie ich liczby powiększa samoindukcję obwodu (długość połączeń).

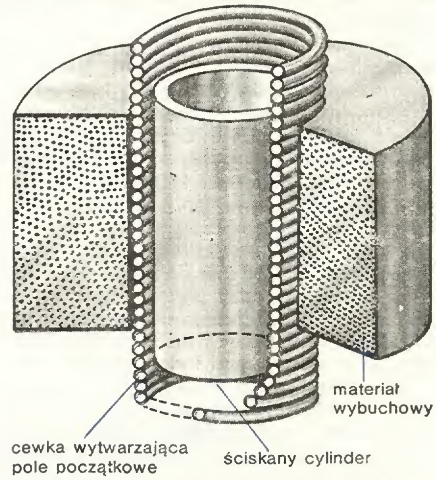
Cewki elektromagnesów impulsowych na ogół są chłodzone cieczami kriogenicznymi w celu obniżenia oporu elektrycznego. W trakcie impulsu prądu ciepło wydzielane w cewce nie może szybko wydostać się na zewnątrz, a zatem cewka się ogrzewa. W wypadku impulsów najkrótszych prąd płynie tylko po powierzchni, ponadto pole magnetyczne dyfundując w głąb solenoidu nadtapia warstwę naskórkową metalu. Z tego względu, a także z uwagi na naprężenia mechaniczne przekraczające wytrzymałość materiału, cewki wytwarzające pola powyżej 100 T ulegają każdorazowo rozerwaniu lub deformacji.

ściskanie strumienia magnetycznego

Do otrzymywania pól jeszcze wyższych stosuje się inne metody. Wyobraźmy sobie, że wytworzymy pole o indukcji 10 T w cienkościennym cylindrze metalowym o osi skierowanej wzdłuż pola, a następnie będziemy go szybko ściskać promieniście ku osi. Wskutek tego zmniejszać się będzie przekrój cylindra, zachowany natomiast zostanie strumień indukcji magnetycznej. Ponieważ strumień magnetyczny jest to iloczyn indukcji magnetycznej i powierzchni obejmowanej przez przewodnik (w tym wypadku — pole przekroju cylindra), zachowanie stałej wartości strumienia przy n-krotnym zmniejszeniu powierzchni przekroju powoduje n-krotny wzrost indukcji magnetycznej. Przy typowych wartościach średnic cylindra: początkowej 10 cm i końcowej 0,5 cm, pole przekroju maleje 400 razy; odpowiednio do tego pole magnetyczne powinno wzrosnąć z 10 T do 4000 T. W rzeczywistości osiąga się wzrost dwustukrotny, z powodu częściowego wypływu strumienia magnetycznego przez ścianki cylindra. Do ściskania cylindra stosuje się eksplozje ładunków wybuchowych (rys. 9).

zastosowanie ładunków wybuchowych

Ograniczeniem tej metody jest wartość ciśnienia wytwarzanego przez istniejące materiały wybuchowe, wskutek czego pierścieniowi można nadać tylko ograniczone przyspieszenie. Jest ono ograniczone jego masą oraz siłami elektrodynamicznymi wynikającymi z ruchu przewodnika w polu magnetycznym. W miarę wzrostu pola ruch ścianek cylindra jest hamowany rosnącym przeciwdziałaniem pola magnetycznego (równoważne ciśnieniu 10^{12} Pa w polu 1500 T). A więc dalszy rozwój wymaga albo stosowania silniejszych materiałów wybuchowych, albo nowych metod. Metodą ściskania strumienia magnetycznego uzyskuje się pola do 1500 T, przy czym rekordowe pole wynosiło 2500 T (rys. 1).



Rys. 9. Schemat jednego z możliwych urządzeń do uzyskiwania najsilniejszych pól magnetycznych metodą ściskania strumienia magnetycznego. Zużywa się 10–20 kg materiału wybuchowego

zastosowanie pól impulsowych

Pola impulsowe stosuje się we wszystkich zagadnieniach opisanych w rozdziale o zastosowaniach silnych pól magnetycznych, a więc zarówno do badań ciał stałych, jak też w fizyce jądrowej czy termojądrowej. Jednakże długość impulsu pola musi być dostosowana do danego zagadnienia fizycznego, a to z kolei ogranicza wielkość stosowanych pól magnetycznych. Można podać orientacyjnie, że w urządzeniach fizyki jądrowej i termojądrowej stosuje się pola do 30 T, w badaniach ciał stałych — do kilkuset T. Pola najwyższe, powyżej 1000 T stosuje się w badaniach plazmy oraz do wytwarzania dużych ciśnień.

P. Byszewski, A. Szymborski *Ekstremalne pola magnetyczne*, Problemy nr 1, 334 (1974); D. S. Parasnis *Magnetyzm*, Warszawa 1970; J. Szpilecki *Megagausowe pola magnetyczne*, Post. Fiz. 19, 411 (1968); W. Zawadzki *O silnych polach magnetycznych*, Problemy nr 4, 217 (1964).

ELEKTRONIKA WSPÓŁCZESNA

Co to jest współczesna elektronika · Fizyka przyrządów półprzewodnikowych · Przyrządy półprzewodnikowe dyskretne · Optoelektronika półprzewodnikowa · Mikroelektronika · Generacja mikrofal · Komputer jako narzędzie fizyków

Co to jest współczesna elektronika

Jarosław Świderski

Wydarzenia w kosmosie i w świecie wielkich energii fizyki jądrowej usunęły w cień właściwą rewolucję techniczną przebiegającą na naszych oczach, rewolucję, bez której te wielkie osiągnięcia naszej doby nie byłyby w ogóle możliwe. Mowa — oczywiście — o elektronice. Hasło elektronika kojarzy się najczęściej z telewizją i radiotechniką, z komputerami (elektronowymi mózgami) i automatyką przemysłową, radiolokacją i elektromedycyną. Szybki postęp w rozwoju tych urządzeń uświadamiamy sobie, gdy nasze dzieci grymaszą, że odbiornik radiowy nie dość łatwo mieści im się w kieszonce lub że kolory na pokazie telewizyjnym nie są dość naturalne. Od czasu do czasu prasa stara się nas olśnić możliwościami „myślących" maszyn, a gdy przypadkiem trafimy pod opiekę medycyny, przerażają nas coraz bardziej wymyślne urządzenia badające człowieka swymi elektrycznymi mackami.

elektronika techniczna

U źródeł tego rozwoju stoi skromnie właściwa elektronika, zwana inaczej elektroniką techniczną, zajmująca się wytwarzaniem bardzo nieefektownych przyrządów, jak lampy elektronowe (w tym lampy radiowe) i przyrządy półprzewodnikowe (diody, tranzystory, hallotrony i in.). Pierwsze przyrządy elektronowe pojawiły się w początkach naszego stulecia i z punktu narzuciły burzliwy rozwój radiotechnice. Ale zajmujący się nimi naukowcy niemal od razu spostrzegli, że rola ich będzie znacznie większa. Istotą działania tych przyrządów było i jest wykorzystanie ruchu swobodnych elektronów do wykrywania, selekcjonowania, wzmacniania i przetwarzania docierających do nich bodźców-sygnałów. A to przecież jest istotą działania poszczególnych komórek i całych systemów nerwowych, jest istotą procesu myślenia.

Jak wiadomo, wspaniałe działanie naszego mózgu zawdzięczamy ogromnej liczbie elementarnych komórek mikroskopijnej wielkości. Pierwsze lampy elektronowe miały kilkanaście cm wysokości, toteż zestawienie bodaj kilkuset takich „komórek" stawało się dosłownie monstrualnym problemem.

tranzystor — przełom w elektronice

Zasadnicza zmiana nastąpiła z chwilą, gdy wynaleziono tranzystor — wzmacniający przyrząd elektronowy wykorzystujący ruch elektronów nie w próżni, jak to było w lampach, lecz w ciele stałym (a ściślej — w półprzewodniku). Szybkość rozwoju produkcji tranzystorów, a wraz z nimi innych przyrządów elektronowych, jest czymś, nawet w skali pojęć dwudziestowiecznych tak wyjątkowym, że śmiało można mówić o „eksplozji" elektroniki. Jeśli się przyjmie, że w 1950 r. na całym świecie wyprodukowano przyrządów elektronowych za sumę ok. 5 mld umownych jednostek monetarnych, to w 1960 r. wyprodukowano ich już za 25 mld, a w 1970 — za 48 mld. W 1980 r. wartość wyprodukowanych przyrządów elektronowych wyniosła już 183 mld takich umownych jednostek.

Zastosowania przyrządów półprzewodnikowych obejmują, jak powódź, coraz to nowe dziedziny, przy czym ekspansja techniczna łączy się tu z ekspansją ekonomiczną. I tak np. zbudowany w 1966 r. pierwszy kalkulator elektroniczny zawierał w sobie pokaźną liczbę diod i tranzystorów łącznej wartości 170 dolarów. Osiem lat później wewnątrz takiego samego kalkulatora jest jeden tylko przyrząd półprzewodnikowy — układ scalony wielkiej skali integracji, który kosztuje 3,5 dolara; układ ten jest przy tym znacznie bardziej niezawodny i eliminuje prawie całkowicie czynności montażowe w produkcji kalkulatora. W tych warunkach staje się zrozumiałe, że w 1971 r. w USA sprzedawano takie kalkulatory na sztuki, a dwa lata później, w 1973 r., sprzedano ich ok. 6,5 mln sztuk za 700 mln dolarów. Podobnie wygląda rynek zegarków elektronicznych. Takie przykłady można mnożyć i mnożyć.

Entuzjaści elektroniki twierdzą, że postęp, który wprowadza ona do techniki i w ogóle do ludzkiego życia, da się porównać jedynie z postępem, który nastąpił po wprowadzeniu metali w miejsce narzędzi i przedmiotów użytkowych wykonywanych z kamienia. Możliwe, trzeba jednak pamiętać, że tamten proces trwał wiele wieków.

Co da człowiekowi elektronika za następne 50 lat? Jak zmieni oblicze naszej planety w XXI wieku? Ryzykowna byłaby przepowiednia, zwłaszcza po tym, co wyżej powiedziano o czasie minionym. Już w tej chwili jest niemal pewne, że następnym krokiem będą tzw. przyrządy funkcjonalne. Przebiegające w nich procesy będą odwzorowaniem fizycznym lub matematycznym zjawisk rządzących otaczającym nas światem. Dzięki takim przyrządom pojedynczy człowiek zostanie być może wyposażony w dodatkowy mózg, „mózg elektronowy", który zwielokrotni jego możliwości zarówno w dziedzinie ochrony zdrowia, organizacji pracy i odpoczynku, jak i, miejmy nadzieję, w służbie innym ludziom.

przyrządy funkcjonalne

Elektronika półprzewodnikowa w zbliżeniu

Cóż to jest ta elektronika półprzewodnikowa? Według definicji jest to dziedzina nauki i techniki zajmująca się praktycznym wykorzystaniem zjawisk, w których

podstawowe znaczenie ma dający się sterować ruch elektronów w półprzewodnikach. Obejmuje ona teorię, konstrukcję i technologię przyrządów półprzewodnikowych. Jak niemal każda dziedzina techniki, wyrosła na gruncie osiągnięć fizyki, a zwłaszcza fizyki ciała stałego. Natomiast w przeciwieństwie do większości innych dziedzin — elektronika półprzewodników wykorzystała i wykorzystuje osiągnięcia fizyki (odkrycia zjawisk) w bardzo krótkim czasie po ich sformułowaniu, a często nawet korzysta ze zjawisk będących dopiero przedmiotem badań. Stąd ścisłe powiązanie tych dwu gałęzi wiedzy, stąd liczne rzesze fizyków pracujących jako elektronicy i na odwrót.

pierwsze przyrządy — dioda i tranzystor

Elektronika półprzewodników rozpoczęła swój niezależny byt od skonstruowania dwóch elementarnych przyrządów półprzewodnikowych: diody (początek naszego stulecia) i tranzystora (1948 r.). Istnienie tych przyrządów stymulowało badania nad materiałami półprzewodnikowymi i nad technologią struktur półprzewodnikowych. Rozwój tych ostatnich doprowadził do wynalezienia kilkudziesięciu innych przyrządów, bądź wykorzystujących nowe zjawiska w strukturze jedno- lub dwuzłączowej, bądź stanowiących zupełnie nowe formy (kombinacja złącz, kontaktów i kondensatorów), bądź wreszcie będących powieleniem przyrządów klasycznych w monolityczny układ scalony. Rozwój technologii doprowadził w latach sześćdziesiątych do nowego skoku jakościowego, do technologii epiplanarnej i MOS, umożliwiających stosunkowo łatwe i niedrogie (wielkoseryjne) wytwarzanie bardzo skomplikowanych układów. Wówczas

mikroelektronika

to wyodrębniono pojęcie mikroelektroniki — tej dziedziny elektroniki półprzewodnikowej, która się zajmuje układami scalonymi. Oczywiście metody technologiczne mikroelektroniki umożliwiają wytwarzanie również wielu innych przyrządów półprzewodnikowych, przy czym dają możność łączenia ich z gotowymi wzmacniaczami, przetwornikami itp. Można w ten sposób skonstruować fotodetektor z układem zasilająco-detekcyjnym, hallotron ze wzmacniaczem itp. Słuszniej byłoby wobec tego mówić dziś o mikroelektronice jako dziedzinie podstawowej, a o pozostałych dziedzinach elektroniki półprzewodnikowej — jako o częściach czy szczególnych przypadkach mikroelektroniki.

podział elektroniki półprzewodnikowej

Tradycyjnie jednak (jeśli można mówić o tradycji w dyscyplinie liczącej sobie tak naprawdę niespełna 30 lat) dzieli się elektronikę półprzewodnikową inaczej, przyjmując za kryterium rodzaj zastosowania. W ten sposób wyodrębniamy: klasyczną elektronikę półprzewodników, zajmującą się diodami i tranzystorami w postaci dyskretnej (tj. w postaci odrębnych przyrządów); mikroelektronikę, zajmującą się układami scalonymi; optoelektronikę półprzewodnikową, zajmującą się źródłami (diody elektroluminescencyjne, lasery, wskaźniki alfanumeryczne itp.) i detektorami promieniowania oraz ich kombinacjami; elektronikę półprzewodnikowych przyrządów bezzłączowych (hallotrony, gausotrony, termistory itp.); półprzewodnikową elektronikę mikrofalową, zajmującą się przyrządami pracującymi w zakresie mikrofal.

W dalszych rozdziałach omówimy dokładniej ważniejsze z tych dziedzin. Tu warto sobie jedynie uświadomić ogromną różnorodność „elektronicznych dzieci" (il. 8, tabl. 2), by spotykając je dziś w najdziwniejszych nieraz postaciach i we wszystkich gałęziach ludzkiej działalności, wiedzieć, że pochodzą z jednej rodziny.

Fizyka przyrządów półprzewodnikowych

Stanisław Sikorski

Podstawę działania przyrządów półprzewodnikowych stanowią określone zjawiska fizyczne. Jednym z takich zjawisk jest zwiększanie się lub zmniejszanie koncentracji nośników ładunku, a co za tym idzie — możliwość regulacji przewodności właściwej w półprzewodnikach. Najprostszym sposobem zmian koncentracji jest zmiana temperatury materiału, jej podwyższenie powoduje w półprzewodnikach wzrost koncentracji nośników (→ Półprzewodniki). Zjawisko to zostało wykorzystane przy konstrukcji termistorów, tj. przyrządów stosowanych m.in. w układach do pomiaru temperatury.

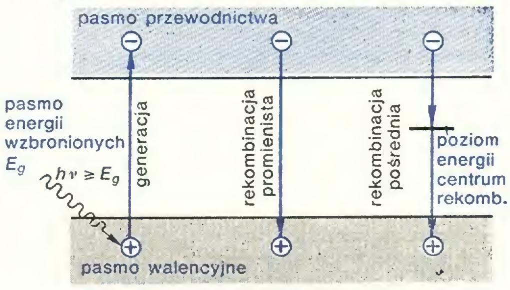
Istnieje jednak możliwość regulowania przewodnictwa w stałej temperaturze, „na zimno" (co jest wykorzystywane w większości przyrządów półprzewodnikowych). Dodatkowe nośniki ładunku, które powodują zmianę przewodnictwa bez podwyższania temperatury, nazywamy nadmiarowymi.

Nośniki nadmiarowe

Znane są dwa zjawiska fizyczne powodujące występowanie nośników nadmiarowych: zjawisko fotoelektryczne wewnętrzne i wstrzykiwanie nośników występujące w złączach *p-n*.

przewodnictwo selenu

W 1873 r. W. Smith zaobserwował silny wpływ światła na przewodnictwo selenu. Nie znano wtedy kwantowego charakteru tego zjawiska. Obecnie dzięki teorii pasmowej (→ Struktura elektronowa ciał stałych) opis jest jasny. Rzucając strumień światła na półprzewodnik, wprowadzamy do jego wnętrza kwanty światła, fotony, o energii $h\nu$ (h — stała Plancka, ν — częstość drgań fali świetlnej). Energetyczny obraz zjawiska przedstawia rys. 1. Wewnątrz półprzewodnika w pasmie przewodnictwa pojawia się dodatkowy



Rys. 1. Przejścia elektronów przez pasmo zabronione: a) zjawisko fotoelektryczne wewnętrzne — generacja pary dziura–elektron, b) rekombinacja bezpośrednia — promienista, c) rekombinacja pośrednia — przez centrum rekombinacji

fotoprzewodnictwo

elektron, a w pasmie podstawowym (walencyjnym) — puste miejsce po wyrwanym elektronie, dziura, która stanowi dodatni nośnik ładunku. Następuje więc generacja pary dziura–elektron. Powstałe pary nośników o różnych znakach tworzą pewien nadmiar koncentracji w stosunku do koncentracji wynikającej z pobudzenia cieplnego, dlatego nośniki te nazywamy nadmiarowymi. Koncentrację nadmiarową dziur oznaczamy Δp, elektronów Δn, przy czym:

$$\Delta p = \Delta n, \qquad (1)$$

gdyż generacja odbywa się parami. Korzystając z ogól-

nego wzoru na elektryczną przewodność właściwą σ półprzewodnika ($\sigma = en\mu$, e — ładunek elektryczny, n — koncentracja nośników, μ — ruchliwość; → Półprzewodniki) i z powyższej zależności, możemy łatwo obliczyć przyrost przewodności właściwej wywołanej oddziaływaniem światła:

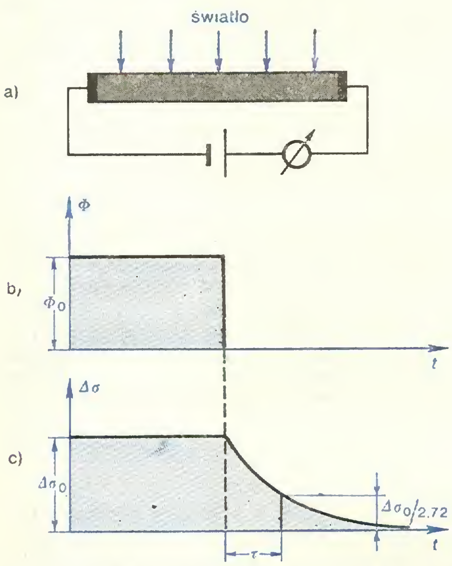
$$\Delta\sigma = e(\mu_n \Delta n + \mu_p \Delta p) = e(\mu_n + \mu_p)\Delta p, \quad (2)$$

gdzie μ_p i μ_n — to odpowiednio ruchliwość dziur i elektronów. Przyrost przewodności właściwej pod wpływem światła nosi nazwę fotoprzewodnictwa. Zjawisko to wykorzystuje się w przyrządach półprzewodnikowych — fotoopornikach (fotorezystorach), używanych do wykrywania strumienia świetlnego i pomiaru jego natężenia.

W zjawisku fotoprzewodnictwa bardzo duże znaczenie ma fakt, że istnieją nośniki ładunku dodatniego i ujemnego (dwunośnikowy mechanizm przewodzenia prądu). Pozwala to na wytwarzanie dużych koncentracji nadmiarowych nośników ładunku i zachowanie jednocześnie wypadkowej neutralności elektrycznej nie tylko w całości układu, ale i lokalnie (warto tu zwrócić uwagę, że zjawisko fotoprzewodnictwa nie występuje w metalach, w których jest tylko jeden rodzaj nośników ładunku elektrycznego).

rekombinacja

Po wyłączeniu światła fotoprzewodnictwo nie zanika natychmiast (rys. 2), lecz stopniowo, a zanik ma charakter wykładniczy, podobnie jak np. rozpad promieniotwórczy. Zjawisko zaniku nośników nadmiarowych nazywa się rekombinacją. Czas τ, po upływie



Rys. 2. Zanik fotoprzewodnictwa po wyłączeniu oświetlenia: a) oświetlona próbka półprzewodnikowa; b) zależność natężenia strumienia świetlnego Φ od czasu t, c) zależność fotoprzewodności $\Delta\sigma$ od czasu t

czas życia

którego $\Delta\sigma$ zmniejszy się e = 2,72 raza, nosi nazwę czasu życia nośników nadmiarowych. Czasy życia nośników nadmiarowych mogą być bardzo różne; $\tau = 10^{-8}$ s uznajemy za krótki, $\tau = 10^{-3}$ s — za bardzo długi.

rekombinacja promienista

Rekombinacja może następować wskutek samorzutnych przejść elektronów z pasma przewodnictwa do pasma walencyjnego (rys. 1b). Wytwarzającą się przy tym energię przejmuje powstający foton o energii $h\nu$, równej przerwie energetycznej E_g. Ten rodzaj rekombinacji nazywamy rekombinacją bezpośrednią lub promienistą. Ma ona bardzo duże znaczenie jako podstawa działania przyrządów optoelektronicznych (→ Optoelektronika półprzewodnikowa).

W krzemie i germanie stwierdzono bardzo słabe promieniowanie rekombinacyjne. W tych półprzewodnikach występuje więc inny sposób rekombinacji,

rekombinacja pośrednia

tzw. rekombinacja pośrednia (rys. 1c). W przerwie energetycznej oprócz poziomów lokalnych, położonych w pobliżu pasma przewodnictwa czy pasma walencyjnego (poziomy donorowe i akceptorowe), powstają poziomy pochodzące od defektów sieci krystalicznej lub atomów domieszkowych niektórych pierwiastków (np. miedzi), stanowiących centra rekombinacji. Poziomy te są zlokalizowane w pobliżu środka przerwy energetycznej. Elektrony nie przechodzą bezpośrednio do pasma podstawowego, lecz najpierw spadają na poziom energetyczny centrum rekombinacji, a stamtąd — prawie natychmiast — do pasma podstawowego, zajmując poziom energetyczny dziury. W rezultacie znika elektron w paśmie przewodnictwa i dziura w paśmie podstawowym. Szereg pierwiastków wprowadzonych do materiału półprzewodnikowego powoduje znaczną rekombinację. Tak np. miedź przy koncentracji $4 \cdot 10^{14}$ atomów w 1 cm³ kryształu germanu sprawia, że τ jest rzędu 10^{-5} s. Ponieważ 1 cm³ germanu zawiera $4 \cdot 10^{22}$ atomów, zatem wystarczy zaledwie 1 atom miedzi na 10^8 atomów germanu, ażeby wywierać decydujący wpływ na rekombinację. Zwiększenie koncentracji takiej domieszki zmniejsza, oczywiście, proporcjonalnie czas życia. Ta sama koncentracja miedzi w krzemie daje dziesięciokrotnie większy skutek, a więc wymagania co do czystości są jeszcze większe.

Jak już powiedzieliśmy, nośniki nadmiarowe odgrywają ogromną rolę w pracy przyrządów półprzewodnikowych, dlatego to wielkość czasu życia τ w danym materiale ma duże znaczenie praktyczne. Ten parametr materiału jest bardzo istotny także i dlatego, że dostarcza cennych informacji o rodzaju domieszek i defektów. Warto zwrócić uwagę, że jest to parametr niezależny od przewodności właściwej σ, ponieważ inne domieszki decydują zazwyczaj o czasie życia, a inne o przewodności.

wpływ powierzchni

Czas zaniku fotoprzewodnictwa w oświetlanych próbkach nie zależy od kształtu próbek i elektrod, natomiast jest w znacznym stopniu zależny od stanu powierzchni (→ Stany powierzchniowe w ciałach stałych). Jeśli zatem chcemy poznać własności rekombinacyjne całej objętości kryształu, musimy zredukować wpływ powierzchni. Kiedy dokonamy bardzo starannej obróbki powierzchni (polerowanie, trawienie chemiczne itp.), czas zaniku fotoprzewodnictwa próbki jako całości będzie równy czasowi życia τ nośników nadmiarowych. Źle obrobiona powierzchnia płytki półprzewodnikowej zawiera ogromne zagęszczenie zakłóceń sieci krystalicznej, działających jako powierzchniowe centra rekombinacji, niezależnie od centrów rekombinacji wewnątrz próbki. Czas zaniku fotoprzewodnictwa jest wówczas dużo mniejszy niż czas życia objętościowy. Jest to tzw. efektywny czas życia.

rekombinacja powierzchniowa

Parametrem określającym rekombinacyjne własności powierzchni jest szybkość rekombinacji powierzchniowej S mierzona w cm/s (jest to szybkość dopływu strumienia nośników nadmiarowych rekombinujących na powierzchni). Jeśli powierzchnia jest dobrze przygotowana (gładka i oczyszczona), $S = 100$ cm/s, jeśli chropowata, S może dochodzić do 10^5 cm/s.

Rekombinacja powierzchniowa zależy nie tylko od samej obróbki, ale także od stanu otaczającej atmosfery (temperatury, wilgotności i innych czynników), a ponadto zmienia się z upływem dni. Ma to ogromne znaczenie dla pracy przyrządów półprzewodnikowych. Wskutek różnych, mniej lub bardziej skutecznych zabiegów technologicznych uzyskuje się powierzchnie o dobrych właściwościach, a przede wszystkim doprowadza się do stabilizacji tych właściwości, do których należy w pierwszym rzędzie szybkość rekombinacji powierzchniowej. Wielkim osiągnięciem nowoczesnej technologii półprzewodników stało się wykorzystanie specjalnie wytwarzanej na płytkach krzemu warstwy tlenku krzemu, która trwale zabezpiecza powierzchnię i świetnie stabilizuje jej własności fizyczne.

Dyfuzja i droga dyfuzji nośników nadmiarowych

Rozważając własności i znaczenie nośników nadmiarowych, nie zajmowaliśmy się ich rozkładem przestrzennym. Ważna była ich ogólna liczba w próbce, ona bowiem decydowała o ogólnym fotoprzewodnictwie. Ponadto wychodziliśmy z założenia (rys. 2a), że próbka jest przeniknięta strumieniem świetlnym wytwarzającym równomierny rozkład nośników nadmiarowych i że oświetlamy całą próbkę. A przecież w wielu wypadkach może być oświetlona tylko część próbki, przy czym jeśli nawet oświetlimy całą próbkę, światło nierównomiernie dociera do wszystkich punktów w jej głębi. Co więcej, bardzo często światło wnika na głębokość dużo mniejszą niż grubość próbki. Wtedy nadmiarowe elektrony i dziury są generowane tuż pod powierzchnią. Jednakże nie wszystkie nośniki nadmiarowe są skupione tam, gdzie dociera światło — mają one możliwość przemieszczania się dzięki dyfuzji.

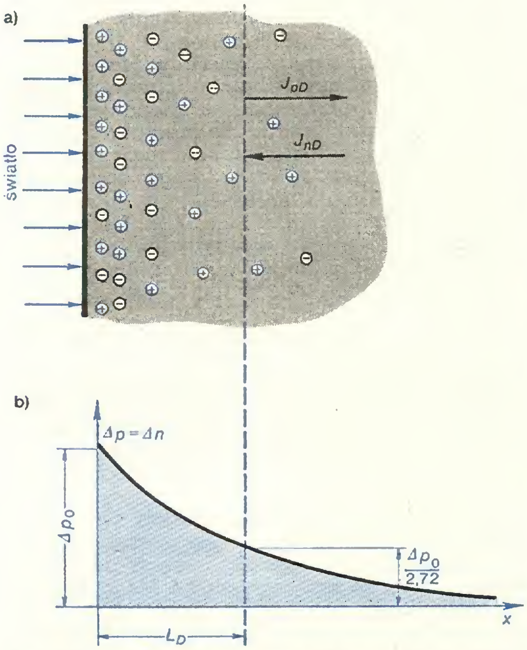
Matematycznym wyrazem dyfuzji jest prawo Ficka, stwierdzające, że gęstość strumienia nośników jest proporcjonalna do przestrzennego spadku koncentracji nośników (pochodna względem współrzędnej ze znakiem minus). Przepływ nośników odbywa się więc w kierunku od większych koncentracji do mniejszych koncentracji.

gęstość prądu dyfuzyjnego dziur

Przypuśćmy, że dziury są rozłożone w ten sposób, że spadek koncentracji występuje tylko w kierunku osi x. Przepływ nośników ładunku dodatniego stanowi jednocześnie pewien prąd elektryczny o tym samym zwrocie, a zatem na mocy prawa Ficka gęstość prądu dyfuzyjnego dziur:

$$J_{pD} = -eD_p \frac{dp}{dx}, \qquad (3)$$

gdzie D_p jest stałą dyfuzji dziur, dp/dx — przyrostem ich koncentracji.



Rys. 3. Dyfuzja nośników nadmiarowych w głąb półprzewodnika

gęstość prądu dyfuzyjnego elektronów

Podobnie dyfuzji elektronów towarzyszy prąd elektryczny o gęstości:

$$J_{nD} = +eD_n \frac{dn}{dx},$$

gdzie D_n — stała dyfuzji elektronów.

Znając ruchliwości μ_p i μ_n nośników, można obliczyć stałe dyfuzji D_p i D_n z zależności Einsteina–Smoluchowskiego:

$$D_p = \frac{kT}{e}\mu_p, \quad D_n = \frac{kT}{e}\mu_n \qquad (4)$$

($kT/e = 0{,}026$ V w temperaturze pokojowej). Stałe dyfuzji nie są równe, gdyż ruchliwości elektronów μ_n i dziur μ_p nie są równe (zależą od masy efektywnej; → Dynamika elektronu w ciałach stałych). Nośniki bardziej ruchliwe mają więc większą stałą dyfuzji.

Wskutek dyfuzji nośniki nadmiarowe generowane tuż pod powierzchnią próbki półprzewodnikowej przemieszczają się w głąb półprzewodnika, dążąc do wypełnienia całej objętości. Jednakże po drodze napotykają centra rekombinacji i ulegają zanikowi. Im więcej ich przybywa do danego miejsca, tym szybsza jest rekombinacja. Po pewnym czasie (kilkakrotna wartość τ) ustala się w każdym punkcie próbki stan równowagi: tyle nośników dopływa, ile ich rekombinuje. Równowagowy rozkład koncentracji nośników nadmiarowych pokazany jest na rys. 3 zarówno w sposób poglądowy, jak i na wykresie (rys. 3b). Przy powierzchni jest najwięcej nośników, w miarę oddalania się w głąb koncentracja ich maleje w sposób wykładniczy. Średnią odległość, którą nośniki przebywają od powierzchni w głąb w czasie swego życia, nazywamy drogą dyfuzji i oznaczamy symbolem L_D. Jest to odległość, na której koncentracja początkowa (Δp_0 — dziur, czy Δn_0 — elektronów) zmniejsza się e ≈ 2,72 raza. Droga dyfuzji jest tym większa, im większa jest stała dyfuzji oraz czas życia τ nośników nadmiarowych. Wielkość ta wyraża się wzorami:

droga dyfuzji

dla półprzewodników o przewodnictwie dziurowym (typ p)

$$L_D = \sqrt{D_n \tau};$$

dla półprzewodników o przewodnictwie elektronowym (typ n)

$$L_D = \sqrt{D_p \tau}.$$

Z tych wzorów wynika, że o drodze dyfuzji nośników nadmiarowych decyduje stała dyfuzja nośników mniejszościowych. Przy wartościach granicznych $\tau = 10^{-8}$–-10^{-3} s droga dyfuzji nośników nadmiarowych zawiera się w granicach 0,01–1 mm.

Opisane wyżej dyfuzyjne wnikanie w głąb materiału nośników nadmiarowych połączone z ich jednoczesną rekombinacją odgrywa dużą rolę w pracy przyrządów półprzewodnikowych, w szczególności w złączu p-n.

Złącze *p-n*

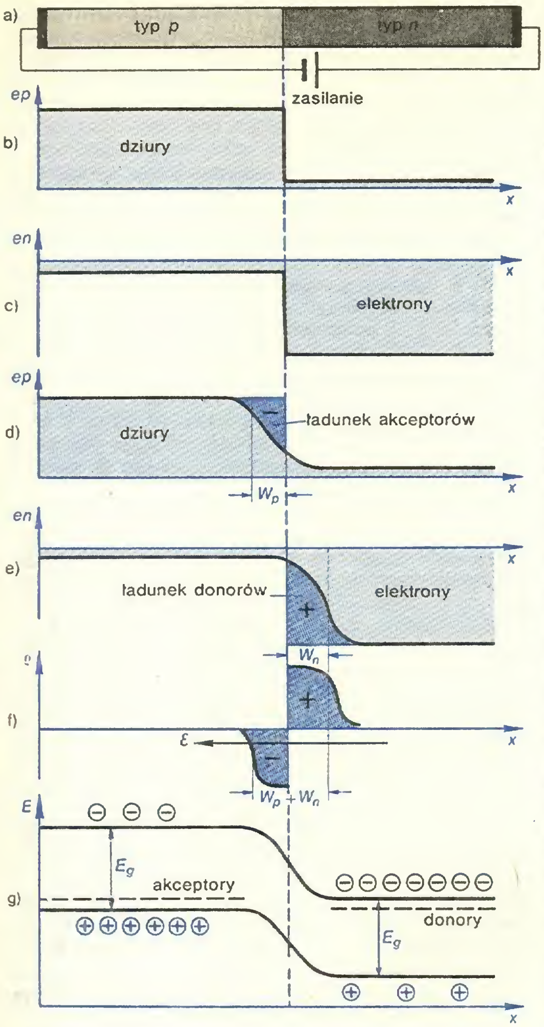
Dotychczas rozważaliśmy przepływ prądu przez próbki półprzewodnikowe jednorodne. Oznacza to, że rozmieszczenie domieszek donorowych lub akceptorowych, a co za tym idzie — koncentracje nośników są jednakowe we wszystkich miejscach próbki.

struktura półprzewodnikowa

Najprostszym rodzajem półprzewodnika niejednorodnie domieszkowanego, zwanego strukturą półprzewodnikową, jest bryła półprzewodnikowa, w której wytworzono dwa sąsiadujące ze sobą obszary o różnym typie przewodnictwa (obszar typu p i obszar typu n), czyli złącze p-n. Model złącza p-n przedstawiono na rys. 4a. Do obydwu części, p i n, bryły półprzewodnika dołączono elektrody dla umożliwienia przepuszczania prądu przez złącze. W konkretnym przyrządzie półprzewodnikowym może występować jedno lub więcej takich złącz.

model złącza *p-n*

Niemożliwe jest oczywiście uzyskanie złącza przez zwykłe zetknięcie dwóch bryłek o różnym typie przewodnictwa. Możemy jednak wykonać taki eksperyment myślowo, łącząc płaszczyznami dwie jednorodne płytki półprzewodnikowe typu n oraz typu p. Koncentracje, a właściwie ładunki elektryczne dziur i elektronów przed ścisłym zetknięciem obu obszarów,



Rys. 4. Powstawanie bariery potencjału w złączu *p-n*: a) model przestrzenny, b) rozkład ładunku dziur przed połączeniem obszarów *p* i *n*, c) rozkład ładunku elektronów przed połączeniem obszarów *p* i *n*, d) rozkład ładunku dziur po połączeniu obszarów *p* i *n*, e) rozkład ładunku elektronów po połączeniu obszarów *p* i *n*, f) rozkład gęstości ładunku przestrzennego *ρ* w warstwie podwójnej, g) przebieg pasm energii elektronów w złączu *p-n* bez przepływu prądu

przedstawiają rys. 4b i c. Po połączeniu obszarów *n* i *p*, umożliwiającym przepływ nośników ładunku, taki rozkład koncentracji nie utrzymuje się i dziury z obszaru *p*, gdzie ich jest bardzo dużo, dyfundują do obszaru *n*, gdzie jest ich bardzo mało. Rekombinują tam z elektronami, ponieważ nośniki nadmiarowe zakłócają równowagę termodynamiczną. W rezultacie — w obszarze *n* przybywa niewiele dziur, natomiast ubywa ich w obszarze *p*, a więc nośniki większościowe odsuwają się niejako od płaszczyzny rozgraniczającej obszar *p* od obszaru *n*, zostawiając po sobie warstwę zubożoną, tzn. obszar o znacznie zmniejszonej koncentracji nośników większościowych przy niewiele zmienionej koncentracji nośników mniejszościowych. Dziury, cofając się w obszar *p*, zostawiają po sobie nieruchome zjonizowane akceptory. Powstaje więc obszar przestrzennego ładunku ujemnego (zaznaczony na rys. 4d kolorem niebieskim), wzrastającego w miarę pogrubiania się warstwy zubożonej. Ładunek ten wytwarza bardzo silne pole elektryczne, przeciwdziałające dyfuzji dziur. Grubość warstwy zubożonej W_p zależy od stałej dyfuzji nośników ładunku oraz od własności dielektrycznych półprzewodników. Warstwa W_p w półprzewodnikach o oporze właściwym ok. 1 Ω cm wynosi mniej niż 1 μm = 10^{-6} m. Jest to wielkość do tysiąca razy mniejsza niż droga dyfuzji nośników nadmiarowych (0,01–1 mm).

przestrzenny ładunek ujemny

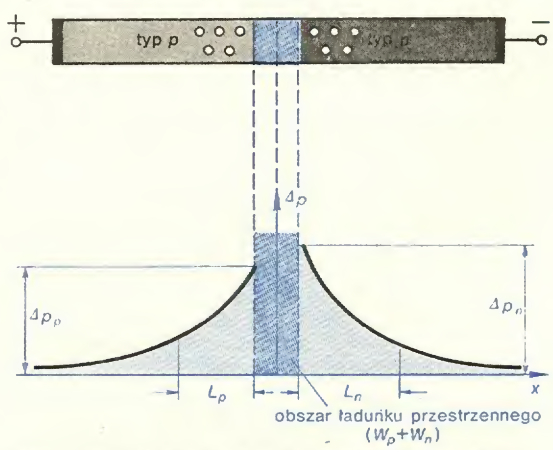
przestrzenny ładunek dodatni

Podobnie zachowują się elektrony w obszarze *n*. Tam także powstaje warstwa przestrzennego ładunku elektrycznego, ale dodatniego (rys. 4e), ponieważ cofające się elektrony zostawiają po sobie dodatnio zjonizowane donory. W rezultacie — po jednej stronie powierzchni granicznej *p-n* powstaje warstwa ładunków ujemnych, po drugiej — dodatnich. Wytwarza się więc układ ładunków taki jak w naładowanym kondensatorze, gdzie na jednej okładce występuje warstwa ładunków dodatnich, na drugiej zaś ujemnych. Naładowanie kondensatora powoduje powstanie napięcia elektrycznego pomiędzy ładunkami okładek. Oznacza to, że energie potencjalne ładunków jednej i drugiej warstwy są różne. Analogiczna sytuacja zachodzi w złączu *p-n*: wytworzenie się pola elektrycznego w obszarze między warstwami ładunków i powstanie różnicy energii potencjalnych nośników ładunku między jedną a drugą stroną złącza. Energia potencjalna elektronów jako nośników ładunku ujemnego jest wyższa po stronie warstwy naładowanej dodatnio. Przebieg tej energii odwzorowują poziomy energii pasm — powstaje w ten sposób bariera energetyczna złącza *p-n* przedstawiona na rys. 4g. Jeśli przez złącze nie przepływa prąd, wysokość bariery jest nieco mniejsza od wielkości przerwy energetycznej E_g. Jeśli do elektrod złącza przykładamy napięcie ze źródła prądu (tzw. napięcie polaryzacji), następuje podwyższenie lub obniżenie bariery. Ma to miejsce, gdy do obszaru *p* przykładamy dodatni, a do obszaru *n* ujemny biegun źródła. Dziury z obszaru *p* przepływają wtedy do obszaru *n*, czyli powstaje prąd dziurowy. Elektrony poruszają się w stronę przeciwną, tworząc prąd o zwrocie zgodnym z prądem dziurowym (ponieważ znak ich ładunku jest przeciwny). Dzięki obniżeniu bariery prądy te osiągają znaczne wartości, w związku z czym kierunek od *p* do *n* nazywamy kierunkiem przewodzenia.

bariera potencjału

Wskutek dyfuzji nośników większościowych z jednego obszaru do drugiego w sąsiadujących ze sobą obszarach o przeciwnym typie przewodnictwa, po obu stronach powierzchni granicznej powstają obszary o nierównowagowej koncentracji nośników, wnikające na głębokości równe drogom dyfuzji L_D nośników nadmiarowych. Jest to zjawisko analogiczne do zjawiska powstawania nośników nadmiarowych pod wpływem światła, a następnie ich dyfuzji w głąb półprzewodnika na głębokość właśnie równą L_D. Teraz jednak nośniki nadmiarowe powstają wskutek przepływu prądu przez złącze, są więc niejako „wstrzykiwane" zarówno z obszaru *p* do *n*, jak z *n* do *p*. Zjawisko to ma ogromne znaczenie w przyrządach półprzewodnikowych (np. diodach elektrolumine-

wstrzykiwanie nośników nadmiarowych



Rys. 5. Rozkład wstrzykiwanych nośników nadmiarowych w złączu *p-n* spolaryzowanym w kierunku przewodzenia

scencyjnych, tranzystorach) i nosi nazwę wstrzykiwania nośników nadmiarowych. Przykładowy rozkład nośników nadmiarowych w złączu pracującym w kierunku przewodzenia przedstawia rys. 5. Koncentracja nośników nadmiarowych zależy wykładniczo od napięcia polaryzacji U, a ściśle — od energii nośników eU dzielonej przez kT. Wyraża to funkcja $e^{eU/kT}$. Gęstość prądu dyfuzji nośników większościowych J_d jest więc proporcjonalna do funkcji wykładniczej:

rozkład nośników nadmiarowych

$$J_d = J_s e^{eU/kT}, \tag{5}$$

przy czym współczynnikiem proporcjonalności jest pewna wartość gęstości prądu J_s. W temperaturze pokojowej (290 K):

$$J_d = J_s e^{40U} = J_s \cdot 10^{17,4\,U}. \tag{6}$$

Podwyższenie napięcia polaryzacji U zaledwie o 0,06 V powoduje dziesięciokrotny wzrost natężenia prądu. Jednakże wzór ten można stosować tylko do wartości eU mniejszych od wartości energii wzbronionej E_g, w przeciwnym bowiem razie bariera znika i prąd podlega prawu Ohma, jeśli w ogóle złącze nie ulegnie zniszczeniu wskutek przegrzania.

polaryzacja zaporowa

Jeżeli przyłączymy zewnętrzne napięcie polaryzacji złącza w kierunku przeciwnym niż w dotychczasowym rozważaniu, a więc w ten sposób, że do elektrody części p złącza dołączony będzie ujemny biegun źródła prądu, to taką polaryzację nazwiemy wsteczną lub zaporową. Dyfuzja nośników większościowych już przy niewielkich napięciach wstecznych będzie całkowicie zahamowana przez silne pole elektryczne. Ponadto dziury w części p zostaną przyciągnięte do elektrody ujemnej, elektrony zaś w części n — do elektrody dodatniej, tak że nośniki większościowe obu znaków cofną się znacznie bardziej w głąb swoich obszarów niż w warunkach równowagi. Zostawią one po sobie, jak to przedstawia rys. 6, obszary naładowane dodatnio po stronie n, ujemnie zaś po stronie p. W ten sposób ładunek przestrzenny po obu stronach warstwy podwójnej będzie znacznie zwiększony i wniknie na duże głębokości w obszary p i n. Bariera potencjału, którą wytworzy taka warstwa podwójna, będzie więc wyższa niż w warunkach równowagi, a wielkość jej podwyższenia będzie równa przyłożonemu napięciu polaryzacji wstecznej. W kierunku zaporowym można przykładać bardzo duże napięcie — w odpowiednio wykonanych złączach przekraczające 1000 V. Można uważać, że złącze spolaryzowane w kierunku wstecznym składa się z przewodzących obszarów p i n przedzielonych warstwą izolatora grubości $W_n + W_p$, zwaną warstwą zaporową, ponieważ w tym obszarze, pozbawionym prawie nośników, półprzewodnik ma właściwości izolatora. W ten sposób złącze p-n jest kondensatorem i to takim, w którym możemy zmieniać grubość warstwy izolatora, a więc o zmiennej pojemności. Zjawisko to ma zastosowanie w tzw. waraktorach.

warstwa zaporowa

przebicie lawinowe

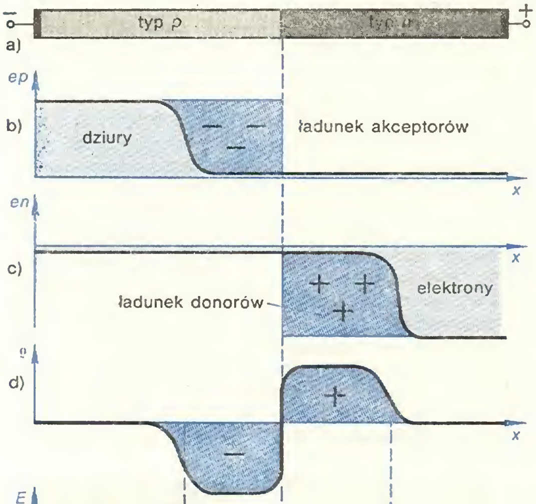
W miarę wzrostu napięcia wstecznego wzrasta natężenie pola elektrycznego wewnątrz warstwy zaporowej. Gdy osiąga ono pewną wartość, następuje tzw. lawinowe przebicie złącza, połączone z gwałtownym wzrostem prądu. Jeśli się nie przekroczy pewnych wartości prądu, złącze nie ulegnie zniszczeniu. Zjawisko to jest wykorzystywane w diodach lawinowych.

prąd nasycenia

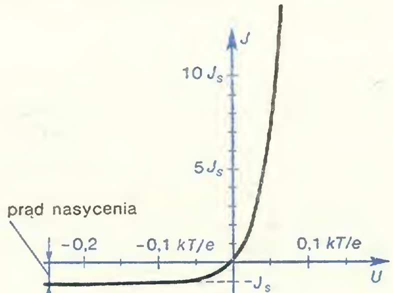
Mimo że warstwa zaporowa ma — wskutek zahamowania dyfuzji nośników większościowych — własności izolacyjne, przepływa przez nią prąd dyfuzji nośników mniejszościowych, zwany prądem nasycenia. Prąd ten jest niezależny od przyłożonego napięcia (zarówno w kierunku zaporowym, jak i przewodzenia), stanowi stałą wartość na tle prądu przewodzenia, a przy tym ma zwrot przeciwny, należy więc jego wartość odjąć od wartości prądu określonej wzorem (5). Ponieważ wypadkowy prąd płynący przez złącze przy napięciu $U = 0$ winien być równy zeru, przeto wartość gęstości prądu nasycenia musi być równa J_s. Zatem gęstość wypadkowego prądu płynącego przez złącze wynosi:

$$J = J_s e^{eU/kT} - J_s = J_s\,(e^{eU/kT} - 1). \tag{7}$$

Widać, że spełniony jest warunek równowagi złącza przy braku zewnętrznego napięcia: gdy $U = 0$, to $J = 0$. Powyższy wzór określa zależność natężenia przepływającego prądu od napięcia przyłożonego do złącza, zilustrowaną na wykresie (rys. 7).



Rys. 6. Podwyższenie bariery potencjału w złączu p-n spolaryzowanym wstecznie: a) model przestrzenny, b) rozkład ładunku dziur, c) rozkład ładunku elektronów, d) rozkład gęstości ładunku przestrzennego ρ w warstwie podwójnej, e) przebieg pasm energii elektronów



Rys. 7. Charakterystyka prądowo-napięciowa idealnego złącza p-n w temperaturze $T = 290$ K

Prąd nasycenia o gęstości J_s zależy od właściwości obu materiałów p i n w sposób następujący:

$$J_s = \frac{eD_p n_p}{L_p} + \frac{eD_n p_n}{L_n}, \tag{8}$$

gdzie n_p jest koncentracją nośników mniejszościowych (elektronów) w obszarze p, p_n — koncentracją dziur w obszarze n, a L_p, L_n — drogi dyfuzji nośników mniejszościowych odpowiednio w obszarze p i n.

Aby złącze przepuszczało mały prąd w kierunku wstecznym, prąd nasycenia musi być również mały. Drogi dyfuzji L_p i L_n, a co za tym idzie — czasy życia τ_p i τ_n nośników mniejszościowych muszą więc być duże. Można to uzyskać przez ograniczenie rekombinacji w obszarze złącza dzięki dobrze opracowanej technologii, przy której struktura kryształu jest jak najmniej naruszona i nie wprowadza się niepożądanych domieszek. Fakt, że opór złącza w kierunku przewodzenia jest mały, a w kierunku zaporowym bardzo duży, jest podstawą jednego z ważniejszych rodzajów zastosowania złącz *p-n*, a mianowicie prostowania prądu zmiennego (→ Przyrządy półprzewodnikowe dyskretne).

Oddziaływanie nośników nadmiarowych na złącze *p-n*. Zjawisko tranzystorowe

Spolaryzowanie złącza w kierunku przewodzenia pociągało za sobą zjawisko wstrzykiwania nośników nadmiarowych do obu części złącza (wskutek obniżenia się jego bariery potencjału). Okazuje się, co potwierdza teoria i doświadczenie, że występuje też zjawisko odwrotne: obniżanie się bariery potencjału pod wpływem wytworzonych w obszarze złącza nośników nadmiarowych. Oznacza to pojawienie się napięcia polaryzacji na elektrodach złącza nie dołączonego do źródła energii elektrycznej. Nośniki nadmiarowe mogą powstawać np. pod wpływem światła, a na elektrodach złącza pojawia się wówczas napięcie elektryczne. Jest to złączowe zjawisko fotowoltaiczne. Znajduje ono zastosowanie w przyrządach do pomiaru natężenia światła oraz do przetwarzania energii świetlnej w energię elektryczną, np. w bateriach słonecznych wykonywanych z krzemu, których sprawność energetyczna przekracza 10%; jest to osiągnięcie o dużym znaczeniu praktycznym.

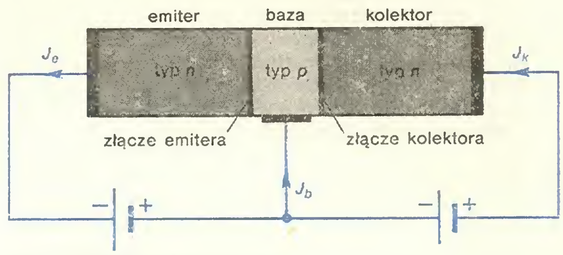
złączowe zjawisko fotowoltaiczne

Do oświetlonego złącza można przyłożyć napięcie. Każdemu napięciu polaryzacji będzie odpowiadać pewne natężenie prądu, inne niż w złączu nie oświetlonym. Wszystkie pary dziura–elektron generowane strumieniem świetlnym w odległości mniejszej niż ich droga dyfuzji dotrą do złącza, gdzie ulegną rozdzieleniu pod wpływem pola. Elektrony będą się poruszać od obszaru *p* do obszaru *n* złącza, dziury zaś w kierunku przeciwnym. Towarzyszące ruchom obu rodzajów nośników prądy elektryczne sumują się ze względu na różnicę znaków ładunku. Wypadkowy prąd ma zwrot zgodny z ruchem dziur od *n* do *p*. Jest to zwrot prądu wstecznego, a więc nośniki nadmiarowe zwiększają (co do bezwzględnej wartości) prąd wsteczny, co psuje własności złącza jako prostownika. Urządzenie, w którym jeden obszar złącza *p-n* może być oświetlony, zostało zastosowane jako przyrząd zwany fotodiodą (→ Optoelektronika półprzewodnikowa).

polaryzacja złącza oświetlonego

fotodioda

W fotodiodzie wykorzystane jest oddziaływanie nośników nadmiarowych wytworzonych za pomocą światła na wstecznie spolaryzowane złącze *p-n*. Zamiast wytwarzania nośników nadmiarowych za pomocą światła można wykorzystać zjawisko ich wstrzykiwania. W tym celu należy w pobliżu wstecznie spolaryzowanego złącza *p-n* umieścić złącze *n-p* i spolaryzować je w kierunku przewodzenia. Powstanie w ten sposób struktura pokazana na rys. 8. Obszar *p*,



Rys. 8. Struktura tranzystora *n-p-n*

wspólny dla obu złącz, jest zaopatrzony w osobną elektrodę i nosi nazwę bazy. Obszar *n* spolaryzowany zaporowo w stosunku do bazy nazywa się kolektorem (łac. *colligere* 'zbierać'), natomiast przeciwległy obszar *n*, spolaryzowany w kierunku przewodzenia, nazywa się emiterem (łac. *emittere* 'wysłać'). Prąd przepływający przez złącza emitera wstrzykuje do obszaru *p* nośniki nadmiarowe, które docierają do złącza kolektora, powodując przepływ dużego prądu przez to złącze (spolaryzowane w kierunku zaporowym). Jest to zjawisko tranzystorowe, odkryte w 1948 r. przez Bardeena i Brattaina. Stało się ono punktem zwrotnym w ogromnym i szybkim rozwoju elektroniki półprzewodników. Najważniejszym zastosowaniem zjawiska tranzystorowego jest wzmacnianie sygnałów elektrycznych. Napięcie między emiterem i bazą wynosi ułamek wolta, a między kolektorem i emiterem może wynosić od kilku do kilkuset woltów. Jednocześnie prąd płynący między kolektorem a emiterem jest dużo większy niż prąd między bazą i emiterem. Dzięki temu sygnał elektryczny małej mocy wprowadzony do obwodu baza–emiter powoduje pojawienie się sygnału elektrycznego większej mocy w obwodzie kolektor–emiter.

zjawisko tranzystorowe

A. K. Jonscher *Podstawy działania przyrządów półprzewodnikowych*, Warszawa 1962; I. J. Kampel *Półprzewodniki. Teoria i zastosowanie*, Warszawa 1974; A. Świt, J. Pułtorak *Przyrządy półprzewodnikowe*, Warszawa 1976.

Przyrządy półprzewodnikowe dyskretne

Jarosław Świderski

Od „kryształka" do kryształka

Elektronika zajmuje się przekształcaniem sygnałów elektrycznych przez wpływanie bezpośrednio na ruch elektronów. Początkowo operacje takie najłatwiej było przeprowadzać w próżni, działając na strumień elektronów polem elektrycznym i magnetycznym sterowanym „makroskopowo" przez regulację wytwarzającego je prądu elektrycznego (elektronika próżniowa — lampy elektronowe).

W siedemdziesiątych latach ubiegłego stulecia odkryto, że styk metalu z półprzewodnikiem ma właściwości prostujące. Fakt ten umożliwił wytworzenie prymitywnych elementów półprzewodnikowych, tzw. kryształków, i budowanie przy ich użyciu kryształkowych radiodobiorników. Kryształki te to nic innego jak pierwsze diody półprzewodnikowe, w których się wykorzystuje odpowiednią konfigurację ciał krystalicznych dla formowania wiązki (strumienia) elektronów płynących przez ciało stałe. Ponieważ w diodzie półprzewodnikowej ruch elektronów odbywa się w obszarze znacznie mniejszym niż w analogicznych przyrządach elektroniki próżniowej, pierwsze poważniejsze zastosowanie półprzewodników związane było

kryształki prostujące

z techniką wielkich częstości (np. radiolokacja), gdzie konieczne jest uzyskanie jak najmniejszych pojemności i indukcyjności czynnego elementu. Drugą cechą szczególną przyrządów półprzewodnikowych, dającą im przewagę nad próżniowymi, było wykorzystywanie ruchu elektronów w ich jak gdyby naturalnym środowisku. Aby sterować elektronami w próżni, trzeba je tam wprowadzić, np. przez termoemisję z rozgrzanego do wysokiej temperatury drutu wolframowego (stąd świecenie katod w lampach radiowych). Swobodne elektrony w półprzewodniku egzystują w normalnej temperaturze w dostatecznej dla „elektronicznych" celów liczbie bez ingerencji człowieka, a więc przyrządy półprzewodnikowe nie wymagają skomplikowanych, energochłonnych i nietrwałych urządzeń zmuszających elektrony do akcji (na przykład układów zasilających żarzenie katod w lampach radiowych).

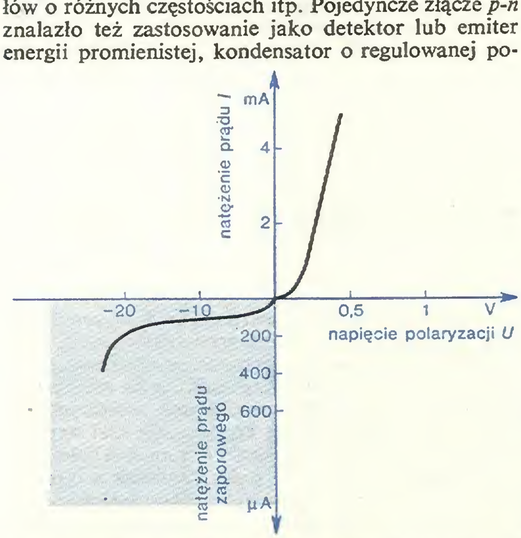
zastosowanie monokryształów

Dalszy rozwój elektroniki półprzewodnikowej to już jedno pasmo triumfów technologii. Niepewne, nie zawsze powtarzalne styki metal–półprzewodnik zastąpiono monokrystalicznym materiałem półprzewodnikowym, w którym minimalne ilości obcych atomów zmieniają typ przewodnictwa w części obszaru czynnego z elektronowego (*n*) na dziurowy (*p*) lub odwrotnie (→ Fizyka przyrządów półprzewodnikowych). Rolę prostownika przejęła powierzchnia graniczna między obszarami o różnym typie przewodnictwa. Granice te, czyli złącza *p-n*, można dziś rozmieszczać w materiale półprzewodnikowym z dokładnością do dziesiątych części mikrometra, a odpowiednio je dobierając — tworzyć struktury dowolnie niemal skomplikowanych przyrządów (→ Mikroelektronika). Z historycznego „kryształka", który był zlepkiem polikrystalicznych ziaren dociskanych metalową igłą, pozostała tylko, i to już coraz rzadziej używana, nazwa: kryształek, czyli kawałeczek (wymiary od dziesiątych części do pojedynczych milimetrów) półprzewodnikowego monokryształu z wbudowanymi złączami *p-n* i metalowymi (najczęściej napylonymi) elektrodami. Ale i ta nazwa ustępuje nowszym: półprzewodnikowa struktura lub z angielska — czip.

Diody i tranzystory

funkcje diody próżniowej

Właściwości prostujące złącza *p-n* lub złącza metal–półprzewodnik wykorzystano w diodzie półprzewodnikowej, przejmującej wszystkie funkcje spełniane poprzednio przez diody próżniowe: prostowanie prądów zmiennych, ich generację, wykrywanie (detekcja) sygnałów elektromagnetycznych, „mieszanie" sygnałów o różnych częstościach itp. Pojedyncze złącze *p-n* znalazło też zastosowanie jako detektor lub emiter energii promienistej, kondensator o regulowanej pojemności, detektor naprężeń mechanicznych, ogranicznik napięcia, skomplikowany przetwornik funkcji matematycznych itd. Niektóre z tych przyrządów zostały szerzej opisane w rozdziałach poświęconych mikroelektronice, optoelektronice czy elektronice mikrofalowej. Jak z tego widać, pojęcie diody i jej zakres zastosowania są bardzo szerokie. Zawsze jednak jest to dwuelektrodowy przyrząd półprzewodnikowy o nieliniowej charakterystyce prądowo-napięciowej (rys. 1), tj. o takiej zależności prądu płynącego przez diodę od przyłożonego napięcia, że można rozróżnić kierunek przewodzenia i kierunek wsteczny (zaporowy). Jeśli do diody półprzewodnikowej przyłożone jest napięcie w ten sposób, że dodatni potencjał znajduje się na elektrodzie połączonej z obszarem typu *n*, mówimy o spolaryzowaniu jej w kierunku zaporowym, tj. w kierunku, w którym ma większy opór elektryczny. Przez diodę płynie wtedy prąd zaporowy (inaczej wsteczny), składający się z prądu ładunków mniejszościowych, powstających pod wpływem generacji termicznej, i z prądu upływu. Pierwszy nie zależy od przyłożonego napięcia, a jedynie od temperatury i parametrów rekombinacyjnych, drugi, proporcjonalny do przyłożonego napięcia, „kryje" w sobie niedoskonałości złącza. W pierwszym przybliżeniu można sobie diodę spolaryzowaną zaporowo wyobrazić jako równoległe połączenie diody idealnej (przez którą płynie prąd zaporowy) i opornika (przez który płynie prąd upływu).



Rys. 1. Przykładowa charakterystyka prądowo-napięciowa diody półprzewodnikowej

dioda półprzewodnikowa

zjawisko Zenera

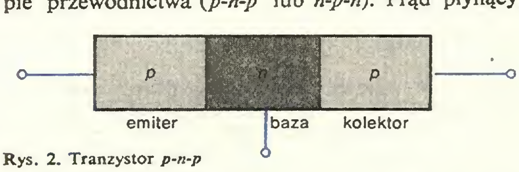
Przy dostatecznie dużym napięciu (tzw. napięciu przebicia) prąd zaporowy zaczyna szybko rosnąć wskutek generacji par elektron–dziura, spowodowanej zerwaniem pewnych wiązań kowalencyjnych między sąsiednimi atomami i tunelowym przenikaniem elektronów walencyjnych przez pasmo wzbronione (tzw. zjawisko Zenera), oraz wskutek powielania lawinowego (generacji par elektron-dziura pod wpływem dostatecznie „rozpędzonych" elektronów). Napięcie przebicia związane ze zjawiskiem Zenera maleje ze wzrostem temperatury, a związane ze zjawiskiem powielania — rośnie.

Przez diodę spolaryzowaną w kierunku przewodzenia, tj. w kierunku, w którym ma ona mniejszy opór, płynie prąd zależny od napięcia początkowo wykładniczo, a następnie liniowo, gdyż dioda zaczyna się wówczas zachowywać jako opornik.

Bardziej złożone struktury, zawierające co najmniej dwa złącza *p-n*, złącze i dwa kontakty omowe (nie prostujące styki metal–półprzewodnik) lub kombinację złącz *p-n*, kontaktów i kondensatorów, umożliwiają wzmacnianie sygnałów elektrycznych lub ich przełączanie.

tranzystor bipolarny

Trójelektrodowy przyrząd półprzewodnikowy zwany tranzystorem (triodą półprzewodnikową) jest drugim, obok diody, podstawowym przyrządem współczesnej elektroniki, elementarnym składnikiem przygniatającej większości układów scalonych (mikroukładów). Najpopularniejsze obecnie rozwiązania — to tranzystor bipolarny o dwóch złączach *p-n* i tranzystor polowy MOS, w którym przepływ ładunków elektrycznych jest sterowany polem kondensatora. Tranzystor bipolarny zbudowany jest z płytki półprzewodnikowej zawierającej trzy obszary (emiter, baza, kolektor; rys. 2) o kolejno zmieniającym się typie przewodnictwa (*p-n-p* lub *n-p-n*). Prąd płynący



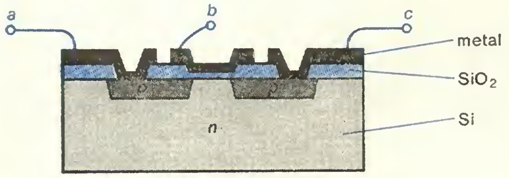
Rys. 2. Tranzystor *p-n-p*

przez spolaryzowaną wstecznie diodę kolektor–baza jest sterowany przez prąd spolaryzowanej w kierunku przewodzenia diody emiter–baza, gdyż nośniki wprowadzane z obszaru emitera dyfundują (lub w szczególnym wypadku są unoszone polem) przez obszar bazy

i zwiększają, jako dodatkowe nośniki mniejszościowe, prąd wsteczny płynący z kolektora. Obwód diody emiter–baza (spolaryzowanej w kierunku przewodzenia) odznacza się oczywiście małym oporem, w przeciwieństwie do obwodu z diodą kolektor–baza.

tranzystor polowy

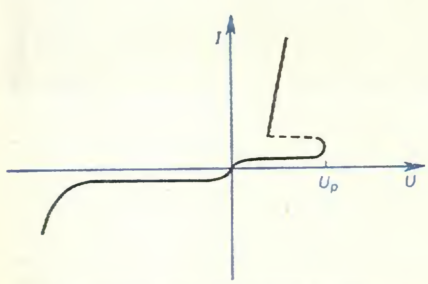
Tranzystor polowy (rys. 3) wykonany jest w zasadzie z krzemu. Nazwa MOS pochodzi od skrótu angielskich słów metal–tlenek–półprzewodnik. Używa się też czasem nazwy MIS, w której słowo tlenek zostało zastąpione bardziej ogólnym — izolator. Oznaczenie p^+ stosuje się do półprzewodnika o przewodnictwie typu p silnie domieszkowanego. Prąd płynący między elektrodami a i c (czyli między dwoma obszarami p^+) w wąskiej, przypowierzchniowej warstewce obszaru n (tzw. kanałem) jest sterowany potencjałem przyłożonym do elektrody b (bramki), która wraz z bardzo cienką (ułamek µm) warstwą tlenku i powierzchnią półprzewodnika stanowi kondensator o małej stratności. Tranzystory MOS pracują przy niższych częstościach niż tranzystory bipolarne, ich wytwarzanie jest jednak znacznie łatwiejsze, a ponadto sterowanie kondensatorowe zużywa znacznie mniej energii. Stąd ich główne zastosowanie w monolitycznych układach scalonych wielkiej skali integracji (→ Mikroelektronika).



Rys. 3. Tranzystor MOS

przyrządy wieloelektrodowe

Tranzystory i diody stanowią elementy podstawowe układów bardziej złożonych, same również przybierają nieraz skomplikowane formy wieloelektrodowe. Można tu wymienić np. tranzystor tyratronowy, czyli tyrystor, spełniający funkcję sterowanego zaworu (dwustanowego elementu przełączającego). Składa się on z czterech obszarów o kolejno zmieniających się typach przewodności: *p-n-p-n*, i ma charakterystykę prądowo-napięciową podobną nieco (rys. 4) do charakterystyki diody — z tą różnicą, że przełączenie go w stan przewodzenia wymaga przyłożenia odpowiedniego napięcia, tzw. napięcia przeskoku (U_p), niższego od napięcia przebicia i dającego się sterować za pomocą prądu bazy.



Rys. 4. Charakterystyka prądowo-napięciowa tyrystora

Miniaturyzacja, mikrominiaturyzacja i co dalej?

Zwiększenie liczby złącz, kontaktów, kondensatorów itp. prowadzi do tworzenia skomplikowanych układów indywidualnych, a następnie do układów scalonych. Zmniejszanie wymiarów elementów czynnych już jest celem dość atrakcyjnym, a łączy się z tym prawie zawsze zmniejszenie ciężaru, zasilania i przede wszystkim zwiększenie niezawodności. Skok jakościowy spowodowany wprowadzeniem elementów półprzewodnikowych można obrazowo przedstawić modnym w pierwszych latach „panowania" półprzewodników porównaniem: „Gdyby w czasie lądowania aliantów w Normandii w 1944 r. były już znane i stosowane tranzystory, ciężar przerzucanego na francuski brzeg ekwipunku przypadającego na jednego żołnierza zmalałby w przybliżeniu o jedną tonę".

Wprowadzenie technologii epiplanarnej umożliwiło następny skok jakościowy. Zamiast wycinania z monokryształów jak najcieńszych płytek, co zawsze naruszało strukturę krystaliczną czynnych obszarów półprzewodnika, zastosowano tzw. warstwy epitaksjalne, tj. bardzo cienkie (od pół do kilkunastu mikrometrów) monokrystaliczne warstwy, narastające w odpowiednio sterowanym procesie na podłożu dającym wytrzymałość mechaniczną i (przeważnie) orientację krystalograficzną. Płytki te (w technologii epiplanarnej stosuje się przede wszystkim krzem) pokrywa się jeszcze o rząd cieńszymi warstewkami świetnego izolatora, jakim jest tlenek krzemu; odgrywa on zarówno rolę ochrony przed atmosferą, jak i maski (otwory wytrawiane po odpowiednim naświetleniu specjalnej emulsji — stąd precyzja, na jaką stać tylko optykę), przez którą się dozuje domieszki dające odpowiedni typ przewodnictwa. Te dwa elementy technologiczne — epitaksja i pokrycie krzemu jego tlenkami, czyli pasywacja — przyniosły rewolucję, powstały diody i tranzystory, w których problem wymiarów sprowadzał się do problemu dotarcia do samej struktury (czipu). Najcieńsze druty były grubsze od niektórych czynnych obszarów, nie mówiąc już o ograniczeniach wymiarów narzucanych przez możliwości manipulowania oprawkami. Doszły też nowe problemy, które poprzednio ledwo dawały o sobie znać: układowe i materiałowe. Do pierwszych należy sprawa iloczynu maksymalnej mocy, jaka może być tracona w danym elemencie (diodzie, tranzystorze itp.), i maksymalnej częstości sygnału, przy jakiej ten element zachowuje jeszcze swoje właściwości. Oczywiście — im mniejszy element, tym łatwiej zachować np. dostatecznie małe pojemności potrzebne przy dużych częstościach. Ale jednocześnie przy tym samym prądzie mniejszy element bardziej się nagrzewa i zmienia swoje właściwości z przyczyn natury termicznej. Stąd tendencja do budowania elementów o wielu odpowiednio połączonych obszarach stanowiących jeden przyrząd. I tak np. nowoczesny tranzystor dużej mocy i dużej częstości, tzw. *overlay*, składa się z kilkudziesięciu elementów, co czyni go podobnym raczej do układu scalonego, i to średniej skali integracji (rys. 5).

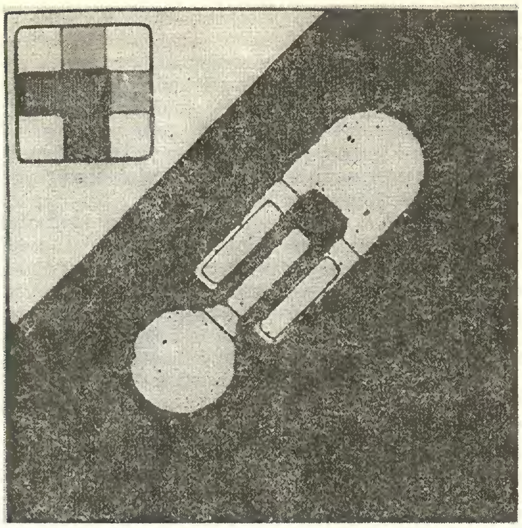
warstwy epitaksjalne

pasywacja

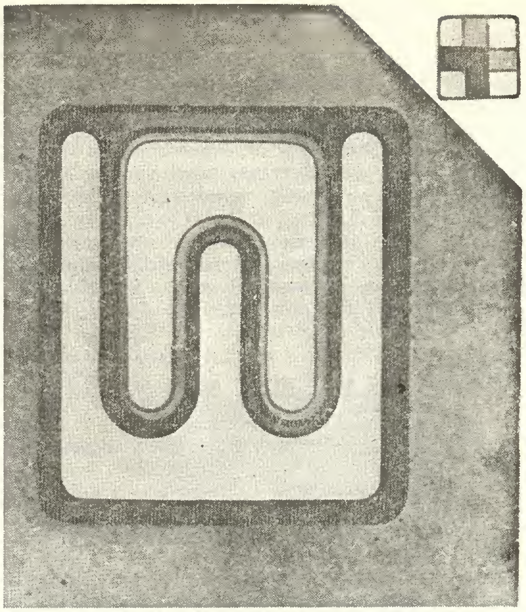
problemy materiałowe

Ogromne znaczenie w całej elektronice półprzewodników mają problemy materiałowe. Wielka czystość pierwiastków i związków półprzewodnikowych, której nie można określić nawet najczulszymi metodami chemicznymi, doskonałość sieci krystalicznej półprzewodników, jednorodność parametrów elektrycznych — oto wymagania, jakich nie stawia żadna inna gałąź techniki. Miniaturyzacja potęguje te wymagania w sposób oczywisty; niejednorodność np. występująca na powierzchni 1 μm^2 mogła być dopuszczalna w złączu o powierzchni 10 mm^2, ale dyskwalifikuje ona złącze o powierzchni milion razy mniejszej. Mikrominiaturyzacja, w ślad za którą idzie tworzenie struktur wielozłączowych, narzuca warunek, by taka niejednorodność nie występowała w żadnym złączu wchodzącym w skład danego przyrządu, gdyż zdyskwalifikuje to całą strukturę. Powyższe wymagania skierowały znaczne potencjały badawcze na zagadnienia związane z pomiarami parametrów elektrycznych i strukturalnych materiałów półprzewodnikowych. Wyposażenie laboratoriów pomiarowych stanowi dziś najpoważniejszą pozycję zarówno w budżecie instytutów, jak zakładów przemysłowych działających w dziedzinie elektroniki półprzewodnikowej.

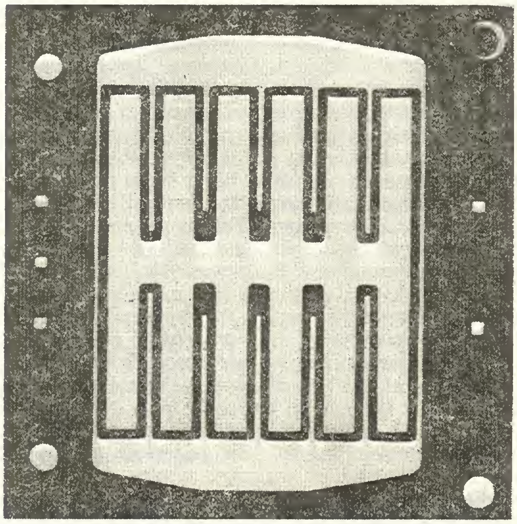
struktury tranzystorów



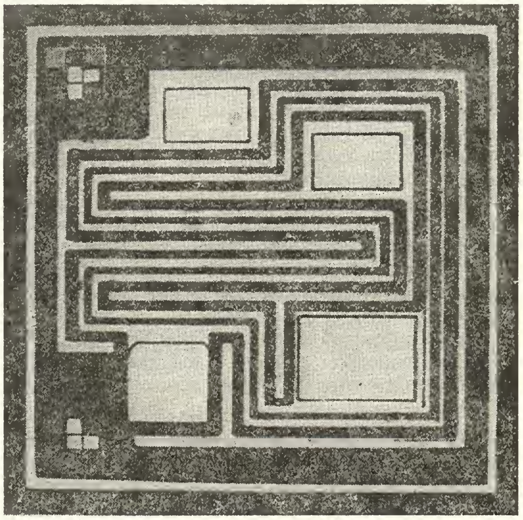
a)



b)



c)

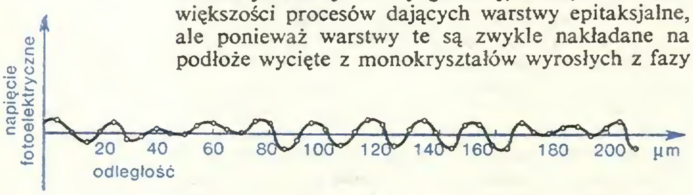


d)

Rys. 5. Struktury tranzystorów: a) i b) tranzystory impulsowe, c) tranzystor overlay, d) tranzystor typu metal-tlenek-półprzewodnik (MOS)

Rozważmy np. zagadnienie tzw. mikroniejednorodności półprzewodników. Otóż wzrost monokryształów z fazy ciekłej nie odbywa się w sposób ciągły, ale jak gdyby drobnymi skokami, co powoduje także skokowy (periodycznie zmieniający się) rozkład domieszek, a więc i większości parametrów elektrycznych półprzewodnika. Rysunek 6 ukazuje zmiany napięcia fotoelektrycznego wzdłuż monokryształu krzemu wywołane mikroniejednorodnościami. Oczywiście w punktach największych odchyleń od średniej koncentracji domieszek oraz w punktach najsilniejszych zmian tych koncentracji (największe „wbudowane" pola elektryczne) najczęściej występują defekty elementów półprzewodnikowych. Wprawdzie mikroniejednorodność przeważnie nie pojawia się w trakcie monokrystalizacji z fazy gazowej, a więc w trakcie większości procesów dających warstwy epitaksjalne, ale ponieważ warstwy te są zwykle nakładane na podłoże wycięte z monokryształów wyrosłych z fazy ciekłej — domieszki dyfundują z tych ostatnich do obszaru czynnego i wytwarzają w nim mikroniejednorodność analogiczną do istniejącej w podłożu. Sama istota powstawania mikroniejednorodności nie jest zbyt dobrze zbadana. Dotychczas udało się uzyskać monokryształy wolne od mikroniejednorodności jedynie w warunkach kosmicznych, w stanie nieważkości (badania nad tym problemem wchodziły między innymi w skład programu pierwszego wspólnego lotu kosmonautów amerykańskich i radzieckich).

mikroniejednorodności



Rys. 6. Rozkład napięcia fotoelektrycznego związany z mikroniejednorodnością, zmierzony wzdłuż monokryształu krzemu

miniaturyzacja

Diody półprzewodnikowe i tranzystory umożliwiły miniaturyzację układów. Skonstruowano miniaturowe odbiorniki radiowe, a przenośne stacje radionadawcze, niegdyś potężne „maszynerie" montowane na ciężarówkach, zmieniły się w chowane w dłoniach pudełka. Powstało wiele miniaturowych urządzeń automatyki przemysłowej.

mikrominiaturyzacja

Dalszy postęp w dziedzinie miniaturyzacji — mikrominiaturyzacja — przyniósł ogromny rozwój elektronicznych maszyn matematycznych, popularnie zwanych komputerami. Wprowadzenie technologii epiplanarnej (i związanej z nią technologii MOS) stworzyło realne szanse budowania komputerów o wymiarach i niezawodności umożliwiających ich obecne rozpowszechnienie, mimo iż nowoczesny

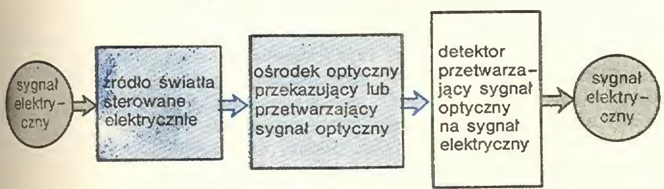
komputer zawiera wiele milionów czynnych elementów. Tradycyjny tranzystor miał wymiary czipu mierzone w milimetrach, a do tego dochodziła spora oprawka z doprowadzeniami. Układ scalony wielkiej skali integracji w strukturze o wymiarach 5×5×0,2 mm mieści kilkanaście tysięcy tranzystorów. Czy technika, wspierana przez fizykę, pójdzie w tym kierunku jeszcze dalej? Na pewno tak.

W. Marciniak *Przyrządy półprzewodnikowe i układy scalone*, Warszawa 1979; W. Rosiński *Zasady działania tranzystorów*, Warszawa 1977; J. Świderski *Techniczne badania właściwości materiałów i struktur półprzewodnikowych*, Warszawa 1975; A. Świt, J. Półtorak *Przyrządy półprzewodnikowe*, Warszawa 1976.

Optoelektronika półprzewodnikowa

Marian A. Herman

Optoelektroniką nazywa się dział elektroniki, którego przedmiotem jest łączne wykorzystanie optycznego jak i elektrycznego sposobu przetwarzania i przekazywania sygnałów. Podstawą optoelektroniki są fizyczne procesy warunkujące przetwarzanie sygnałów elektrycznych na optyczne i sygnałów optycznych na elektryczne oraz procesy wytwarzania, przesyłania, przetwarzania i magazynowania informacji niesionych przez światło. Ogólny schemat układu optoelektronicznego przedstawia rys. 1.



Rys. 1. Schemat układu optoelektronicznego

kanał optyczny

Zaletą wykorzystania kanału optycznego do przesyłania, przetwarzania i magazynowania informacji jest to, że występujące w tym kanale nośniki informacji (fotony) są pozbawione ładunku elektrycznego. Ponieważ czynnikami sterującymi procesem przekazywania lub przetwarzania informacji są w tym wypadku układy wieloatomowe (kryształy lub ciecze, których atomy są naładowane elektrycznie lub wykazują silne właściwości magnetyczne) oddziaływające z fotonami niezbyt silnie (zwłaszcza przy małych natężeniach światła nie wywołujących efektów nieliniowych), to niesiona przez te fotony w kanale optycznym informacja nie ulega niezamierzonym zniekształceniom. Kanał elektryczny szeroko stosowany w klasycznej elektronice, w którym nośniki informacji (elektrony) i czynniki sterujące (pole elektryczne czy pole magnetyczne) oddziałują ze sobą bardzo silnie, jest znacznie bardziej podatny na wszelkiego rodzaju zniekształcenia przesyłanej, przetwarzanej czy magazynowanej informacji. Inną zaletą kanału optycznego jest to, że wysoka częstość sygnałów optycznych (10^{14}–10^{15} Hz) zwiększa o około 10^6 razy pojemność informatyczną tego kanału w stosunku do pojemności kanału elektrycznego. Kanał optyczny zapewnia ponadto doskonałą izolację galwaniczną zacisków wyjściowych, z których odbierana jest informacja, od zacisków wejściowych, do których informacja ta zostaje doprowadzona.

elementy układów optoelektronicznych

Od dawna już znane są takie elementy układów optoelektronicznych, jak różnego rodzaju lampy elektryczne, żarowe czy gazowe, które są sterowanymi elektrycznie źródłami światła, różne materiały dielektryczne ciekłe, szklane czy krystaliczne będące ośrodkami, w których światło się rozchodzi lub jest przetwarzane, oraz takie jak fotokomórka czy fotopowielacz, które są detektorami światła. Zastosowanie tych elementów oraz zbudowanych przy ich użyciu układów było i jest nadal ogromne, zwłaszcza w automatyce czy technice pomiarowej. Duże wymiary wymienionych elementów, a w konsekwencji i duże wymiary całych układów optoelektronicznych, jak i niedoskonałość fizyczna (niespójność) światła używanego jako nośnika nformacji uniemożliwiały jednak realizację w kanale optoelektronicznym tych wszystkich funkcji, które do niedawna realizować można było przy użyciu elektrycznego kanału przekazywania i przetwarzania informacji. Trudno byłoby sobie np. wyobrazić radar optyczny czy telefon optyczny z żarówką lub zwykłą lampą jarzeniową. Dopiero odkrycie na początku lat sześćdziesiątych laserów oraz zastosowanie półprzewodników spowodowało, że optoelektronika zaczęła skutecznie konkurować z klasyczną elektroniką czy telekomunikacją. Wiele danych wskazuje już obecnie na to, że w niedalekiej przyszłości może ona w pewnych wypadkach zastąpić z powodzeniem obie te dziedziny techniki.

optoelektronika półprzewodnikowa

Dział optoelektroniki, w którym do realizacji procesów optoelektronicznych wykorzystuje się półprzewodniki, nosi nazwę optoelektroniki półprzewodnikowej. Znane obecnie półprzewodnikowe przyrządy optoelektroniczne mają takie parametry techniczne, że można je stosować w układzie optoelektronicznym w każdym z przedstawionych na rys. 1 bloków funkcjonalnych (źródło, ośrodek, detektor). Z materiałów półprzewodnikowych można wykonać źródła światła sterowane elektrycznie oraz detektory przetwarzające sygnał świetlny na sygnał elektryczny. Materiały te mogą też stanowić ośrodek optyczny, przez który światło jest przekazywane lub, w którym jest ono przetwarzane na światło o innych właściwościach fizycznych. Największe znaczenie spośród wymienionych półprzewodnikowych przyrządów optoelektronicznych mają źródła i detektory, gdyż mogą one spełniać samodzielnie różne funkcje bez konieczności łączenia ich ze sobą w jeden złożony układ optoelektroniczny.

Półprzewodnikowe źródła światła

Rozwój półprzewodnikowych źródeł światła wiąże się z odkryciem zjawisk katodoluminescencji i elektroluminescencji półprzewodników. Katodoluminescencja — to świecenie (luminescencja) półprzewodnika wywołane napromienieniem go strumieniem elektronów. Elektroluminescencja jest natomiast świeceniem półprzewodnika pod wpływem prądu elektrycznego, który przez niego przepływa. Katodoluminescencja półprzewodników znalazła zastosowanie np. w lampie kineskopowej, która jest nieodłączną częścią składową każdego telewizora. Wykorzystanie tego zjawiska w konstrukcji lampy oscyloskopowej stało się podstawą budowy oscylografu, przyrządu, bez którego trudno byłoby sobie wyobrazić współczesną elektroniczną technikę pomiarową.

katodoluminescencyjne źródła światła

Katodoluminescencyjne źródła światła nie znajdują jednak, jak do tej pory, szerszego zastosowania w układach optoelektronicznych. Jest to związane z ich dużymi wymiarami oraz z tym, że do skutecznego bombardowania półprzewodnika strumieniem elektronów potrzebne jest wytworzenie wysokiej próżni, w której elektrony te mogłyby się poruszać i byłyby odpowiednio przyspieszane.

elektroluminescencyjne źródła światła

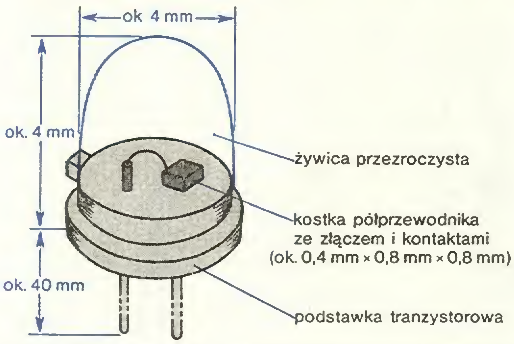
Znacznie szersze zastosowanie w półprzewodnikowych układach optoelektronicznych znalazły elektroluminescencyjne źródła światła. Gdy źródłem światła jest półprzewodnik polikrystaliczny o jednym typie

przewodnictwa, osadzony w postaci cienkich warstw na odpowiednich elektrodach metalowych i świecący pod wpływem przepływającego przezeń stałego lub zmiennego prądu elektrycznego (efekt Destriau), to mamy do czynienia z tak zwaną komórką elektroluminescencyjną lub panelem elektroluminescencyjnym. Gdy natomiast źródłem światła jest półprzewodnikowe złącze *p-n* (→ Fizyka przyrządów półprzewodnikowych) wytworzone w materiale monokrystalicznym i świecące pod wpływem stałego prądu elektrycznego przepływającego przez złącze w kierunku przewodzenia (efekt Łosiewa), to mamy do czynienia z diodą świecącą lub ze złączowym źródłem światła. W zależności od konstrukcji diody oraz od natężenia prądu elektrycznego płynącego przez złącze *p-n*, promieniowanie świetlne emitowane przez źródło złączowe jest albo niespójne — mamy wówczas do czynienia z diodą elektroluminescencyjną (DEL), albo jest ono spójne — źródło jest wówczas laserem złączowym. Diody elektroluminescencyjne i lasery złączowe są obecnie najczęściej stosowane jako źródła światła w układach optoelektronicznych.

komórka elektroluminescencyjna i dioda świecąca

dioda elektroluminescencyjna (DEL)

DEL składa się z małego kawałka monokrystalicznego półprzewodnika, w którym wytworzono złącze *p-n*, przymocowanego za pomocą odpowiedniego spoiwa do podstawki metalowej, oraz z osłony otaczającej i hermetyzującej ten półprzewodnik, wykonanej najczęściej z barwionej żywicy (rys. 2, il. 6 na tabl. 2). Najważniejszym elementem DEL jest złącze



Rys. 2. DEL zamocowana na podstawce tranzystorowej i zanurzona w przezroczystej żywicy

p-n. Gdy zostanie ono spolaryzowane elektrycznie w kierunku przewodzenia, to w obszarze typu *p*, w warstwie o grubości rzędu 1 μm, wytwarza się stan inwersji obsadzeń poziomów energetycznych. Więcej elektronów znajduje się wówczas w pasmie przewodnictwa (o większej energii) niż na górnych poziomach pasma walencyjnego (o mniejszej energii). Oznacza to, że elektrony mogą przejść na puste poziomy pasma walencyjnego i rekombinować z dziurami znajdującymi się po stronie *p* złącza *p-n*. Gdy rekombinacji tej towarzyszy emisja promieniowania elektromagnetycznego, mówimy o rekombinacji promienistej dziur i elektronów w złączu *p-n*. Gdy energia rekombinacji nośników ładunku przekazywana jest sieci krystalicznej półprzewodnika, mówimy o rekombinacji niepromienistej. Podstawowym kryterium wyboru materiałów do produkcji DEL jest stosunek prawdopodobieństw wystąpienia w danym półprzewodniku rekombinacji promienistej i rekombinacji niepromienistej. Im większe jest prawdopodobieństwo rekombinacji promienistej w stosunku do prawdopodobieństwa rekombinacji niepromienistej, tym bardziej dany materiał półprzewodnikowy nadaje się do produkcji DEL. Jeśli jest to materiał o przerwie energetycznej dostatecznie dużej (większej niż 1,7 eV), to emitowane promieniowanie elektromagnetyczne jest promieniowaniem widzialnym. Materiałami używanymi obecnie do produkcji DEL są półprzewodnikowe związki galu z arsenem lub z fosforem. Arsenek galu (GaAs) domieszkowany krzemem emituje promieniowanie podczerwone o długości fali około 0,97 μm. Fosforek galu (GaP) domieszkowany w obszarze *p* cynkiem i tlenem, a w obszarze *n* tellurem emituje promieniowanie czerwone. GaP domieszkowany w obszarze *p* cynkiem i azotem, a w obszarze *n* siarką i azotem emituje promieniowanie zielone. Diody z GaP mogą być wykonane również w taki sposób, że ich barwa świecenia zmienia się w funkcji natężenia przepływającego przez nie prądu. Jest to związane z tym, że mechanizmy rekombinacji promienistej w DEL czerwonych z GaP i w DEL zielonych z GaP są inne. Rekombinacja promienista warunkująca świecenie czerwone jest związana z przejściami między poziomami domieszkowymi donorów i akceptorów. Dlatego też, gdy rośnie prąd płynący przez złącze, następuje zjawisko nasycenia natężenia świecenia, wynikające ze skończonej liczby centrów rekombinacyjnych (par donor–akceptor). Rekombinacja warunkująca świecenie zielone związana jest z przejściami z płytkiego poziomu domieszkowego azotu do pasma walencyjnego. Ponieważ nie ma tu ograniczenia liczby centrów rekombinacyjnych nie obserwujemy też zjawiska nasycenia natężenia luminescencji wraz ze wzrostem prądu płynącego przez złącze. Ponadto czerwone świecenie występuje tylko po stronie *p* złącza *p-n* zaś zielone świecenie występuje po obu stronach złącza *p-n*. Gdy więc obszar *n* DEL z GaP jest domieszkowany siarką i azotem (świecenie zielone), zaś obszar *p* cynkiem i tlenem (świecenie czerwone), barwa świecenia diody zależy od natężenia płynącego przez nią prądu. Przy małym natężeniu prądu DEL świeci intensywnie światłem czerwonym a słabo światłem zielonym. Gdy natężenie prądu wzrasta, rośnie też natężenie światła zielonego, natomiast natężenie światła czerwonego nie zmienia się. W wyniku tego następuje zmiana barwy emitowanego światła — przez pomarańczową, żółtą aż do zielonej. Do produkcji DEL emitujących światło widzialne stosuje się też często związek mieszany $GaAs_{1-x}P_x$, z odpowiednimi domieszkami.

materiały do produkcji DEL

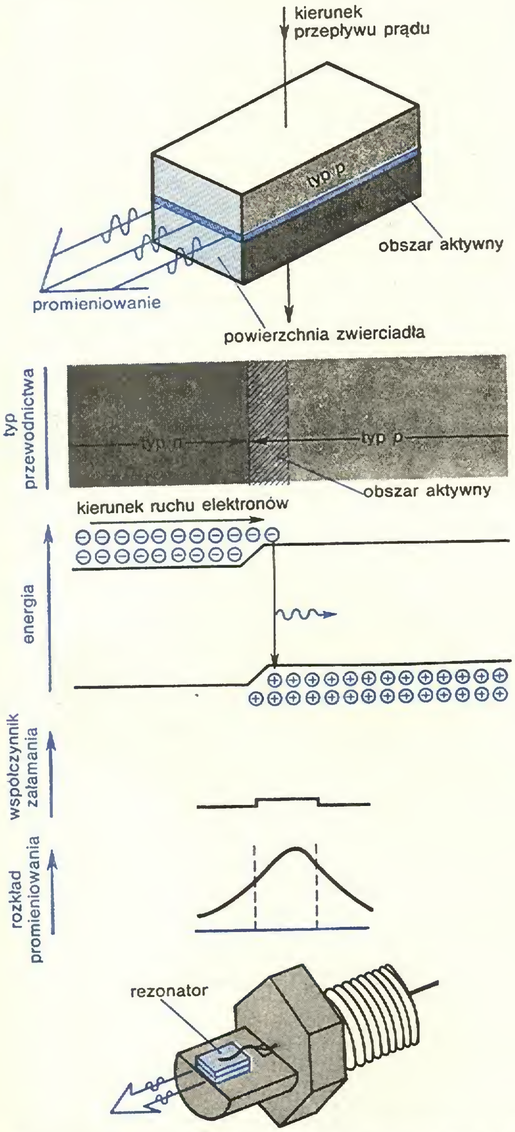
DEL są źródłami światła niespójnego. Półprzewodnik z wytworzonym w nim złączem *p-n* zdolnym emitować promieniowanie elektromagnetyczne, można ukształtować tak, by powstał rezonator optyczny, tzn. układ, w którym promieniowanie elektromagnetyczne tworzy fale stojące. Gdy tak ukształtowany półprzewodnik pobudzimy znacznie silniejszym prądem elektrycznym niż DEL, to może powstać w nim proces laserowy, tzn. proces generacji spójnego promieniowania elektromagnetycznego w wyniku wymuszonych przejść elektronów z pasma przewodnictwa do pasma walencyjnego (→ Lasery — podstawy działania). W najprostszym przypadku półprzewodnik ma kształt rezonatora Fabry'ego–Pérota tzn. jest prostopadłościanem, którego boki prostopadłe do płaszczyzny złącza *p-n* tworzą zwierciadła rezonatora. Promieniowanie wytworzone w tym rezonatorze może zostać rozproszone w półprzewodniku, nie dając wkładu do tworzącej się w nim fali stojącej. Może ono zostać zaabsorbowane, może być niedokładnie odbite na zwierciadłach rezonatora lub też może się wydostać z półprzewodnika jako promieniowanie niespójne. Powstające straty promieniowania stwarzają konieczność wstrzyknięcia przez złącze tak dużej liczby elektronów, by zagwarantowana została przewaga aktów wytwarzania promieniowania nad aktami jego rozpraszania. Zatem dopiero po przekroczeniu pewnej progowej wartości natężenia prądu płynącego przez złącze *p-n* diody laserowej emituje ona promieniowanie spójne i monochromatyczne. Najlepsze lasery złączowe pracujące w temperaturze pokojowej wymagają prądów o natężeniach rzędu kilkudziesięciu mA, co odpowiada progowej gęstości prądu rzędu kilkuset A/cm². Przy wartościach natężenia prądu mniejszych od wartości progowej laser jest zwykłą DEL. Emituje on wówczas promieniowanie niespójne o dość szerokiej linii widmowej i znacznie mniejszym natężeniu. Ten dwustanowy charakter pracy jest bardzo ważną cechą laserów złączowych wykorzystaną

dioda laserowa

w układach do przetwarzania danych. Wartość progowego natężenia prądu lasera złączowego lub, co znacznie częściej się podaje, wartość progowej gęstości prądu, jest istotnym parametrem eksploatacyjnym tego przyrządu. Przy zbyt wysokich wartościach prądów progowych lasery złączowe mogą bowiem pracować tylko impulsowo, gdyż występujące wówczas efekty cieplne uniemożliwiają pracę ciągłą (przy zasilaniu prądem stałym).

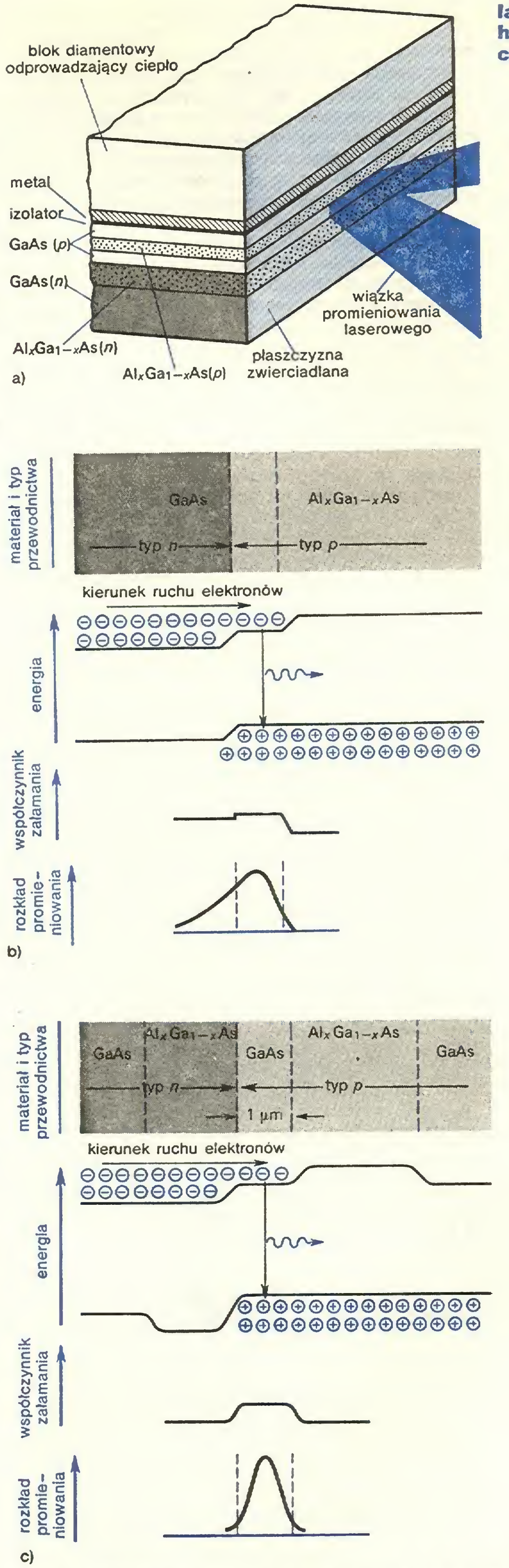
laser homozłączowy

Najprostszy a zarazem najbardziej do niedawna rozpowszechniony typ lasera złączowego, tzw. laser homozłączowy, ma postać przedstawioną schematycznie na rys. 3. Akcja laserowa występuje w cienkim obszarze (obszar aktywny) przylegającym do złącza *p-n*, wytworzonego w jednego rodzaju materiale półprzewodnikowym, a wiązka promieniowania laserowego jest skierowana prostopadle do pary doskonale gładkich płaszczyzn półprzewodnika odgrywających rolę zwierciadeł rezonatora Fabry'ego–Pérota. Lasery homozłączowe są najczęściej wykonane z GaAs, mają bardzo duże progowe gęstości prądów (rzędu 10^4–10^5 A/cm^2), wskutek czego w temperaturze pokojowej mogą pracować tylko impulsowo.



Rys. 3. Najprostszy laser złączowy (laser homozłączowy) z rezonatorem Fabry'ego–Pérota oraz jego struktura pasmowa, współczynnik załamania i rozkład promieniowania w obszarze aktywnym; na dole przykład konstrukcji z rezonatorem zmontowanym na sześciokątnym korpusie z gwintem

laser heterozłączowy



Rys. 4. Laser heterozłączowy: a) schemat budowy lasera z paskową geometrią kontaktu omowego, b) struktura pasmowa i rozkład promieniowania lasera monoheterozłączowego, c) struktura pasmowa i rozkład promieniowania lasera bihetorozłączowego

Dokładne badania przeprowadzone w ostatnich latach doprowadziły do odkrycia, że dostatecznie niskie wartości prądów progowych można uzyskać tylko w półprzewodnikach o złożonej strukturze składu chemicznego. Najlepsze rezultaty uzyskano w strukturach złożonych z warstwy GaAs oraz z dwóch warstw, w których część atomów Ga zastępują atomy Al, przez co powstaje związek trójskładnikowy $Al_xGa_{1-x}As$. Złącza powstałe na styku dwóch warstw o różnym składzie chemicznym nazywa się heterozłączami. Ponieważ przerwy energetyczne w GaAs i w $Al_xGa_{1-x}As$ mają różne wartości, to w heterozłączu utworzonym z tych materiałów powstają znacznie wyższe bariery energetyczne (potencjału) dla elektronów niż w złączu *p-n* wytworzonym w materiale jednego rodzaju, np. w GaAs. Powoduje to, że elektrony wstrzyknięte przez złącze *p-n* w GaAs do obszaru *p* tego materiału zostają odbite od bariery energetycznej występującej na heterozłączu GaAs–$Al_xGa_{1-x}As$ ograniczającym grubość obszaru *p* w GaAs (rys. 4). Rozkład przestrzenny elektronów w strukturze warstwowej lasera heterozłączowego zostaje więc ograniczony do obszaru znajdującego się między heterozłączami tej struktury. Różnice występujące we współczynnikach załamania między GaAs i $Al_xGa_{1-x}As$ powodują, że i światło odbija się od heterozłącza między tymi dwoma związkami półprzewodnikowymi. W strukturze warstwowej z dwoma heterozłączami zostaje więc ograniczony rozkład przestrzenny zarówno elektronów wstrzykniętych do obszaru *p* przez złącze jak i promieniowania elektromagnetycznego wytworzonego w wyniku wymuszonych przejść emisyjnych tych elektronów. Te dwa ograniczenia przestrzenne, równoznaczne ze zmniejszeniem strat optycznych w laserze, prowadzą do uzyskania wartości progowej gęstości prądu rzędu kilkuset A/cm^2, co umożliwia ciągłą pracę lasera heterozłączowego w temperaturze pokojowej.

heterozłącza

Złączowe elektroluminescencyjne źródła promieniowania elektromagnetycznego wyróżniają następujące właściwości: 1) duża sprawność przemiany energii elektrycznej w energię promienistą — w najlepszych laserach złączowych sięga ona 40%, w DEL świecących zielono skuteczność świetlna może osiągać granicę teoretycznie możliwą 680 lm/W; 2) czas pracy użytecznej (tzn. okres, po upływie którego natężenie świecenia maleje do połowy swej wartości początkowej) w najlepszych obecnie egzemplarzach laserów złączowych jest rzędu 10 000 godzin (pracy nieprzerwanej), w DEL zaś aż do ok. 20 lat (DEL są ponadto odporne na wstrząsy i warunki atmosferyczne); 3) mały pobór mocy ze źródła zasilania (kilkanaście lub kilkadziesiąt mW); 4) duża sprawność i łatwość modulacji emitowanego promieniowania z częstościami sięgającymi wartości 10^9 Hz — modulacja jest realizowana bezpośrednio przez zmianę prądu płynącego przez diodę; 5) małe wymiary — półprzewodnik z wytworzonym w nim złączem ma zwykle kształt prostopadłościanu o bokach 0,2 mm, 0,5 mm, 0,5 mm.

właściwości złącz elektroluminescencyjnych

Półprzewodnikowe detektory promieniowania elektromagnetycznego

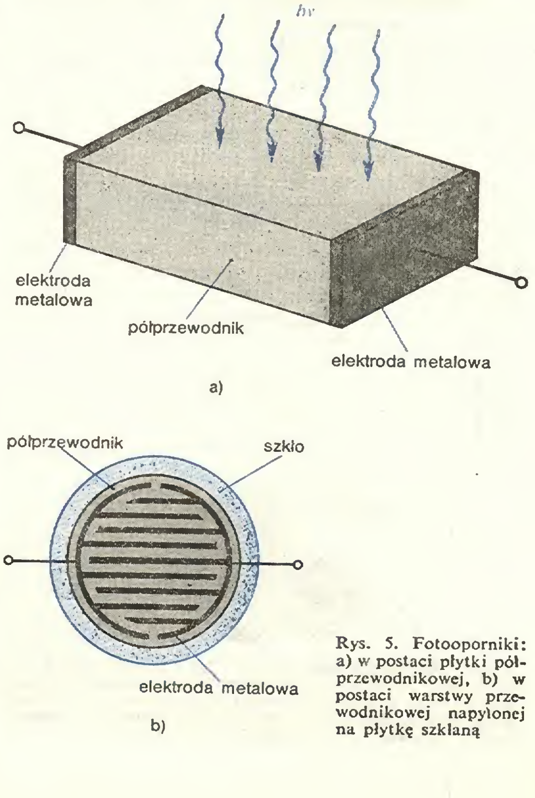
Promieniowanie elektromagnetyczne padające na jakąś substancję może zostać odbite, zaabsorbowane lub też może przejść przez tę substancję z pewnym niewielkim tylko osłabieniem jego natężenia. Detektory promieniowania elektromagnetycznego konstruuje się w taki sposób, by maksymalna część padającego na nie promieniowania została w nich zaabsorbowana. Ta zasada jest ogólna dla wszystkich rodzajów detektorów, a więc i dla detektorów półprzewodnikowych. Różne są zjawiska, które zachodzą w półprzewodniku pod wpływem padającego nań promieniowania elektromagnetycznego i różne są też rodzaje detektorów półprzewodnikowych. Gdy energia, którą przekazuje promieniowanie elektromagnetyczne elektronom półprzewodnika podczas procesu jonizacji atomów w krysztale, wystarcza na to, by elektron pokonał siły wiążące go wewnątrz półprzewodnika i opuścił napromieniony półprzewodnik, mamy do czynienia ze zjawiskiem fotoelektrycznym zewnętrznym. Zjawisko to jest wykorzystywane we wszelkiego rodzaju fotokomórkach, czy fotopowielaczach. Powstawanie pod wpływem napromienienia swobodnych nośników ładunku wewnątrz półprzewodnika nazywa się natomiast zjawiskiem fotoelektrycznym wewnętrznym. Różne są skutki zjawiska fotoelektrycznego wewnętrznego w półprzewodniku. Zależą one przy tym od struktury półprzewodnika, od obecności w nim określonych domieszek, jak też od występowania w półprzewodniku wewnętrznych pól elektrycznych. Zjawisko fotoelektryczne wewnętrzne wywołuje obniżenie oporu elektrycznego jednorodnego półprzewodnika w postaci płytki lub warstwy napylonej na jakiś inny materiał. W półprzewodniku z wytworzonym złączem *p-n*, nie spolaryzowanym zewnętrznym napięciem elektrycznym, zjawisko to wywołuje pojawienie się na złączu stałego napięcia fotoelektrycznego (efekt fotowoltaiczny). Gdy natomiast złącze *p-n* zostanie spolaryzowane wstępnie w kierunku zaporowym, to zjawisko fotoelektryczne wewnętrzne powoduje zmianę nieliniowego oporu złącza. Fotodetektory półprzewodnikowe, w których wykorzystano opisane skutki wewnętrznego zjawiska fotoelektrycznego noszą odpowiednio nazwy: fotoopornik (fotorezystor), fotoogniwo (bateria słoneczna) oraz fotodioda. W takich fotodetektorach zawsze występują następujące procesy fizyczne: wytworzenie nośników ładunku przez promieniowanie elektromagnetyczne padające na półprzewodnik, przeniesienie (transport) tych nośników przez obszar półprzewodnika do kontaktów metalowych łączących ten półprzewodnik z zewnętrznym obwodem elektrycznym, oraz oddziaływanie fotoprądu dopływającego do kontaktów z zewnętrznym obwodem elektrycznym. Charakter wymienionych procesów jest zwykle różny w różnych typach fotodetektorów półprzewodnikowych. Warunkuje on parametry eksploatacyjne tych fotodetektorów oraz określa możliwości ich zastosowań.

zjawiska wywołane przez promieniowanie

fotodetektory półprzewodnikowe

Najprostszym typem fotodetektora jest fotoopornik. Jest to przyrząd złożony z płytki półprzewodni-

fotoopornik



Rys. 5. Fotooporniki: a) w postaci płytki półprzewodnikowej, b) w postaci warstwy przewodnikowej napylonej na płytkę szklaną

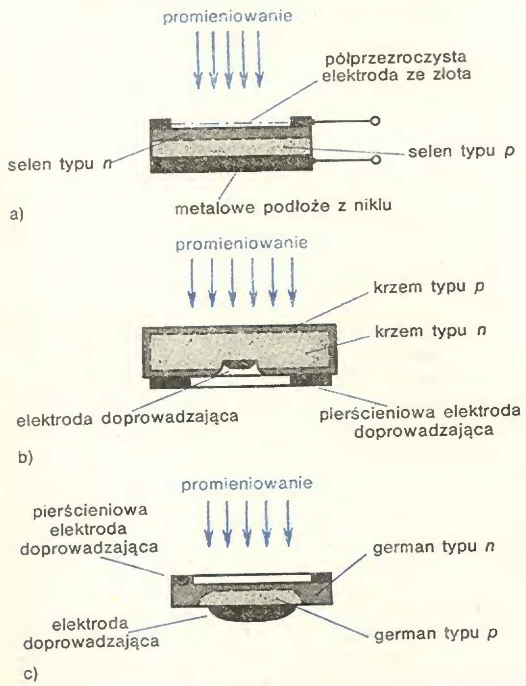
kowej lub z warstwy półprzewodnikowej napylonej na płytkę szklaną. Na przeciwległych końcach tej warstwy lub płytki wykonane są liniowe (nieprostujące) kontakty metalowe (rys. 5). W czasie padania na powierzchnię fotoopornika promieniowania elektromagnetycznego zmniejsza się jego opór elektryczny. Pochłonięte promieniowanie wytwarza w nim nośniki ładunku albo w wyniku przejść międzypasmowych, albo w wyniku przejść z udziałem poziomów energetycznych domieszkowych leżących w przerwie energetycznej półprzewodnika. Fotoprzewodnictwo w tego typu fotodetektorach może się pojawić jednak tylko wtedy, gdy energia pochłoniętych fotonów jest równa lub większa od energii oddzielającej poziom domieszkowy leżący w przerwie energetycznej od krawędzi pasma przewodnictwa lub pasma walencyjnego, albo też jest co najmniej równa przerwie energetycznej półprzewodnika. Ten fakt określa tzw. długofalową granicę pracy fotoopornika.

fotoogniwo

Fotoogniwo jest przyrządem o stosunkowo dużej powierzchni oświetlonej. Złącze *p-n* znajduje się w bezpośrednim sąsiedztwie (na głębokości rzędu 1 μm) oświetlanej powierzchni. Padające na złącze fotony o energii większej od szerokości przerwy energetycznej półprzewodnika powodują powstanie, w miejscu gdzie są pochłaniane, par elektron–dziura. Pole elektryczne wewnątrz półprzewodnika, związane z obecnością złącza *p-n*, przesuwa nośniki różnych rodzajów w różne strony. Elektrony trafiają do obszaru *n*, dziury zaś do obszaru *p*. Rozdzielenie nośników ładunku w złączu powoduje powstanie na nim zewnętrznego napięcia elektrycznego. Ponieważ rozdzielone nośniki są nośnikami nadmiarowymi (mają nieskończony czas życia), a napięcie na złączu *p-n* jest stałe, oświetlone złącze działa jako ogniwo elektryczne.

fotodioda

Fotodioda jest również elementem półprzewodnikowym ze złączem *p-n*. Teraz jednak złącze jest spolaryzowane wstępnie w kierunku zaporowym. Wskutek oświetlenia prąd zaporowy złącza jest modulowany przez powstające w złączu pary elektron–dziura, co



Rys. 6. Fotodiody: a) selenowa, b) krzemowa, c) germanowa

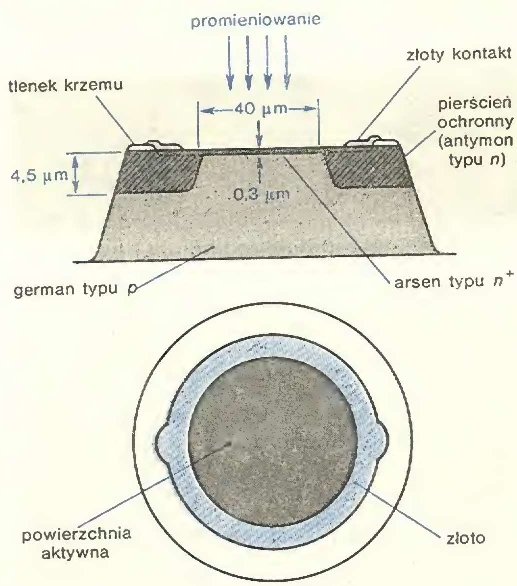
powoduje odpowiednie zmiany nieliniowego oporu tego złącza. Konstrukcje kilku rodzajów fotodiod przedstawiono na rys. 6. Rodzina fotodiod jest bardzo duża. Zaliczają się do niej diody typu *p-i-n*, diody z barierą Schottky'ego, diody z heterozłączami, diody z punktowymi kontaktami metalicznymi oraz diody lawinowe. Wszystkie te fotodiody mają duże i różnorodne zastosowania praktyczne. Jednak fotodiody lawinowe, które odznaczają się silnym wewnętrznym wzmocnieniem fotoprądu mają przed sobą, jak się wydaje, największe perspektywy zastosowań, zwłaszcza w układach komunikacji optycznej.

fotodioda lawinowa

Fotodioda lawinowa pracuje przy dużych napięciach zaporowych, przy których oświetlenie złącza *p-n* powoduje przebicie lawinowe. Jeżeli wytworzony przez pochłonięty foton nośnik mniejszościowy, np. elektron, zostanie przyspieszony w polu elektrycznym spolaryzowanego zaporowo złącza *p-n* do energii kinetycznej większej ok. 1,5 raza od energii równej przerwie energetycznej półprzewodnika, to może on w wyniku zderzenia niesprężystego przekazać część swej energii elektronowi z pasma walencyjnego powodując jego przejście do pasma przewodnictwa. W ten sposób następuje generacja pary elektron–dziura, kosztem pewnego obniżenia energii elektronu pierwotnego, który jednak nadal pozostaje w pasmie przewodnictwa. Powstaje układ dwóch elektronów przewodnictwa i jednej dziury, które to nośniki ponownie mogą nabywać energię kinetyczną w polu złącza *p-n*. Po osiągnięciu energii większej od przerwy energetycznej, każdy z tych ładunków może wytworzyć następną parę elektron–dziura. Proces ten powtarza się wielokrotnie w sposób lawinowy, dając na kontaktach metalicznych takiej diody silny sygnał elektryczny. Przebicie lawinowe bywa często zlokalizowane w kilku obszarach, w których tworzą się tzw. mikroplazmy.

uzyskiwanie dużego wzmocnienia

Aby uzyskać w diodzie lawinowej duże wzmocnienie fotoprądu, konieczne jest takie wykonanie złącza *p-n*, aby przebicie to zachodziło jednocześnie w całym obszarze złącza oraz by efekty brzegowe w miejscu wyjścia złącza na powierzchnię półprzewodnika nie spowodowały prądów upływowych. Wymaga to dobrania materiału o małej ilości wtrąceń i dużej doskonałości sieci krystalicznej (małej ilości dyslokacji) oraz zastosowania pierścienia ochronnego (rys. 7).



Rys. 7. Fotodioda lawinowa

Zastosowania półprzewodnikowych źródeł i detektorów promieniowania elektromagnetycznego

Półprzewodnikowe źródła i detektory promieniowania elektromagnetycznego znajdują różnorodne zastosowania. DEL emitujące światło widzialne stosuje

lampki sygnalizacyjne

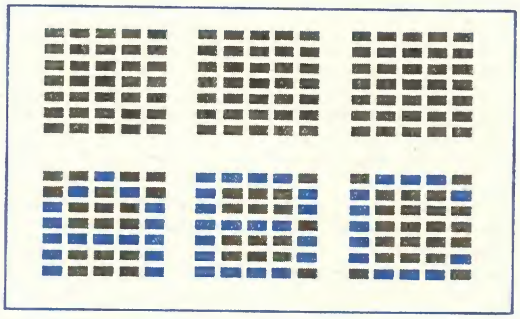
się jako miniaturowe lampki sygnalizacyjne małej mocy o barwach świecenia czerwonej, żółtej czy zielonej w różnego rodzaju aparaturze elektronicznej, zwłaszcza przenośnej. DEL mogą wskazywać, że zostało włączone napięcie w aparaturze, mogą podświetlać przezroczyste klawisze z określonymi napisami, np. w aparatach telefonicznych, mogą być oświetlaczami przyrządów wskazujących poziom paliwa w baku samochodowym czy szybkość samochodu, mogą też być źródłem promieniowania w przenośnych dalmierzach optycznych. DEL są również stosowane we wskaźnikach alfanumerycznych. Diody te, połączone w jedną mozaikę świecącą i sterowane przez sygnały elektryczne trafiają na specjalny układ kodujący, pozwalają wyświetlać dowolną cyfrę od zera do dziewięciu lub dowolną literę alfabetu. Układy takich wskaźników mogą tworzyć albo swoistą tablicę świetlną, albo układ wielocyfrowy stosowany np. w woltomierzach cyfrowych, w kalkulatorach elektronicznych czy w zegarkach elektronicznych. Pojedynczą mozaikę złożoną z 35 DEL przedstawia rys. 8, natomiast całą tablicę świetlną rys. 9.

wskaźniki alfanumeryczne

DEL emitujące promieniowanie podczerwone słabo tłumione przez atmosferę (długość fali $\lambda \approx 0,9$ μm) są stosowane jako źródła promieniowania podczerwonego, m.in. w dalmierzach optycznych, w przenośnych telefonach optycznych krótkiego zasięgu (bez połączenia kablowego) czy też w czytnikach taśmy perforowanej w maszynach matematycznych (rys. 10). Możliwość szerokiego zastosowania tych źródeł promieniowania wiąże się również z przyrządami zwanymi transoptorami lub optronami. Są to umieszczone w jednej obudowie i izolowane elektrycznie od siebie dwa elementy układu optoelektronicznego — źródło i detektor promieniowania elektromagnetycznego (rys. 11). W urządzeniach tych źródłem jest DEL (zwykle wykonana z GaAs) emitująca promieniowanie podczerwone, a detektorem — fotodioda (wykonana z krzemu), której czułość widmowa jest prawie maksymalna dla emitowanej fali. Transoptor jest podstawowym elementem strukturalnym układów optoelektronicznych mogącym realizować różne funkcje obwodowe, takie jak np. wzmacnianie, generacja, przełączanie czy formowanie sygnałów. Może być ponadto używany jako transformator prądu stałego lub jako przekaźnik. Transoptory są szczególnie przydatne tam, gdzie występuje styk dwóch różnych sieci elektrycznych, np. wysokiego i niskiego napięcia, wymagających zastosowania galwanicznej separacji między nimi. Są więc stosowane w układach kontroli

czytniki taśmy perforowanej

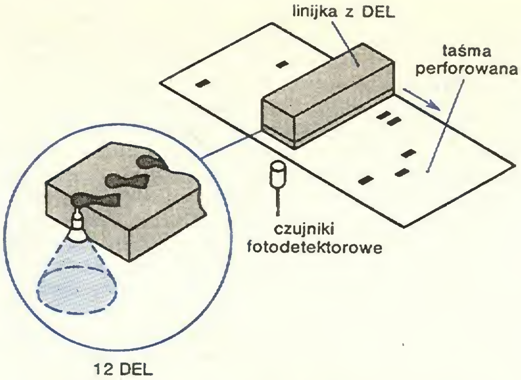
transoptory



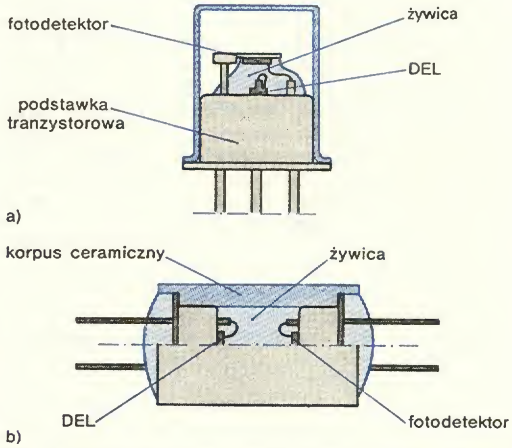
Rys. 8. Wskaźnik literowy złożony z trzech 35-elementowych mozaik DEL oraz litery wyświetlane przez ten wskaźnik



Rys. 9. Tablica świetlna wykonana z DEL



Rys. 10. Schemat czytnika taśmy perforowanej. Taśma perforowana przechodząca pomiędzy linijką wykonaną z 12 DEL a układem fotodetektorów, powoduje pojawienie się sygnału elektrycznego tylko z tej fotodiody, która znajduje się bezpośrednio pod otworem w taśmie (tylko wówczas światło z DEL dochodzi do fotodiody). Sygnały elektryczne odbierane z układu fotodiod odtwarzają rozmieszczenie otworów na taśmie



Rys. 11. Transoptory: a) na podstawce tranzystorowej, b) w obudowie ceramicznej

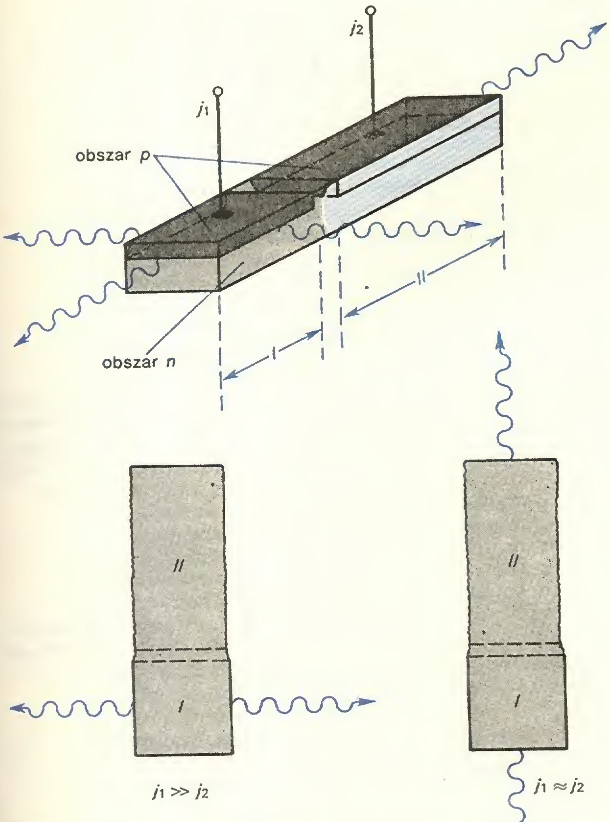
wysokiego napięcia, w aparaturze medycznej, która musi być izolowana od pacjenta, w układach zabezpieczających maszyny górnicze przed iskrzeniem czy też urządzeniach zabezpieczających przed przeciążeniami prądowymi.

zastosowania laserów złączowych

Lasery złączowe nie mają, jak do tej pory, tak masowego zastosowania jak DEL. Zarysowują się jednak już na obecnym etapie rozwoju trzy główne kierunki zastosowania tych przyrządów. Dwustanowy charakter pracy umożliwia ich zastosowanie w układach do przetwarzania danych, wykonujących określone operacje logiczne. Przełączenie ze stanu pracy „off" (DEL) do stanu pracy „on" (laser) może się odbywać nie tylko za pomocą elektrycznych sygnałów. Okazuje się, że można też laser przełączyć z jednego stanu w drugi za pomocą sygnału świetlnego, emitowanego z drugiego lasera złączowego. Fakt ten umożliwił skonstruowanie laserów sprzężonych optycznie. Lasery sprzężone optycznie są wykonywane zwykle w ten sposób, że w jednej monokrystalicznej płytce GaAs wytwarza się dwa laserujące złącza *p-n*. Aby zwiększyć sprzężenie optyczne między laserami, oba złącza umieszcza się w tej samej płaszczyźnie, a ponadto tylko strony *p* złącz są od siebie oddzielone elektrycznie (przez odpowiednie wdyfundowanie domieszek akceptorowych do płytki GaAs). Oba złącza mają różną długość, zaś ścianki boczne płytki, prostopadłe do osi optycznej układu, są dodatkowo zmatowione w obszarze złącza dłuższego. W obszarze złącza krótszego ścianki boczne mają gładkość lustrzaną. Zapobiega to możliwości

lasery sprzężone optycznie

generacji światła w kierunku prostopadłym do osi układu w laserze dłuższym. Gdy gęstość prądu przepływającego przez laser krótszy (obszar *I* na rys. 12) jest znacznie większa od gęstości prądu płynącego przez laser dłuższy ($j_1 \gg j_2$), to akcja laserowa występuje tylko w laserze krótszym i to w kierunku poprzecznym. Dzieje się tak dlatego, że w laserze dłuższym, pracującym wówczas poniżej progu laserowego, promieniowanie wychodzące z lasera krótszego w kierunku równoległym do osi układu jest silnie absorbowane. Jeżeli jednak gęstość prądu w obszarze złącza dłuższego (obszar *II*) wzrasta do wartości progowej akcji laserowej w tym złączu ($j_1 \approx j_2$), to promieniowanie tego złącza spowoduje zmniejszenie inwersji obsadzeń w obszarze złącza krótszego, a w konsekwencji wygaszenie procesu laserowego w kierunku poprzecznym do osi optycznej układu. Układ będzie teraz emitował promieniowanie jedynie w kierunku równoległym do osi układu. Jak widać, zmieniając gęstość prądów zasilających każde ze złączy *p-n* układu, można przełączać promieniowanie laserowe na wyjściu z kierunku prostopadłego do osi układu na kierunek równoległy do tej osi.

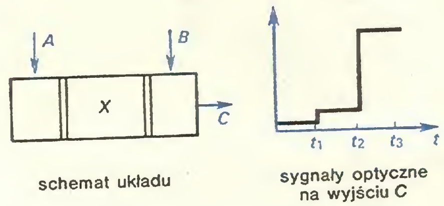


Rys. 12. Układ przełączający zbudowany na laserach złączowych sprzężonych optycznie

układy logiczne

Na tej zasadzie można konstruować układy realizujące różne operacje logiczne, np. operację „i" oraz operację „lub". Operację „i" (rys. 13a) realizuje laser z trzema elektrycznie odizolowanymi częściami znajdującymi się w jednym wspólnym rezonatorze. Stały prąd trafia na część środkową *X*, co powoduje promieniowanie układu jako DEL. Dwie pozostałe części układu spełniają rolę elementów pochłaniających promieniowanie w rezonatorze — nie doprowadza się bowiem do nich prądu. Gdy na układ trafiają sygnały optyczne *A* i *B*, to wskutek działania ich energii promienistej na złącza *p-n* absorbentów maleje współczynnik absorpcji w tych obszarach, w wyniku czego w rezonatorze wzbudza się akcja laserowa, a na wyjściu pojawia się silny laserowy sygnał optyczny *C*. Poziom zasilania części środkowej jest tak dobrany, by sygnał laserowy na wyjściu *C* pojawił się tylko wtedy, gdy sygnały *A* i *B* trafiają na układ jednocześnie. Przy jednym sygnale *A* lub *B* układ pracuje jako DEL. W podobny sposób jest realizowana w laserze z dwiema odizolowanymi elektrycznie częściami operacja „lub" (rys. 13b).

operacja „i"

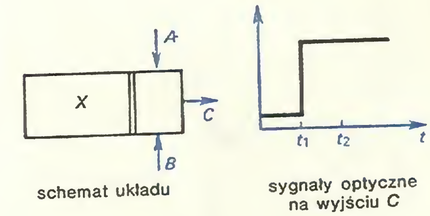


	$t<t_1$	$t_1<t<t_2$		$t>t_2$
A	0	0	1	1
B	0	1	0	1
C	0	0	0	1

schemat układu — sygnały optyczne na wyjściu C — rozkład czasowy sygnałów w układzie

a)

operacja „lub"



	$t<t_1$	$t>t_1$		
A	0	1	0	1
B	0	0	1	1
C	0	1	1	1

schemat układu — sygnały optyczne na wyjściu C — rozkład czasowy sygnałów w układzie

b)

Rys. 13. Przykłady laserowych układów logicznych

telekomunikacja optyczna

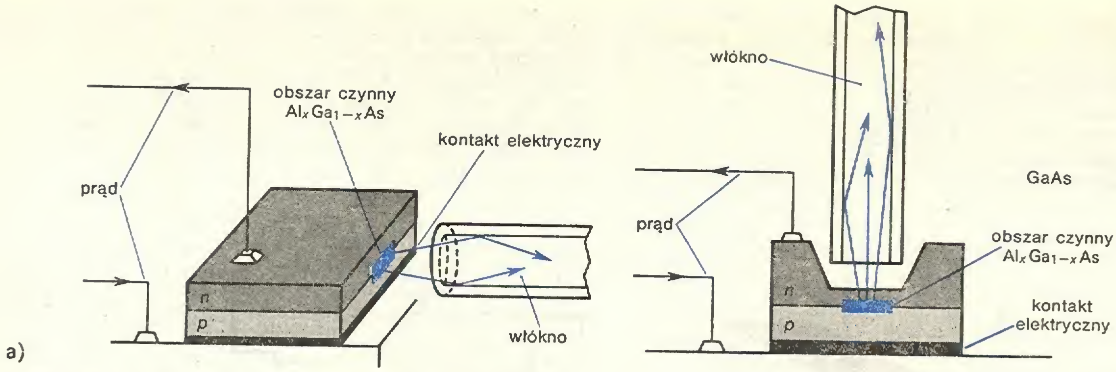
Małe wymiary i mała masa oraz dostatecznie duża moc emitowanego promieniowania (które z łatwością można modulować), umożliwiają zastosowanie laserów złączowych połączonych ze światłowodem, np. z włóknem szklanym lub ze światłowodem z tworzywa sztucznego (rys. 14), zarówno jako nadajnika sygnałów optycznych jak i stacji przekaźnikowej (wzmacniającej i retransmitującej sygnały optyczne) do celów telekomunikacji optycznej. Perspektywa tego zastosowania laserów złączowych może bardzo poszerzyć możliwości zastosowania tych przyrządów, rewolucjonizując jednocześnie telekomunikację.

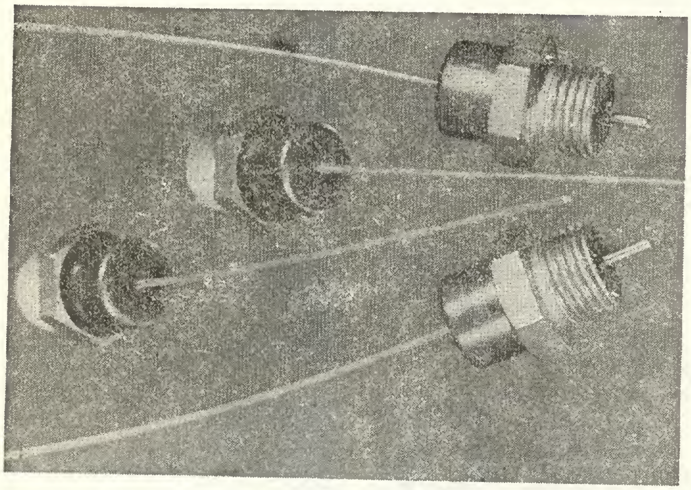
radar, zapalniki pocisków, holografia

Trzecim wreszcie kierunkiem zastosowania laserów złączowych jest zastosowanie o bardziej specjalnym przeznaczeniu. Lasery te mogą być np. stosowane jako głowice lokacyjne w radarach optycznych umieszczonych w samolotach odrzutowych, mogą też być zapalnikami pocisków balistycznych, mogą wreszcie służyć jako miniaturowe czytniki hologramów i stanowić elementy holograficznych układów pamięciowych w maszynach matematycznych.

zastosowania półprzewodnikowych detektorów

Jeszcze szersze zastosowanie niż źródła promieniowania elektromagnetycznego mają półprzewodnikowe detektory tego promieniowania. Są one obecnie stosowane niemal w każdej dziedzinie techniki. W regulacji automatycznej stosuje się je jako różnego rodzaju receptory selektywne, reagujące tylko na promieniowanie o określonej długości fali, oraz nieselektywne — reagujące na promieniowanie o różnych długościach fal. W technice pomiarowej znajdują zastosowanie jako mierniki natężenia sygnałów optycznych, natomiast w kosmonautyce wykorzystuje się je jako baterie słoneczne, tzn. jako przetworniki energii promieniowania słonecznego w energię elektryczną zasilającą aparaturę znajdującą się w pojeździe kosmicznym. Wieloelementowe mozaiki zawierające od kilku do kilkunastu tysięcy światłoczułych elementów są stosowane jako przetworniki obrazów w kamerach telewizyjnych. Mogą one też być zastosowane do przetwarzania obrazu optycznego na bodźce czuciowe, stając się wówczas częścią składową urządzeń zastępujących niewidomym wzrok. Szerokie zastoso-





b)

Rys. 14. a) Przykład sprzężenia światła emitowanego przez laser złączowy (z lewej) i DEL (z prawej) z włóknem szklanym, b) lasery heterozłączowe połączone ze światłowodem wykonanym z włókna z tworzywa sztucznego (średnica światłowodu ok. 2 mm)

wanie fotodetektorów (głównie krzemowych) wiąże się również ze wspomnianymi już wyżej transoptorami czy czytnikami taśmy perforowanej, w których współpracują one ściśle z półprzewodnikowymi źródłami promieniowania elektromagnetycznego.

noktowizja termografia

Ważne zastosowanie znalazły też półprzewodnikowe detektory promieniowania podczerwonego o długości fali równej kilka lub kilkanaście µm. Są one stosowane w noktowizji (widzenie w ciemności na podstawie rejestracji promieniowania podczerwonego emitowanego przez różne ciała lub obiekty i przetworzenia go na promieniowanie widzialne) oraz w termografii (rejestracja rozkładu temperatury na powierzchni rozgrzanego obiektu na podstawie rejestracji natężenia promieniowania podczerwonego emitowanego przez elementy powierzchni tego obiektu).

Zakres zastosowań półprzewodnikowych elementów optoelektronicznych jest ogromny i przytoczone powyżej przykłady nie wyczerpują wszystkich możliwości.

Kierunki rozwoju optoelektroniki półprzewodnikowej

Rozwój optoelektroniki półprzewodnikowej jest stymulowany z jednej strony nowymi zastosowaniami przyrządów i układów optoelektronicznych w różnych dziedzinach techniki, z drugiej zaś strony rozwojem innych gałęzi elektroniki, np. automatyki, informatyki czy telekomunikacji, które wykorzystują przyrządy optoelektroniczne i narzucają odpowiednie wymagania w odniesieniu do niezawodności, sprawności energetycznej, miniaturyzacji czy funkcjonalności tych urządzeń. Rozwój mikroelektroniki (→ Mikroelektronika) spowodował np. konieczność opracowania zminiaturyzowanego i niezawodnego źródła światła, by wskaźniki świetlne stosowane w układach mikroelektronicznych były porównywalne swymi wymiarami i niezawodnością z tymi układami. Doprowadziło to do konstrukcji pierwszych DEL.

zwiększenie zakresu widmowego

Różne są tendencje rozwojowe poszczególnych rodzajów przyrządów optoelektronicznych. W zakresie półprzewodnikowych źródeł promieniowania elektromagnetycznego można przewidywać rozwój w kierunku rozszerzenia zakresu widmowego emitowanych fal i to zarówno w stronę ultrafioletu, jak i w stronę podczerwieni (bardzo dalekiej), w kierunku zwiększenia mocy emitowanego promieniowania oraz zwiększenia sprawności energetycznej procesu emisji promieniowania. Innym kierunkiem rozwoju źródeł jest tworzenie dużych matryc świecących, zmieniających od punktu do punktu swą barwę i natężenie pod wpływem sterujących sygnałów elektrycznych. Matryce takie mogą znaleźć zastosowanie jako duże, płaskie ekrany telewizji kolorowej, zawieszane na ścianie.

matryce świecące

rozpoznawanie druku, obrazów barwnych itp.

Jeśli chodzi o półprzewodnikowe detektory promieniowania elektromagnetycznego, to dąży się zarówno do znacznego zwiększenia ich czułości energetycznej i widmowej, jak też do tworzenia matryc i mozaik o bardzo dużym upakowaniu elementów detekcyjnych, co umożliwi skonstruowanie urządzeń do rozpoznawania druku, obrazów barwnych i przedmiotów. Ważnym kierunkiem rozwoju detektorów będzie też prawdopodobnie trójwymiarowy zapis informacji optycznej (w całej objętości kryształu) jak i przetwarzanie energii promieniowania słonecznego w energię elektryczną za pomocą dużych powierzchni światłoczułych.

układy pamięciowe

Duże nadzieje wiąże się z optoelektroniką półprzewodnikową w zakresie konstrukcji układów pamięciowych maszyn matematycznych przewidując możliwość budowy systemów o bardzo szybkim dostępie i pojemności większej niż 10^{12} bitów na cm^2. Wynika to z jednej strony z faktu, że teoretyczna granica gęstości upakowania bitów informacji jest ustalona przez długość fali, na jakiej jest ta informacja zapisywana, z drugiej zaś strony przez możliwość wykorzystania zalet techniki holograficznej.

monolityczna optyka zintegrowana

Innym kierunkiem rozwoju optoelektroniki półprzewodnikowej jest tzw. monolityczna optyka zintegrowana (scalona). Jest to dziedzina zajmująca się mikrominiaturyzacją układów optoelektronicznych na zasadzie analogicznej jak w obwodach scalonych. Dąży się tu do skonstruowania całego układu optoelektronicznego, tj. źródła, elementu przetwarzającego lub przesyłającego oraz detektora w postaci jednego monolitycznego elementu krystalicznego o wymiarach rzędu ułamka milimetra. Rozwój tej dziedziny może doprowadzić do nowych, trudnych jeszcze do przewidzenia zastosowań układów optoelektronicznych.

J. I. Pankove *Zjawiska optyczne w półprzewodnikach*, Warszawa 1974; C. H. Gooch *Przyrządy elektroluminescencyjne ze złączem p-n*, Warszawa 1977.

Mikroelektronika

Andrzej Zawadzki

W ciągu zaledwie kilkunastu lat prawie we wszystkich dziedzinach elektroniki zastąpiono lampy elektronowe elementami półprzewodnikowymi: diodami, tranzystorami, tyrystorami (tzw. dyskretnymi elementami półprzewodnikowymi). Mają one znacznie mniejsze rozmiary, większą niezawodność, zużywają mniej energii oraz są tańsze dzięki zastosowaniu w dużym stopniu zautomatyzowanych metod produkcji.

Początkowo stosowano proste, pojedyncze tranzystory i diody. Następnie opracowano elementy półprzewodnikowe spełniające inne funkcje (jak np. tyrystory), a tranzystory i diody wyspecjalizowano w najróżnorodniejsze typy — mocy, wielkiej częstości, przełączające, mikrofalowe itp. Szybko jednak osiągnięty został kres miniaturyzacji i niezawodności układów elektronicznych zbudowanych z dyskretnych elementów półprzewodnikowych, bowiem muszą one mieć wymiary umożliwiające montaż, a duża liczba połączeń określa pewien poziom niezawodności tych układów, który trudno zwiększyć. Na przykład, wg danych statystycznych, przy 1000 połączeń przynajmniej jedno zawodzi raz na 5 godzin pracy układu.

miniaturyzacja

Tymczasem rozwój pewnych dziedzin elektroniki uzależniony był od dalszego radykalnego postępu w dziedzinie miniaturyzacji, zmniejszenia zużycia energii oraz zwiększenia niezawodności układów elektronicznych. Takimi dziedzinami były przede wszystkim informatyka, telekomunikacja satelitarna i automatyka.

Potrzeba opracowania nowej koncepcji konstrukcji układów elektronicznych pojawiła się w sytuacji, gdy istniały już możliwości techniczne realizacji takiej koncepcji. Mianowicie doprowadzona była do perfekcji technologia planarna, służąca do wytwarzania większości tranzystorów, diod i tyrystorów półprzewodnikowych. Technologia ta polega na wytwarzaniu złącz *p-n* lub *n-p* w płytce monokrystalicznej przez kolejne stosowanie takich procesów jak dyfuzja i epitaksja. Na płytce o średnicy ok. 5 cm, wyciętej z monokryształu krzemu, wytwarza się jednocześnie (grupowo) wiele setek lub nawet tysięcy np. tranzystorów o praktycznie identycznych właściwościach (il. 101, tabl. 25).

technologia planarna

Technologia planarna jest bardzo szybka i ekonomiczna. Doskonaląc ją osiągano coraz lepsze parametry wytwarzanych elementów i coraz większą ich różnorodność. Dalszym krokiem było przejście do wytwarzania na ciągłym podłożu zespołów elementów elektronicznych wraz z połączeniami tak, aby tworzyły gotowe miniaturowe układy — tzw. układy scalone. Nie trzeba wówczas zaopatrywać poszczególnych elementów w wyprowadzenia, a następnie łączyć wyprowadzeń ze sobą. Odpowiednio zaprojektowane układy można wytwarzać grupowo — jak pojedyncze elementy. Tak narodziła się mikroelektronika, czyli technika projektowania i wytwarzania układów scalonych. Obecnie produkowane są układy scalone zawierające na kawałku krzemu o wymiarach rzędu 5×5 mm obwody elektroniczne zawierające kilkadziesiąt tysięcy tranzystorów, diod i oporników.

układ scalony

W skład układu scalonego wchodzą elementy czynne, czyli tranzystory, oraz bierne, jak diody, oporniki i kondensatory, które nie mają właściwości przekształcania sygnałów elektrycznych. Elementy czynne, a więc tranzystory, są w układzie scalonym (tak jak i w każdym układzie elektronicznym) wzmacniaczami prądu lub napięcia sygnałów elektrycznych, wzmacniaczami mocy, detektorami, dyskryminatorami czy też generatorami sygnałów o określonym kształcie i częstości. Elementy czynne mają wiele parametrów opisujących ich właściwości, jak np. współczynniki wzmocnienia, przebieg zależności prądu płynącego przez ten element od przyłożonego napięcia (zależności te mają charakter nieliniowy), czy też częstości, z jaką dany element może pracować itp. W odróżnieniu od elementów czynnych, elementy bierne są w zasadzie opisywane przez jedną cechę — jeden parametr (np. oporniki — przez wartość oporu elektrycznego, kondensatory — przez wartość pojemności).

elementy czynne i bierne

skale integracji układów

Stopień złożoności układów scalonych określany jest mianem skali integracji. Pierwsze układy scalone wyprodukowane w USA na początku lat sześćdziesiątych spełniały proste funkcje i zawierały kilka lub kilkanaście tranzystorów i oporników. Były to układy małej skali integracji — SSI (ang. *Small Scale Integration*).

Następnym krokiem w historycznym rozwoju techniki układów scalonych były układy wykonujące kilka funkcji, jak np. wzmacniacz pośredniej częstotliwości i detektor FM (modulacji częstotliwości) mające zastosowanie w odbiornikach radiowych, czy np. układy cyfrowe stosowane w technice maszyn i urządzeń cyfrowych. Układy takie, zawierające powyżej 100 tranzystorów oraz podobną liczbę oporników i diod, są nazywane układami średniej skali integracji MSI (ang. *Medium Scale Integration*).

W ostatnich latach rozpoczęto wytwarzanie układów wielkiej skali integracji (ang. *Large Scale Integration*), to znaczy scalonych układów spełniających określone funkcje. Są to układy zawierające wiele tysięcy tranzystorów i elementów biernych. Produkuje się w ten sposób np. pamięci o pojemności od 64 do 16 384 bitów, a firma IBM wyposaża swoje komputery w pamięci centralne, zbudowane z układów scalonych o pojemności ponad 60 000 bitów. Trzeba przy tym pamiętać, że jesteśmy dopiero na początku tej drogi, pamięci półprzewodnikowe produkuje się zaledwie od kilku lat i wielu ekspertów twierdzi, że w niedługiej przyszłości stosowane będą wyłącznie pamięci półprzewodnikowe.

pamięci półprzewodnikowe

Innymi układami wielkiej i bardzo wielkiej skali integracji są układy stosowane np. w kieszonkowych i stołowych kalkulatorach, minikomputerach, systemach telefonicznych i telekomunikacyjnych, systemach informatycznych i automatyki. Najnowszym osiągnięciem techniki cyfrowych układów scalonych wielkiej skali integracji są mikroprocesory, to znaczy układy scalone o możliwościach obliczeniowych małych maszyn matematycznych. Wiele firm, w szczególności amerykańskich, jak Motorola, Intel, Fairchild, Texas Instruments, produkuje zestawy mikrokomputerów składające się z kilku układów scalonych (mikroprocesorów, pamięci, układów wejściowych i wyjściowych), do których dołącza instrukcje programowania. Użytkownik według tych instrukcji może tak zaprogramować zestaw, aby spełniał on samodzielne zadania obliczeniowo-sterujące, jak sterowanie ruchem ulicznym, obliczanie ceny sprzedawanego towaru i wydawanie reszty, sterowanie procesami technologicznymi itp.

mikroprocesory

Projektowanie układów scalonych

Projektowanie układów wymaga zastosowania komputerów, gdyż inaczej czas zaprojektowania układu scalonego byłby niewyobrażalnie długi lub wręcz byłoby to niemożliwe do wykonania. Oczywiście właściwym projektantem układu scalonego jest człowiek — konstruktor. Opracowano wiele różnych programów obliczeniowych, które pozwalają ocenić, czy zaprojektowany układ będzie spełniał swoją funkcję zgodnie z założonymi parametrami, uwzględniając rozrzut właściwości układu wynikający z warunków produkcji, wpływu temperatury i np. warunków pracy układu. Inne programy pozwalają zaprojektować roz-

mieszczenie elementów w układzie scalonym tak, aby zajmowały one jak najmniejszą powierzchnię, dawały się połączyć jedną warstwą metalizacji (chociaż znane są sposoby wykonywania połączeń skrzyżowanych) i aby połączenia pomiędzy elementami układu były jak najkrótsze. Przy projektowaniu układów scalonych wykorzystuje się następujące zalety technologii tych układów:

zalety i ograniczenia technologii układów scalonych

— połaczenia o bardzo dużej niezawodności.

— taka sama zależność parametrów od warunków pracy elementów,

— możliwość dobierania elementów o takich właściwościach i parametrach, jakie są potrzebne.

Istnieją także ograniczenia, które konstruktor musi uwzględnić w swoim projekcie, a mianowicie:

— ograniczenie maksymalnych wartości oporów i pojemności,

— trudności w uzyskaniu tranzystorów *p-n-p* o parametrach zbliżonych do tranzystorów *n-p-n*,

— występowanie pasożytniczych pojemności oraz tranzystorów *p-n-p* (rys. 1) ograniczających częstość pracy układu,

— brak możliwości wykonywania elementów indukcyjnych w układzie scalonym wynikający z niemożności uzyskania elementów magazynujących pole magnetyczne w półprzewodniku.

Niektóre z tych ograniczeń są już usuwane. Stosowanie np. izolacji dielektrycznej (oddzielenie wysp elementów dwutlenkiem krzemu lub przerwą powietrzną) pozwala na znaczne zredukowanie pojemności pasożytniczych, a co za tym idzie — osiągnięcie wyższych częstości pracy układu. Znane są także takie układy, które mogą zastąpić elementy indukcyjne wytwarzając określone przesunięcie fazowe prądu i napięcia. Do takich układów należą żyratory oraz układy o sprzężeniu fazowym.

Wytwarzanie układów scalonych

układy monolityczne

Istnieje kilka zasadniczych technik wytwarzania układów scalonych. Ogromną większość produkowanych na świecie układów scalonych stanowią układy monolityczne, czyli takie, w których wszystkie elementy są wykonane w jednym kryształku krzemu zwanym strukturą, o rozmiarach rzędu kilku milimetrów kwadratowych. W układach monolitycznych osiągane jest największe zagęszczenie (tzw. gęstość upakowania) elementów elektronicznych — obecnie na kryształku o wymiarach ok. 5×5 mm można zmieścić ponad 60 000 tranzystorów tworzących określony układ elektroniczny. Układy wielkiej skali integracji, szczególnie zbudowane z tranzystorów MOS, zawierają prawie wyłącznie elementy czynne (np. oporniki zajmują zbyt dużą powierzchnię krzemu, szczególnie te, które mają wartość powyżej 10 kΩ).

układy cienko- i grubowarstwowe

Poza układami monolitycznymi, produkuje się w dość ograniczonym zakresie układy scalone cienko- i grubowarstwowe oraz hybrydowe. Układy cienkowarstwowe wytwarza się przez naparowanie elementów oraz połączeń między nimi w odpowiednich miejscach podłoża (szklana lub ceramiczna płytka) w postaci cienkich warstw różnych materiałów (niepółprzewodników). Układy grubowarstwowe wykonuje się przy użyciu past przewodzących metodą druku sitowego. Układy hybrydowe stanowią połączenie techniki monolitycznej i warstwowej, gdyż składają się z dyskretnych półprzewodnikowych elementów czynnych, a nawet układów mikroelektronicznych dolutowanych do połączeń będących ścieżkami cienkich warstw metalu naparowanych na ceramicznym podłożu. Oporniki i kondensatory w układach hybrydowych są wytworzone, podobnie jak połączenia, w postaci cienkich warstewek odpowiedniego materiału, naparowanych na podłoże.

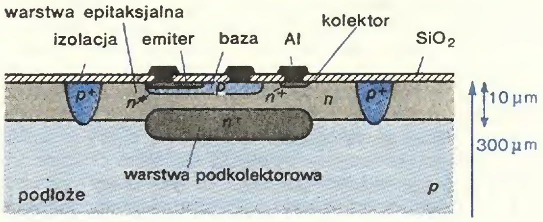
układy hybrydowe

Układy hybrydowe stanowią niewielki procent wszystkich produkowanych na świecie układów scalonych. Stosowane są one jako wzmacniacze dużych mocy do 100 W, przetworniki analogowo-cyfrowe i cyfrowo-analogowe o dużej dokładności. Wytwarza się je na specjalne zamówienia. Układy warstwowe mają dziś znaczenie raczej historyczne i stanowią jedynie niewiele znaczący margines.

Monolityczne układy scalone podzielić można ogólnie na wytwarzane techniką elementów bipolarnych oraz elementów unipolarnych, przy czym ostatnio zaznacza się tendencja do łączenia tych technik.

Układy scalone bipolarne

Elementami czynnymi w scalonych układach bipolarnych są tranzystory bipolarne, tzn. dwuzłączowe elementy *n-p-n* (rzadziej *p-n-p*), których nazwa związana jest z faktem, że udział w przewodzeniu prądu biorą nośniki ładunku obu rodzajów, tj. elektrony i dziury. Przekrój przez strukturę typowego tranzystora *n-p-n* w układzie scalonym ukazuje rys. 1.



Rys. 1. Tranzystor *n-p-n* w układzie scalonym

Przykładowy proces wytwarzania płytek z układami scalonymi przebiega następująco: płytka krzemu monokrystalicznego o przewodnictwie typu *p* (płytka podłożowa) jest szlifowana, polerowana i trawiona chemicznie w celu usunięcia wszelkich nierówności powierzchni, oczyszczenia jej i nadania idealnej gładkości i równoległości płaszczyzn. Następnie cała powierzchnia płytki zostaje utleniona w celu wytworzenia tzw. maski tlenkowej. Taka warstwa tlenku (maska) skutecznie zapobiega przedostawaniu się atomów pierwiastka użytego jako domieszka w procesie dyfuzji do obszaru krzemu znajdującego się pod nią. Obszary krzemu, do których chcemy wprowadzić daną domieszkę, zostają odsłonięte w procesie trawienia fotolitograficznego.

maska

proces fotolitograficzny

Na płytkę krzemu nakłada się emulsję światłoczułą, przesłania się ją maską w postaci szklanej płytki pokrytej czarną emulsją, na której wytworzone są w skali 1:1 wzory geometryczne stanowiące topografię struktury układu scalonego i się ją naświetla. Następnie się emulsję światłoczułą wywołuje. Z miejsc nie naświetlonych emulsję, wraz ze znajdującym się pod nią tlenkiem, usuwa się za pomocą odpowiednich rozpuszczalników. Opis powyższy dotyczy tzw. emulsji negatywowej. Stosowana jest także emulsja pozytywowa, przy użyciu której warstwa tlenku zostaje usunięta z obszarów naświetlonych. Każdy z tych rodzajów emulsji ma swoje określone cechy i wybór jednej z nich podyktowany jest określonymi wymaganiami technologicznymi.

wytwarzanie warstw podkolektorowych

Opisana powyżej technika trawienia fotolitograficznego stosowana jest przy wytwarzaniu wszystkich obszarów dyfuzyjnych w krzemie (w strukturze). Jako pierwsze wytwarzane są tzw. warstwy podkolektorowe (zagrzebane, ang. *Buried Layer*).

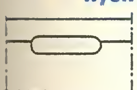
Obszary warstw podkolektorowych, a także obszary baz i emiterów tranzystorów, oraz obszary izolujące poszczególne elementy układu, są wytwarzane za pomocą procesu dyfuzji w ciele stałym. Płytki umieszczone zostają w piecu o specjalnej konstrukcji, utrzymującym temperaturę rzędu 1000°C z dokładnością do 0,5°C. Do pojemnika zawierającego płytki wprowadza się w odpowiedni sposób gaz (wodór lub azot) unoszący atomy odpowiedniego pierwiastka chemicznego, stanowiącego domieszkę dla materiału półprze-

wodnikowego, jakim jest krzem. W celu wytworzenia obszarów o typie przewodnictwa *n* stosuje się w technologii układów scalonych arsen, antymon i fosfor; natomiast do wytworzenia obszarów o typie przewodnictwa *p* stosuje się bor. Atomy pierwiastka domieszkującego osiadają na płytkach i wnikają w krzem w miejscach odsłoniętych (tzw. okna tlenkowe) w procesie trawienia fotolitograficznego, tworząc obszary o wymaganym typie przewodnictwa, a pomiędzy nimi złącza *p-n*. Oczywiście wprowadzone atomy domieszki nie naruszają budowy krystalicznej półprzewodnika.

okna tlenkowe

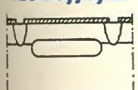
Warstwy podkolektorowe są obszarami o przewodnictwie typu *n*, silnie domieszkowanymi, a więc mającymi niski opór. Wytwarzane są one w takich miejscach, że obszary baz tranzystorów *n-p-n* znajdą się dokładnie nad nimi. Zadaniem warstw podkolektorowych jest usprawnienie transportu nośników prądu (zmniejszenie oporu obszaru łączącego). Uzyskane dzięki temu zmniejszenie oporu kolektorów powoduje polepszenie wzmacniających właściwości tranzystora, zmniejszenie napięcia nasycenia tranzystora, zmniejszenie pasożytniczej pojemności pomiędzy kolektorem a podłożem, czyli „masą" układu, ograniczającej częstość pracy układu.

wytwarzanie warstw epitaksjalnych



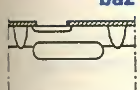
Po wytworzeniu warstw podkolektorowych usunięty zostaje tlenek z całej powierzchni płytki i zostaje ona przygotowana do następnego procesu technologicznego, to jest do osadzania warstwy epitaksjalnej. Proces epitaksji polega na takim osadzaniu atomów (krzemu) z fazy gazowej na monokrystalicznym podłożu struktury, aby powstająca warstwa wiernie i bez zniekształceń zachowała budowę krystaliczną podłoża. Podwyższona temperatura procesu (rzędu 1200°C) nadaje atomom osiadającym na podłożu pewną zdolność przemieszczania się, co ułatwia im zajmowanie miejsc określonych przez budowę krystaliczną krzemu, a więc umożliwia wytworzenie warstwy będącej kopią podłoża. Wysoka temperatura procesu epitaksji powoduje także ruch atomów domieszkujących warstwę podkolektorową, co w efekcie daje wybrzuszenie w głąb warstwy epitaksjalnej.

wytwarzanie obszarów izolacyjnych

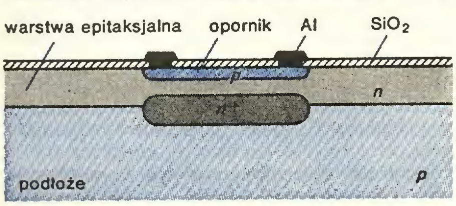


W wytworzonej warstwie epitaksjalnej znajdować się będą wszystkie elementy czynne i bierne układu, zachodzi więc konieczność wzajemnego ich odseparowania. W tym celu odsłonięte za pomocą procesu fotolitograficznego obszary warstwy epitaksjalnej domieszkuje się dyfuzyjnie tak, aby wytworzyły się obszary izolacyjne o przewodnictwie typu *p* łączące się z podłożem. Pozostałe obszary — o typie *n* — mają wtedy kształt odizolowanych wysp. W gotowym układzie scalonym obszary izolujące zostają przyłączone do najniższego potencjału, zaś obszary wysp — do potencjału najwyższego. W ten sposób poszczególne wyspy zostają wzajemnie oddzielone przez zaporowo spolaryzowane złącza *p-n*. Ze względu na rodzaj zastosowanej izolacji układy takie nazywane są układami z izolacją złączową. Nie jest to optymalny rodzaj izolacji zarówno ze względu na zajmowaną powierzchnię płytki, jak i na jej jakość (prądy upływu, napięcie przebicia, pojemności pasożytnicze); jednakże z powodu najłatwiejszego wytwarzania jest on w chwili obecnej powszechnie stosowany.

wytwarzanie baz

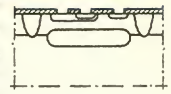


Po otrzymaniu odseparowanych obszarów, następnym krokiem prowadzącym do otrzymania układu scalonego jest wytworzenie poszczególnych jego elementów. W tym celu po otwarciu odpowiednich okien tlenkowych do obszaru warstwy epitaksjalnej wprowadza się dyfuzyjnie taką domieszkę, aby część tych obszarów zmieniła typ przewodnictwa na *p*, tworząc w ten sposób bazy tranzystorów *n-p-n*. Jednocześnie w tym samym procesie wytwarza się oporniki w postaci pojedynczych pasków o określonej szerokości i długości (rys. 2).



Rys. 2. Opornik w układzie scalonym

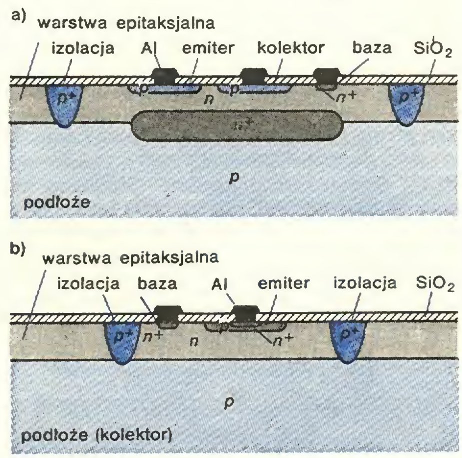
wytwarzanie emiterów



Emitery tranzystorów powstają przez wytworzenie w części obszarów baz obszarów o przewodnictwie typu *n*. Obszar emitera jest domieszkowany silniej od obszaru bazy dla uzyskania większego wzmocnienia prądowego tranzystora, ma więc znacznie niższy opór. Właściwość tę czasem wykorzystuje się do otrzymywania oporników o małych wartościach oporu. W tym samym procesie powstaje silnie domieszkowany obszar typu *n* w miejscu, gdzie następnie będzie wykonany kontakt kolektorowy do warstwy epitaksjalnej. Poprawia to znacznie właściwości kontaktu.

tranzystor boczny (lateralny)

W układzie monolitycznym można wytworzyć dwa rodzaje tranzystorów *p-n-p*: tzw. tranzystor boczny, czyli lateralny, oraz tranzystor podłożowy, czyli wertykalny. Tranzystor boczny, pokazany na rys. 3a, wytwarzany jest w ten sposób, że jego emiter i kolektor wykonuje się jednocześnie z bazami tranzystorów *n-p-n*, w związku z czym bazę takiego tranzystora stanowi warstwa epitaksjalna, a oba złącza, kolektorowe i emiterowe, mają tę samą głębokość. W tranzystorach tych, podobnie jak w wypadku kolektora tranzystora *n-p-n* wytwarza się silnie domieszkowany



Rys. 3. Tranzystor *p-n-p* w układzie scalonym: a) boczny, b) podłożowy

obszar dyfuzyjny typu *n*, ułatwiający kontakt do warstwy epitaksjalnej — bazy. W takim tranzystorze przepływ nośników w obszarze bazy odbywa się w kierunku równoległym do powierzchni struktury. Współczynnik wzmocnienia prądowego takiego tranzystora oraz maksymalna częstość jego pracy są znacznie mniejsze niż dla tranzystora *n-p-n*.

tranzystor podłożowy (wertykalny)

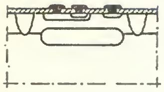
Nieco lepsze parametry mają tranzystory podłożowe (rys. 3b), które powstają z pasożytniczego tranzystora *p-n-p*, składającego się z emitera (obszar typu *p*, w którym dodatkowo wytwarza się obszar typu *n*, silnie domieszkowany), obszaru *n* warstwy epitaksjalnej oraz podłoża (typu *p*) będącego kolektorem tranzystora. Tranzystory podłożowe mogą pracować tylko w układzie ze wspólnym kolektorem, co stanowi poważne ograniczenie ich zastosowania.

wytwarzanie diod i kondensatorów

Diody w układzie scalonym są uzyskiwane przez wykorzystanie złącza *p-n* pomiędzy obszarami baz i emiterów (złącze emiter–baza ma lepszą charakterystykę w kierunku przewodzenia od złącza kolektor–baza). Natomiast kondensatory w układzie scalonym można otrzymywać wykorzystując pojemność złącza *p-n* lub też jako struktury MOS, gdzie jedną okładkę kondensatora stanowi np. dyfuzyjna warstwa o przewodnictwie typu *n* (emiter), drugą okładkę — warstwa aluminiowa, zaś dielektrykiem jest warstwa tlenku krzemu SiO_2.

Jak widać z powyższego przeglądu, technologia planarna umożliwia uzyskanie następujących elementów układu scalonego: 1) tranzystorów *n-p-n* cechujących się w zależności od potrzeb częstością maksymalną pracy rzędu 500 MHz lub mocą rzędu kilku watów, 2) diod, 3) oporników o oporze od kilkudziesięciu Ω do kilkuset kΩ, 4) kondensatorów o wartościach rzędu kilkudziesięciu pF, 5) tranzystorów *p-n-p* o współczynnikach wzmocnienia prądowego rzędu 10 i częstościach granicznych rzędu kilku MHz.

wytwarzanie połączeń



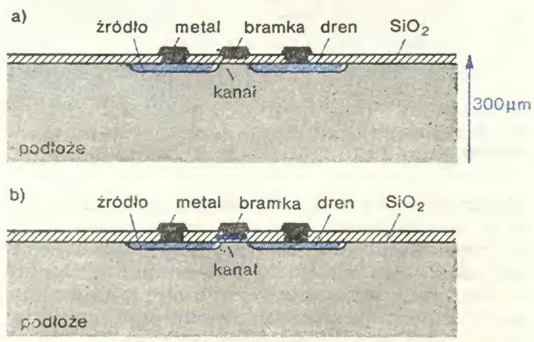
Gdy wszystkie elementy układu scalonego są już wytworzone, należy wykonać odpowiednią sieć połączeń elektrycznych pomiędzy nimi. Cała płytka zostaje przykryta warstwą tlenku. W procesie fotolitograficznym zostają wytworzone okna tlenkowe w miejscach, gdzie mają powstać kontakty elektryczne. Następnie na całą płytkę zostaje nałożona warstwa aluminium, z której po nowym procesie fotolitograficznym pozostają tylko ścieżki metaliczne stanowiące połączenia pomiędzy poszczególnymi elementami układu.

Dalsze etapy otrzymywania bipolarnych układów scalonych, tzn. sprawdzanie i obudowa są podobne, jak w wypadku niżej opisanych unipolarnych układów scalonych.

Układy unipolarne

tranzystor MOS

Funkcję analogiczną do tej, jaką w bipolarnych układach scalonych spełniają tranzystory *n-p-n*, w układach unipolarnych pełnią tranzystory polowe typu MOS (metal–tlenek–półprzewodnik, ang. *Metal–Oxide–Semi-Conductor Field–Effect–Transistor, MOS–FET*). Proces technologiczny produkcji układów scalonych typu MOS jest zasadniczo identyczny z opisanym poprzednio procesem technologii planarnej scalonych układów bipolarnych, z tą różnicą, że potrzebny jest tylko jeden proces dyfuzji, w którym wykonywane są zarówno źródło jak i dren (rys. 4) tranzystora MOS.



Rys. 4. Tranzystor MOS w układzie scalonym: a) tranzystor z bramką metalową, b) tranzystor z bramką krzemową

bramki tranzystorów MOS

Powierzchnię odpowiednio przygotowanej płytki monokrystalicznego krzemu utlenia się. Następnie w procesie fotolitograficznym odsłania się obszary, w których mają być tranzystory MOS. Po czym nakłada się na część tych obszarów nową warstwę tlenku, ale znacznie cieńszą od pierwotnej. Cienkie warstwy tlenku pokrywa się z kolei polikrystalicznym krzemem odpowiednio domieszkowanym w celu uzyskania dobrej przewodności elektrycznej. W ten sposób uzyskuje się bramki tranzystorów MOS. Przykładane do bramki napięcie wywołuje w obszarze pod nią (w tzw. kanale) pole elektryczne, którym można wpływać na prąd elektryczny płynący przez tranzystor MOS.

Bramka jest więc elektrodą sterującą i spełnia funkcję analogiczną do tej, którą w lampie próżniowej spełnia siatka sterująca. Istnieje kilka typów tranzystorów MOS, różniących się między sobą sposobem wytwarzania bramki.

W następnym etapie procesu wytwarzania tranzystorów MOS w układzie scalonym uzyskana bramka krzemowa służy jako maska osłaniająca kanał. Zaletą takiej technologii jest znaczne zmniejszenie pojemności pasożytniczych, a także ułatwienie procesu fotolitograficznego. Bramki z polikrystalicznego krzemu można także wykorzystać jako dodatkowe połączenia elementów układu scalonego (szczególnie przydatne, gdy trzeba stosować połączenia skrzyżowane). Niekiedy wykonuje się bramki z trudno topliwego metalu, jak molibden lub tytan, co polepsza pewne właściwości tranzystorów MOS, lecz technologia ta jest znacznie kosztowniejsza.

źródło i dren

Nie przesłonięte bramką obszary tranzystora, tzw. źródło i dren, domieszkuje się w jednym procesie dyfuzji, a więc mają one taki sam typ przewodnictwa.

W pojedynczym tranzystorze MOS w zależności od przyłożonego napięcia źródło nośników ładunku może być drenem, czyli ujściem prądu, natomiast w układzie scalonym zawsze spełnia tę samą funkcję.

Standardowe tranzystory MOS z bramką metalową wykonuje się najczęściej z aluminium, ale wtedy ulega zmianie, ze względu na niską temperaturę topnienia aluminium, kolejność wykonywania elementów. Najpierw w procesach fotolitograficznym i dyfuzyjnym otrzymuje się dreny i źródła, a następnie na cienką warstwę tlenku nakłada się warstwę aluminium (zob. il. 102, tabl. 25).

Rozróżniamy dwa rodzaje tranzystorów MOS z kanałem typu *n* i z kanałem typu *p*. W tranzystorze z kanałem *n* nośnikami ładunku są elektrony, natomiast w kanale *p* — dziury. W jednym układzie scalonym można wytwarzać oba rodzaje tranzystorów MOS. Układy takie, zwane komplementarnymi układami MOS (COS/MOS lub CMOS) mają wiele zalet, jak np. większą szybkość działania niż w zwykłych układach MOS, zmniejszone zużycie mocy, co jest szczególnie ważne w wypadku układów wielkiej skali integracji.

komplementarne układy MOS

zalety i wady układów MOS

W porównaniu z układami bipolarnymi, układy MOS mają mniejsze wymiary elementów czynnych i nie mają izolacji pomiędzy nimi, a więc są znacznie mniejsze. Wadami układów unipolarnych są: wolniejsze działanie (pracują na niższych częstościach) i stosunkowo nieduża wartość prądów wyjściowych, co jest poważną wadą w dużych systemach cyfrowych.

W układach MOS praktycznie nie ma elementów biernych, w razie potrzeby mogą być wytworzone w identyczny sposób jak w układach unipolarnych.

Montaż i pomiary układów scalonych

Z chwilą ukończenia procesów technologicznych wytwarzających układy scalone (unipolarne lub bipolarne) w płytce półprzewodnikowej rozpoczyna się proces montażu. Jest on poprzedzony pomiarami parametrów elektrycznych każdego poszczególnego układu.

program pomiarowy

Do odpowiednich punktów metalizacji doprowadzone są cienkie igły stanowiące część skomplikowanego systemu pomiarowego złożonego z zasilaczy prądowych i napięciowych oraz przyrządów pomiarowych, sterowanych określonym programem pomiarowym. Program taki, złożony z wielu testów parametrów prądowych i funkcjonalnych, jest tak ułożony, że spełnienie wszystkich jego wymagań przez dany układ scalony daje duży stopień pewności, że układ taki po zamknięciu w obudowie będzie działał prawidłowo. Układy nie spełniające wymagań programu są odpowiednio znakowane i odrzucane po podzieleniu płytki.

Sprawdzone układy scalone zamyka się w obudowach w celu mechanicznego zabezpieczenia, zapewnienia niezawodności działania i umożliwienia zainsta-

lowania ich w urządzeniach. Przedtem odpowiednie punkty metalizacji łączy się za pomocą złotych drucików o grubości 0,025 mm z zewnętrznymi wyprowadzeniami obudowy (il. 100, tabl. 25). Obudowane układy scalone podlegają dokładnej kontroli parametrów układu; są to tzw. badania niezawodnościowe, do których należy także np. sprawdzenie, jak działają układy po wygrzewaniu w temperaturze 100°C przez 1000 godzin w czasie zasilania ich energią elektryczną.

Zastosowania układów scalonych

komputery

Największe zastosowanie mają układy scalone w komputerach, których rozwój stał się możliwy właśnie dzięki pojawieniu się układów scalonych. Cyfrowe układy scalone, głównie typu MOS, mogą być układami logicznymi, arytmetycznymi, procesorami, pamięciami, układami sterującymi urządzeniami peryferyjnymi (jak drukarki, dziurkarki, czytniki itp.). Układy scalone są podstawą działania komputerów, minikomputerów, mikrokomputerów, kalkulatorów i in. Kalkulator kieszonkowy firmy amerykańskiej Hewlett-Packard, mający układ scalony (mikrokomputer) o wymiarach rzędu kilku mm^2, wykonuje 86 różnych operacji matematycznych, ma możliwość zapamiętania dowolnego programu obliczeniowego składającego się z 49 operacji. Ilustracja 9 (tabl. 3) przedstawia w powiększeniu strukturę obwodu scalonego mikrokomputera. Układy scalone znalazły też zastosowanie w centralach telefonicznych, maszynach do księgowania i operacji bankowych, w systemach sterowania produkcją, w kontroli przebiegu procesów chemicznych, w diagnostyce medycznej, w przyrządach pomiarowych itp.

mikrokomputery

Mikrokomputery, czyli miniaturowe maszyny matematyczne, znajdują coraz szersze zastosowania tam, gdzie jeszcze kilka lat temu nie myślano o używaniu komputerów. Mikrokomputer sterujący sygnalizacją świetlną ruchu ulicznego będzie uwzględniał charakterystykę poszczególnych skrzyżowań, lokalne przepisy, a także stale zmieniające się czynniki, jak natężenie ruchu, pora dnia i roku. Mikrokomputery stosowane w samochodach kontrolować będą: światła samochodu, stan hamulców i akumulatora, zapięcie pasów bezpieczeństwa przez jadących, system przeciwpoślizgowy, sterowanie zapłonem, regulację mieszanki paliwowej, zawartość szkodliwych składników w spalinach oraz zużycie paliwa. Rozwój telekomunikacji, a szczególnie komunikacji satelitarnej, nie byłby możliwy bez pojawienia się małych, niezawodnych, zużywających znikome ilości energii układów przetwarzających sygnały W sprzęcie wojskowym, aparaturze samolotowej i kosmicznej układy scalone są coraz powszechniej stosowane ze względu na swą niezawodność i wykonywanie skomplikowanych funkcji przy bardzo małym zużyciu mocy.

telekomunikacja

przedmioty powszechnego użytku

Także w sprzęcie przeznaczonym do powszechnego użytku, jak radioodbiorniki, telewizory, magnetofony, gramofony, bliski jest moment, gdy wszystkie funkcje elektroniczne przejęte zostaną przez wielofunkcyjne układy scalone. Przykład zastosowania układów scalonych w sprzęcie radiowo-telewizyjnym ukazuje il. 97 (tabl. 24). Dzięki układom scalonym możliwa stała się produkcja zegarków elektronicznych niezwykle dokładnych (il. 10, tabl. 3), wyświetlających w postaci cyfr i liter godziny, minuty, sekundy, dzień tygodnia, miesiąc, a nawet różnicę stref czasowych i lata przestępne (zob. też il. 11, tabl. 3).

wzmacniacze operacyjne

Istnieją także układy scalone, tzw. wzmacniacze operacyjne, mogące wykonywać oprócz funkcji wzmacniania sygnałów, operacje matematyczne jak logarytmowanie, całkowanie, różniczkowanie. Można je zastosować jako części składowe złożonych, czułych systemów pomiarowych, przetwarzania sygnałów, analogowych maszyn matematycznych itp.

Przyszłość układów scalonych

ulepszanie technologii

Prace nad doskonaleniem układów scalonych prowadzone są wielokierunkowo. Ulepsza się istniejące technologie w celu podniesienia wydajności i obniżenia cen układów scalonych. Uzyskuje się to głównie przez stosowanie coraz większych płytek krzemowych (na których można zmieścić większą liczbę układów scalonych), powiększanie skali integracji, zmniejszenie obszarów poszczególnych elementów w układzie scalonym, zmniejszanie oddziaływań pasożytniczych, itp. Prowadzone są prace nad nowymi technologiami otrzymywania układów scalonych, jak np. implantacja jonów, polegająca na wytwarzaniu żądanych obszarów w strukturze przez wprowadzenie atomów domieszek w głąb materiału za pomocą bombardowania jonami o dużej energii. Ogromne znaczenie ma rozwój nowych technik fotolitograficznych, pozwalających na uzyskiwanie elementów o coraz mniejszych rozmiarach, porównywalnych z długością fali światła widzialnego (np. opracowuje się procesy chemicznego trawienia plazmowego, pozwalające na zmniejszenie rozmiarów elementów poniżej 1 μm). Prowadzone są też badania nad zastosowaniem promieniowania rentgenowskiego lub wiązki elektronów do wytwarzania pożądanej topografii układu w maskach tlenowych.

układy bardziej złożone

Oprócz prac zmierzających do opracowania ulepszonych bądź nowych procesów technologicznych, które mogą być wykorzystywane do wytwarzania układów scalonych w obecnej postaci, trwają prace mające przynieść zupełnie nowe koncepcje konstrukcyjne. Prace z dziedziny mikroelektroniki mają na celu budowę coraz bardziej złożonych układów scalonych, spełniających coraz bardziej skomplikowane funkcje. Układy takie muszą zawierać coraz więcej elementów i coraz poważniejszym ograniczeniem staje się konieczność izolacji poszczególnych elementów układu scalonego rozpatrywanego jako zbiór tranzystorów, diod, oporników. Obecnie myśli się o wytwarzaniu takich układów, w których odpowiednie funkcje byłyby realizowane przez nośniki ładunku przepływające w materiale półprzewodnikowym nie podzielonym na poszczególne obszary typu *p* i *n*, tworzące tradycyjnie pojmowane tranzystory i diody. Przepływ ładunków elektrycznych byłby w odpowiedni sposób kontrolowany przez przykładane potencjały elektryczne.

układy CCD

Zwiastunem tej tendencji są układy o sprzężeniu ładunkowym (CCD, ang. *Charge Coupled Devices*), w których sygnał jest przenoszony przez ruch nośników ładunku odpowiednio sterowanych oddziaływaniem zewnętrznym w ciągłym ośrodku półprzewodnikowym, bez wyraźnie określonych elementów czynnych. Już w tej chwili produkowane są pamięci o dużej pojemności (16 kilobitów) w technice CCD. Układy takie otwierają niewątpliwie drogę nowej rodzinie przyrządów półprzewodnikowych — rodzinie bloków funkcjonalnych.

układy VLSI

W ten niezwykle szeroki zakres prac naukowo-badawczych są zaangażowane wielkie zespoły naukowców i inżynierów dysponujących najnowocześniejszymi osiągnięciami elektroniki, fizyki, chemii i inżynierii materiałowej. W połowie lat osiemdziesiątych mają zostać wprowadzone do produkcji układy o bardzo wielkiej skali integracji (VLSI od ang. *Very-Large Scale Integration*) zawierające powyżej miliona elementów czynnych, a zapowiedzią tego, co przyniosą najbliższe lata, może być fakt wyprodukowania w 1979 r. w laboratoriach firm japońskich pamięci półprzewodnikowej o pojemności 1 Mb (1 milion bitów).

A. Ambroziak *Półprzewodnikowe układy scalone*, Warszawa 1966; K. Badźmirowski, J. Pieńkos, W. Piestrzyński *Cyfrowe układy MOS-LSI*, Warszawa 1979; M. Białko *Układy mikroelektroniczne*, Warszawa 1969; J. Eimbinder *Zastosowania liniowych układów scalonych*, Warszawa 1974; A. Filipkowski *Mikroelektroniczne układy scalone*, Warszawa 1966; E. Keonjian *Mikroelektronika*, Warszawa 1967; Z. Kulka, M. Nadachowski *Liniowe układy scalone i ich zastosowanie*, Warszawa 1975; W. Marciniak *Przyrządy półprzewodnikowe i układy scalone*, Warszawa 1979; J. Millman, C. Halkias *Układy scalone analogowe i cyfrowe*, Warszawa 1976.

Generacja mikrofal

Janusz Konopka

Oddziaływanie promieniowania elektromagnetycznego z ośrodkami materialnymi jest jednym z podstawowych narzędzi badawczych fizyki. Spośród różnych zakresów widmowych mikrofale, a zwłaszcza fale milimetrowe i submilimetrowe (tzn. zakres fal od 1 cm do 0,1 mm) były najdłużej niedostępne dla badań — głównie z powodu trudności technicznych w ich wytwarzaniu.

Prawdziwy rozwój techniki generacji fal centymetrowych i milimetrowych nastąpił dopiero w ostatnim trzydziestoleciu, w ostatnim zaś dziesięcioleciu opanowano dziedzinę wytwarzania fal submilimetrowych, w ten sposób została wypełniona jedna z ostatnich „dziur" w pasmie generowanych spójnie fal od drgań podakustycznych do fal świetlnych. Należy podkreślić, że głównym stymulatorem rozwoju techniki mikrofal był radar i systemy radiokomunikacyjne.

Ostatnie lata przyniosły również zasadnicze zmiany w sposobie generacji mikrofal. Poprzednio wszystkie typy oscylatorów mikrofalowych wykorzystywały oddziaływanie wiązki elektronów w próżni z mikrofalowymi obwodami rezonansowymi (tzw. rezonatorami wnękowymi) przy ewentualnym dodatkowym przyłożeniu stałego pola magnetycznego. W 1963 r. — dzięki odkryciu J. B. Gunna — okazało się, że spójne drgania elektryczne w zakresie mikrofal mogą być emitowane przez pewne kryształy półprzewodnikowe.

źródła mikrofal

W chwili obecnej do najważniejszych źródeł mikrofal zaliczyć należy ciągle jeszcze stosowane lampy mikrofalowe: klistrony, karcinotrony (lampy z falą wsteczną) i magnetrony, urządzenia półprzewodnikowe: oscylator Gunna, zwany inaczej generatorem z przeniesieniem elektronów, oraz różne konstrukcje półprzewodnikowych diod lawinowych, a ostatnio także tranzystory bipolarne i tranzystory polowe.

Artykuł niniejszy stawia sobie za cel omówienie zasad generacji mikrofal za pomocą najbardziej dziś popularnych i mających duże znaczenie praktyczne urządzeń, jakimi są klistron refleksowy i generator Gunna zbudowany na krysztale arsenku galu.

Klistron refleksowy

Próby podwyższenia częstości generacji oscylatorów zbudowanych przy użyciu lamp elektronowych (triody i tetrody) wykazują, że istnieje pewna częstość graniczna, której nie można przekroczyć zmieniając konstrukcję czy też zmniejszając wymiary lamp.

ograniczenia lamp elektronowych

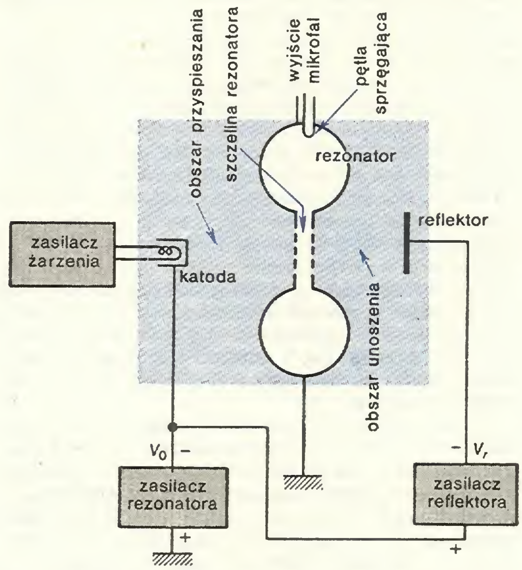
Jedną z przyczyn jest fakt, że powiększenie częstości rezonansowej obwodu oscylatora wiąże się ze zmniejszaniem jego indukcyjności lub pojemności. Pojemność jednak nie może być mniejsza niż pojemność międzyelektrodowa lampy, graniczna zaś indukcyjność sprowadza się do indukcyjności doprowadzeń lampy plus indukcyjności najkrótszych połączeń między końcówkami elektrod. Drugą podstawową przyczyną istnienia częstości granicznej jest skończony czas przelotu elektronów od katody do anody. Oscylatory triodowe nie mogą sprawnie pracować, jeśli ten czas przelotu przekracza choćby tylko małą część, np. $^1/_{10}$ okresu oscylacji, niezależnie od tego, jakie się poczyni modyfikacje w elementach czy dostrojeniu obwodu rezonansowego.

Wymienione trudności nie występują w innego typu lampach, tzw. klistronach. Cechą wyróżniającą klistrony spośród pozostałych lamp elektronowych jest specjalna metoda, za pomocą której napięcie wejściowe kontroluje ruch elektronów w lampie. W odróżnieniu od zwyczajnych lamp, w których elektroda sterująca (siatka) określa liczbę elektronów przelatujących od katody do anody w jednostce czasu, napięcie wejściowe w klistronie oddziałuje jedynie na prędkość lotu elektronów przy niezmienionej ich liczbie. Z tego względu klistrony nazywane są lampami z modulowaną prędkością.

Istnieje wiele typów klistronów w zależności od ich przeznaczenia, zakresu częstości i dostarczanej mocy. Do najbardziej znanych należą: klistron dwuwnękowy, stosowany na ogół jako wzmacniacz i generator wielkich mocy, oraz najpopularniejszy dotychczas generator mikrofal, klistron refleksowy.

budowa klistronu

Klistron refleksowy przedstawiony jest schematycznie na rys. 1. Jak widać, przestrzeń katoda–siatka, która w zwyczajnych lampach elektronowych kontroluje prąd anodowy, została tu zastąpiona przez trzy oddzielne obszary. Każdy z obszarów pełni inną funkcję, można więc mówić o swoistym podziale pracy. Obszar pierwszy, rozciągający się od katody do rezonatora wnękowego, wykorzystywany jest do nadania elektronom pełnego przyspieszenia w stałym polu elektrycznym o natężeniu E rzędu wielu tysięcy woltów na metr. Drugi obszar stanowi przewężenie cylindrycznego lub toroidalnego rezonatora wnękowego i jest przysłonięty najczęściej metalową siatką, przez którą mogą swobodnie przelatywać elektrony. Obszar ten nazywamy szczeliną rezonatora lub szczeliną w.cz. (wielkiej częstości). Nie ma tutaj pól stałych, jest natomiast zmienne pole elektryczne o częstości własnej rezonatora wnękowego.



Rys. 1. Klistron refleksowy w przedstawieniu schematycznym. W obszarze niebieskim panuje próżnia typowa dla lamp elektronowych (10^{-4} Pa). Zasilacz rezonatora dostarcza napięcia przyspieszającego elektrony. Napięcie to musi być stabilne i zawiera się w granicach od 300 do 5000 V przy prądach od 10 mA wzwyż, zależnie od typu klistronu. Zasilacz reflektora wytwarza pole elektrostatyczne hamujące elektrony. Zasilacz pracuje bezprądowo przy napięciach od 100 do 1000 V zależnie od typu klistronu. Musi on być niezwykle stabilny, gdyż decyduje o stabilności częstości drgań mikrofalowych wytwarzanych przez klistron. Napięcie żarzenia wynosi na ogół 6,3 V przy prądach od 0,6 do 3 A

Tu można sobie zadać pytanie, skąd się bierze pole wielkiej częstości w rezonatorze? Jaki jest początek procesu generacji mikrofal? Odpowiedź na to pytanie trudno jest znaleźć w podręczniku czy monografii traktującej opisywany w tym rozdziale temat. Wynika to zapewne z faktu, że za istotne uważa się zazwyczaj to, co się dzieje w tzw. stanie ustalonym. Trwa on poniżej 1 μs i jest bardzo trudny do zaobserwowania.

Proces narastania pola elektromagnetycznego w rezonatorze klistronu, jeśli pominąć pole szumów termicznych, związany jest z przelotem przezeń pierwszych ładunków elektrycznych, które zdąży wyemitować katoda. Szybko poruszający się elektron stanowi impuls prądu elektrycznego o czasie trwania równym czasowi przelotu przez szczelinę rezonatora. Energia takiego impulsu jest sumą energii drgań o wielu częstościach. Im krótszy impuls, tym wyższe częstości są w nim zawarte (lub jak się czasem mówi — tym szersze jest jego widmo fourierowskie). Łatwo więc znaleźć w tym widmie również częstości własne rezonatora. Wiemy następnie, że rezonator raz pobudzony wykonuje, jak potrącona palcem struna, wiele okresów drgań. W ten sposób powstaje pole elektromagnetyczne, które może wpływać na ruch elektronów przelatujących przez rezonator w czasie późniejszym, jak to opisano w dalszej części tego rozdziału. Pole eletromagnetyczne zmienia kierunek co pół okresu drgań rezonatora, może więc przyspieszać bądź opóźniać elektrony. W obszarze tym mamy zatem do czynienia z modulacją prędkości elektronów. Szerokość szczeliny jest bardzo mała, na ogół nie przekracza 100 μm.

modulacja prędkości elektronów

Trzeci obszar, rozciągający się od zewnętrznej okładziny rezonatora do tzw. reflektora, zwany jest obszarem unoszenia lub przestrzenią odbijającą. Sam reflektor jest elektrodą w postaci metalowego krążka i połączony jest ze źródłem napięcia ujemnego w stosunku do katody. Wytwarza więc pole elektryczne silnie hamujące elektrony. Z reguły elektrony zawracają w kierunku szczeliny wielkiej częstości nie osiągnąwszy elektrody reflektora. W obszarze tym nie ma pól zmiennych i na wszystkie elektrony — niezależnie od prędkości, jaką mają przy wejściu do obszaru unoszenia — działa jednakowa siła, równa co do wielkości iloczynowi ładunku i natężenia pola, eE. Ponieważ prędkości elektronów przy wejściu w przestrzeń odbijającą są różne, występuje efekt grupowania czy też paczkowania elektronów. Proces ten, zasadniczy dla pracy klistronu, wymaga bliższego omówienia.

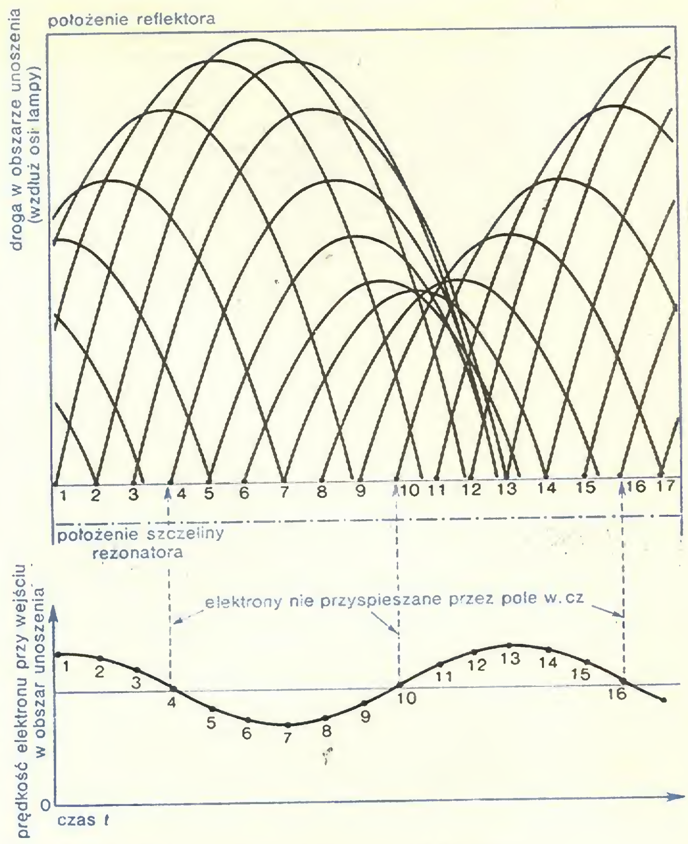
grupowanie elektronów

Elektrony o największej prędkości dolatują najbliżej reflektora i potrzebują najdłuższego czasu, aby powrócić do szczeliny rezonatora. Odwrotnie jest z elektronami wchodzącymi w obszar unoszenia z prędkością najmniejszą. W ten sposób paczka tworzy się wokół elektronu przechodzącego przez szczelinę rezonatora w chwili, gdy pole w.cz. zmienia się z przyspieszonego na opóźniające.

Ruch elektronów jest analogiczny do ruchu kul wyrzuconych w górę przeciw siłom ziemskiego pola grawitacyjnego. Kula wyrzucona z dużą prędkością wzniesie się wyżej niż kula wyrzucona z prędkością mniejszą i będzie potrzebować dłuższego czasu, aby dotknąć z powrotem powierzchni Ziemi. W ten sposób kilka kul wyrzuconych w różnych momentach z różnymi prędkościami może upaść równocześnie na Ziemię.

Przy pierwszym przejściu przez szczelinę rezonatora liczba elektronów przyspieszanych, tzn. pobierających energię od pola w.cz., jest średnio taka sama jak liczba elektronów opóźnionych, tzn. oddających energię polu w.cz. Bilans energetyczny jest więc równy zeru. Przy powrocie do szczeliny rezonatora, tzn. po procesie paczkowania, bilans energetyczny może być dodatni lub ujemny, zależnie od tego, czy paczka będzie spowolniona czy przyspieszona przez pole. Oczywiście największy zysk energii oscylacyjnej, a tym samym maksimum generowanej przez klistron mocy mikrofalowej, wystąpi, gdy środek paczki przechodzić będzie przez szczelinę w momencie największego spowolnienia.

Łatwo wykazać, że maksima mocy mikrofal generowanej przez klistron wystąpią wówczas, gdy czas przebywania elektronów w przestrzeni odbijającej wynosić będzie $n+{}^3/_4$ okresu drgań rezonatora. Wydedukować to można również dla $n = 0$ na podstawie rys. 2.



Rys. 2. Wykresy obrazujące proces grupowania się elektronów w klistronie refleksowym. Dolny wykres obrazuje prędkość ponumerowanych elektronów (*1, 2, 3, ...*) w czasie opuszczania szczeliny rezonatora, tj. przy wejściu w obszar unoszenia. Wykres górny pokazuje lot tych samych elektronów w obszarze unoszenia w funkcji czasu, aż do powrotu do szczeliny rezonatora. Elektron *4*, nie przyspieszany przez pole w.cz., powraca razem z elektronami *3* i *5* po upływie około ³/₄ okresu drgań rezonatora. Elektrony te są silnie spowalniane przez pole w.cz., a tracona przez nie energia kinetyczna przemienia się w energię oscylacji mikrofalowych

Czas przebywania elektronów w przestrzeni odbijającej regulować można bezpośrednio, za pomocą zmian napięcia reflektora, lub pośrednio — przez zmianę napięcia przyspieszającego.

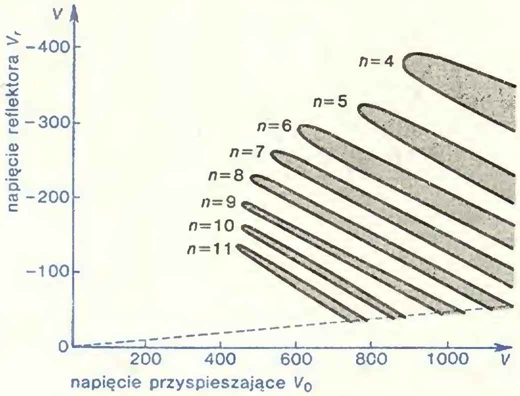
Chociaż spełnienie warunku na czas przejścia jest konieczne do uzyskania oscylacji, nie oznacza to, że warunek ten jest dostateczny. Jeśli liczba elektronów przelatujących przez szczelinę w jednostce czasu, czyli inaczej mówiąc — prąd wiązki elektronów jest zbyt mały, energia dostarczana przez „małe" paczki elektronów może nie być dostateczna do przezwyciężenia strat energii w samym obwodzie rezonatora i strat wynikających ze sprzężenia rezonatora ze światem zewnętrznym (obciążeniem). Zatem przekroczenie pewnego minimum prądu wiązki jest drugim koniecznym warunkiem otrzymania oscylacji. Gdy obydwa warunki są spełnione, klistron dostarcza energii mikrofal do obciążenia.

czas przejścia elektronów

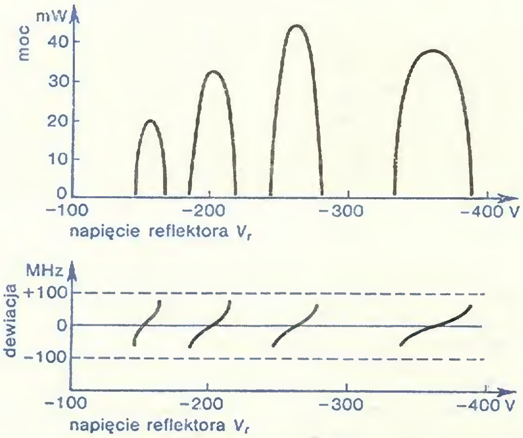
Czas przejścia elektronów przez obszar unoszenia zależy, jak już wspomniano, od napięcia reflektora V_r i napięcia przyspieszającego V_0. Oscylacje powstaną przy pewnych przedziałach wartości tych parametrów. Przedziałom tym można przypisać określoną liczbę całkowitą n, która jest miarą czasu przebywania elektronów w przestrzeni odbijającej, liczonego w okresach drgań własnych rezonatora. Oscylacje odpowiadające różnym n zwane są modami, trybami lub rodzajami pracy klistronu (nie należy ich mylić z rodzajami drgań występujących we wnękach rezonansowych).

mody pracy klistronu

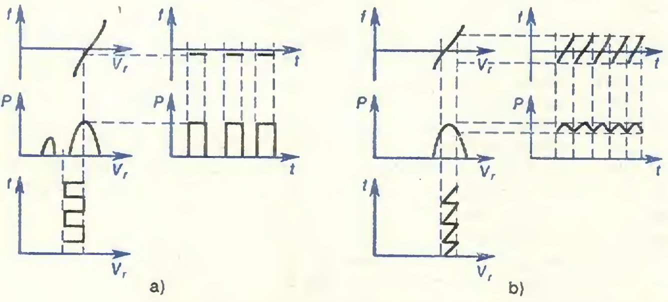
Mody pracy typowego klistronu refleksowego generującego falę długości 1 cm przedstawione są na rys. 3. Należy dodać, że rozkład i szerokość poszczególnych modów zależy również od konstrukcji klistronu (geometrii szczeliny rezonatora i przestrzeni odbijającej oraz od częstości). Z rys. 4 widać, że zarówno moc, jak i częstość w dużym stopniu zależy od napięcia reflektora. Na rys. 5 przedstawiono sposoby modulacji mikrofal generowanych przez klistron w granicach pojedynczego modu — pokazano zarówno modulację amplitudy (AM), jak i modulację częstości (FM). Możliwość łatwej modulacji ma zasadnicze znaczenie przy zastosowaniu klistronów.



Rys. 3. Wykres modów pracy typowego klistronu refleksowego na zakres fal 1 cm. Obszary szare odpowiadają przedziałom oscylacji dla poszczególnych modów *n* od *4* do *11*. (Praca urządzenia w obszarze poniżej linii przerywanej prowadziłaby do uszkodzenia reflektora przez bombardujące elektrony)



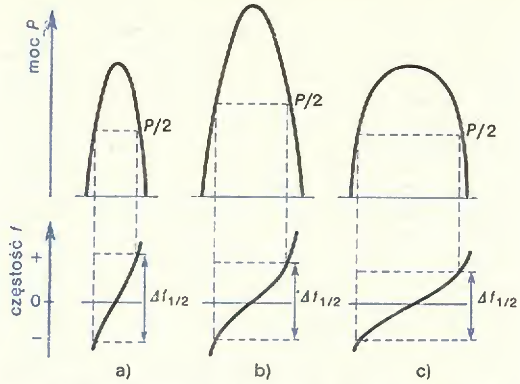
Rys. 4. Wykresy mocy generowanej przez klistron dla modów *4*, *5*, *6*, *7* z rys. 3 przy ustalonym napięciu przyspieszającym 1000 V. Na dole — odchylenie (dewiacja) częstości generowanej od częstości odpowiadającej maksimum mocy dla danego modu



Rys. 5. Sposoby modulacji mikrofal generowanych przez klistron refleksowy: a) modulacja amplitudy (AM) przy ustalonej częstości *f* możliwa jest tylko za pomocą fali prostokątnej; b) modulacja częstości (FM). Gdy dewiacja jest niewielka (rzędu kilku MHz) częstość *f* jest proporcjonalna do napięcia modulującego przy prawie stałej mocy *P*

moc mikrofal

Moc mikrofal generowanych przez klistron w ramach danego modu zależy również w znacznym stopniu od sprzężenia z obwodami zewnętrznymi, czyli z obciążeniem. Wpływ obciążenia na kształt modu ilustruje rys. 6. Przy słabym sprzężeniu (obciążeniu)



Rys. 6. Wpływ obciążenia na kształt modu klistronu refleksowego i dewiację częstości dla punktów połowy mocy: a) obciążenie zbyt wielkie; b) obciążenie optymalne; c) obciążenie za małe

rezonator klistronu ma większą dobroć, mod jest szeroki, a zmiany częstości przy zmianach V_r są w okolicy maksimum mocy nieznaczne. Oczywiście moc dostarczana przez klistron jest wówczas również mała. Przy zwiększaniu sprzężenia przechodzi się przez obszar, gdzie mod jest jeszcze dość szeroki, a dostarczana moc — duża. Dalsze zwiększanie sprzężenia z rezonatorem prowadzi do spadku mocy i znacznego zwężenia modu przy dużych niestabilnościach częstości, wynikających przede wszystkim z małej dobroci silnie obciążonego rezonatora.

Rozważania powyższe obrazuje prosta zależność ilościowa, wiążąca zmiany częstości pracy klistronu z odchyleniami fazy φ paczki elektronów powracających do szczeliny rezonatora od fazy optymalnej $\varphi = 0$ dla mocy maksymalnej oraz z dobrocią rezonatora Q (czyli stosunkiem energii elektromagnetycznej w nim nagromadzonej do sumy energii strat): $\Delta f/f = \mathrm{tg}\,\varphi/2Q$.

sprawność

Klistron refleksowy ma stosunkowo małą sprawność przetwarzania energii prądu stałego na energię mikrofal. Sprawność ta nie przekracza na ogół kilku % i jest mniejsza dla klistronów małej mocy. Zależy ona od napięć zasilających i częstości oraz od ukształtowania geometrycznego klistronu. Sprawność klistronów na fale milimetrowe nie sięga zazwyczaj 1%.

Do dziś nie powiedziano jeszcze ostatniego słowa, jak krótkie fale może generować klistron refleksowy. Modele wytwarzane seryjnie generują fale ok. 1 mm przy mocach kilku miliwatów. Największą wadą klistronów refleksowych jest wąski zakres przestrajania mechanicznego i elektronicznego (na ogół nie przekracza kilkunastu procent). Dużą niedogodnością jest również konieczność stosowania skomplikowanych zasilaczy, dostarczających trzech stabilnych napięć. Napięcie przyspieszające, które w klistronach na fale milimetrowe jest rzędu kilku kV, i napięcie reflektora muszą być bardzo dokładnie stabilizowane.

karcinotrony

W zakresie fal poniżej 1 mm jedynymi lampami mikrofalowymi mającymi znaczenie praktyczne są tzw. lampy z falą wsteczną, zwane inaczej karcinotronami. Dzięki nim osiągnięto zawrotną granicę generacji fali ok. 0,3 mm długości, a więc zbliżoną do częstości pracy lasera cyjanowodorowego (HCN).

Generator Gunna

Jeszcze przed kilku laty wykorzystanie tranzystorów do generacji mikrofal ograniczone było do zakresu poniżej 1 GHz. Obecnie w układach tranzystorowych

osiąga się częstość generacji ok. 30 GHz, a przy 10 GHz uzyskuje się moce rzędu 100 mW, a więc w wielu wypadkach wystarczające. W mikrofalowych generatorach tranzystorowych wykorzystuje się jako element sprzężenia zwrotnego zewnętrzne rezonatory wnękowe lub paskowe i w swej zasadzie działania generatory te nie różnią się zasadniczo od generatorów stosowanych na niższych częstościach.

generatory tranzystorowe

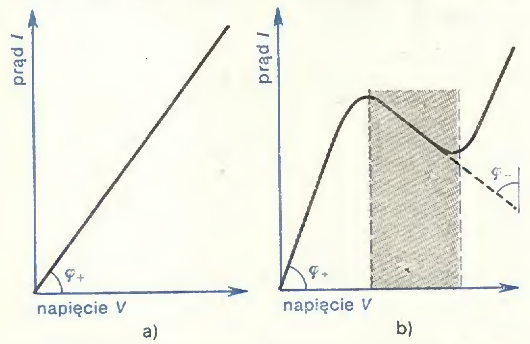
Oprócz generatorów tranzystorowych istnieje cała klasa generatorów półprzewodnikowych działających na zasadzie wytwarzania w płytce półprzewodnikowej tzw. oporu ujemnego. Należą do niej różne typy diod lawinowych, które mogą współpracować z obwodami mikrofalowymi aż do częstości ok. 100 GHz, oraz tzw. generatory Gunna. Te, choć oferują mniejsze moce i mniejsze sprawności przetwarzania energii niż diody lawinowe, pozwalają na uzyskiwanie bezkonkurencyjnej (w porównaniu z innymi generatorami półprzewodnikowymi) czystości spektralnej wytwarzanego promieniowania (szerokość linii poniżej 1 kHz przy częstości 10 GHz).

generatory z oporem ujemnym

Niezwykle interesująca zasada działania generatorów Gunna, jedynych urządzeń półprzewodnikowych generujących mikrofale, a nie zawierających złącz *p-n*, związana jest ze specyficzną strukturą pasmową niektórych półprzewodników 2-i 3-składnikowych.

Jak klistron wyróżnia się oryginalnością swej zasady działania wśród lamp elektronowych — modulacją prędkości i grupowaniem elektronów w przestrzeni unoszenia, tak generator Gunna wyróżnia się spośród urządzeń półprzewodnikowych przenoszeniem elektronów z niższego do wyższego minimum pasma przewodnictwa (gorące elektrony) i tworzeniem się ruchomych domen bardzo silnego pola elektrycznego wewnątrz półprzewodnika.

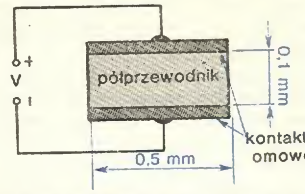
Jest rzeczą wiadomą, że jeśli w jakimś obwodzie elektrycznym znajduje się element o oporze ujemnym, to w obwodzie tym mogą się wzbudzić oscylacje. Opór ujemny nie jest cechą powszechnie spotykaną w przyrodzie. Najczęściej (można tu przytoczyć jako przykład wszystkie metale) prąd płynący przez materiał jest proporcjonalny do przyłożonego napięcia, jak to widać na rys. 7a. Spełnione jest wtedy powszechnie znane prawo Ohma. W pewnych materiałach możliwy jest jednak inny przebieg; ze wzrostem napięcia prąd najpierw rośnie, a następnie maleje pomimo dalszego wzrostu napięcia. Gdy napięcie rośnie jeszcze bardziej, prąd ponownie zaczyna rosnąć, jak to przedstawiono na rys. 7b. Obszar, w którym prąd maleje przy wzroście napięcia, nazywamy obszarem ujemnego oporu lub ściślej — różniczkowego oporu ujemnego ($dV/dI<0$).



Rys. 7. Opór dodatni i ujemny. Normalny przewodnik elektryczności ma opór dodatni w tym sensie, że prąd rośnie ze wzrostem napięcia (a); opór $R = \operatorname{tg}\varphi_+ > 0$. W pewnych materiałach możliwa jest sytuacja taka, że ze wzrostem napięcia prąd najpierw rośnie, a potem maleje, aby przy dalszym wzroście napięcia znów zacząć wzrastać (b). Obszar (niebieski), w którym prąd maleje przy rosnącym napięciu zwany jest obszarem różniczkowego oporu ujemnego; w tym obszarze opór $R = \operatorname{tg}\varphi_- < 0$

opór ujemny

Krzywą z rys. 7b udałoby się uzyskać, gdyby przyspieszając elektrony w półprzewodniku przez przyłożenie pola elektrycznego, można było jakimś sposobem wycofać je z procesu przewodnictwa (np. gdy osiągną pewną energię kinetyczną). Taką możliwość daje użycie arsenku galu (GaAs). Płytka z tego



Rys. 8. Zasadniczy element generatora Gunna: płytka monokrystalicznego arsenku galu. Na kryształ naniesione są kontakty omowe, do których zamocowane są elektrody. Napięcie *V* musi być dostatecznie duże (5–15 V), aby spowodować w półprzewodniku powstanie obszaru o różniczkowym oporze ujemnym

materiału stanowi podstawowy element generatora Gunna (rys. 8). Struktura pasmowa arsenku galu (→ Półprzewodniki str. 533) jest następująca: ponad normalnie zajętym pasmem przewodnictwa, w którym elektrony są bardzo ruchliwe, znajduje się inne pasmo, w którym elektrony nie dają się łatwo przyspieszać. (Ściśle rzecz biorąc, mamy do czynienia z jednym pasmem, które ma dwa minima dla różnych wartości wektora falowego $\vec{k}$, przedzielone lokalnym maksimum). W porównaniu z elektronami w pasmie niższym są one właściwie nieruchome. Ich wkład do prądu ($I = e\sum_i n_i v_i$) płynącego przez półprzewodnik jest minimalny — właśnie ze względu na małą prędkość v. W ten sposób, jeśli zwiększymy napięcie przyłożone do płytki arsenku galu, prąd wzrośnie i dziać się tak będzie aż do momentu, w którym energia elektronów osiągnie taką wartość, że będą się one mogły przedostać do pasma wyższego. Przy dalszym wzroście pola elektrony w pasmie podstawowym poruszać się będą coraz prędzej, podczas gdy elektrony pobudzone do wyższego pasma wypadną z procesu przewodzenia prądu. Jeśli liczba elektronów przechodzących do górnego pasma będzie duża, prąd zmaleje pomimo dalszego wzrostu napięcia.

użycie arsenku galu

Oprócz arsenku galu jest jeszcze kilka związków półprzewodnikowych o strukturze pasmowej odpowiedniej do tego, by mógł w nich wystąpić efekt oporu ujemnego. Z badań optycznych wiadomo, że wyższe pasmo, w którym elektrony są bardzo mało ruchliwe, występuje na ogół w odległości ok. $^1/_3$ eV od normalnego pasma przewodnictwa (lub wyżej). Z prostych obliczeń wynika, że pole elektryczne potrzebne do spowodowania znacznego przepływu elektronów do górnego pasma musiałoby być bardzo duże — tak duże, że wskutek strat na ciepło Joule'a mogłoby wystąpić silne przegrzanie, a nawet stopienie materiału. Generatory Gunna działają w sposób stabilny i długotrwały dzięki pewnemu faktowi wynikającemu wprost z zasad elektromagnetyzmu. W krysztale w którym w jakimś miejscu wytwarza się obszar o oporze ujemnym, rozkład pola staje się natychmiast niejednorodny. W szczególności w płytce półprzewodnikowej wytwarza się mały obszar zwany domeną, w którym pole jest dostatecznie silne, aby przenieść elektrony do górnego pasma, podczas gdy w pozostałej części płytki, poza domeną, pole jest małe. Domena nie jest tworem przestrzennie stabilnym. Pchana siłami zewnętrznego pola elektrycznego przemieszcza się przez półprzewodnik od elektrody ujemnej do dodatniej, gdzie znika, a na jej miejsce tworzy się nowa domena, na ogół w pobliżu elektrody ujemnej. Proces ten się powtarza, dopóki do półprzewodnika przyłożone jest odpowiedniej wielkości napięcie.

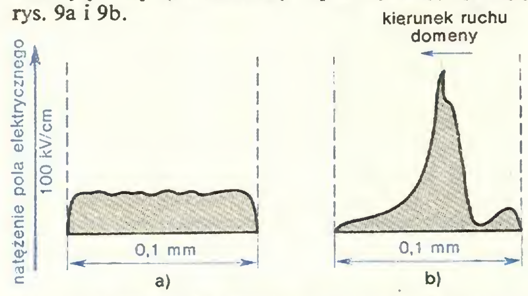
domeny pola elektrycznego

Określenie rozkładu pola elektrycznego wewnątrz płytki półprzewodnikowej jest jednym z głównych problemów teorii oscylatorów Gunna. Przyczyny tworzenia się ruchomych domen można w sposób bardzo uproszczony wyjaśnić następująco: W materiale o dodatnim oporze różniczkowym mamy do czynienia z odpychaniem się ładunków jednoimiennych (elek-

czas relaksacji dielektrycznej

tronów). Każde zaburzenie w postaci zagęszczenia ładunku przestrzennego zanika wykładniczo ze stałą czasu równą czasowi tzw. relaksacji dielektrycznej: $\tau_r = \varepsilon/en\mu$, gdzie ε jest przenikalnością elektryczną, n — koncentracją nośników ładunku, μ — ich ruchliwością.

Tworzenie się w jakimś miejscu półprzewodnika obszaru o różniczkowym oporze ujemnym powoduje zagęszczanie elektronów w tym obszarze. Czas relaksacji dielektrycznej staje się ujemny, ładunek zaś przestrzenny rośnie w czasie z prędkością e^{t/τ_r} Pole elektryczne związane z rosnącym ładunkiem przestrzennym jest skomplikowaną funkcją czasu i położenia wewnątrz półprzewodnika, którą obliczyć można jedynie za pomocą maszyn cyfrowych. Przykład rozkładu pola elektrycznego dla polaryzacji podkrytycznej (różniczkowy opór dodatni) i polaryzacji nadkrytycznej (różniczkowy opór ujemny) podają rys. 9a i 9b.



Rys. 9. Pole elektryczne w arsenku galu przy polaryzacji: a) podkrytycznej, b) nadkrytycznej, gdy zaczyna się tworzyć obszar o oporze ujemnym. Rysunek pokazuje prawdopodobny kształt domeny pola elektrycznego; domena porusza się przez kryształ z prędkością zbliżoną do prędkości unoszenia elektronów. Liniami przerywanymi oznaczono położenie elektrod

analogia z przebiciem elektrycznym

Dla lepszego wyjaśnienia, jak powstaje niejednorodny rozkład pola elektrycznego w półprzewodniku, podać można przykład z dziedziny pokrewnej. W półprzewodniku lub np. w gazie może się pojawić opór ujemny, jeśli przez przyspieszenie elektronów do bardzo wysokiej energii, w czasie zderzeń tych elektronów z atomami wystąpi jonizacja atomów, czyli nagły wzrost liczby elektronów (powielanie). Gdy efekt ten nastąpi, napięcie potrzebne do przepływu prądu określonej wielkości zmaleje. (Efekt ten nazywany jest w różnych kontekstach przebiciem elektrycznym lub przebiciem lawinowym). Charakterystyka prądowo-napięciowa w tym wypadku ma również obszar oporu ujemnego, z tym zastrzeżeniem że tym razem prąd jest wielowartościową funkcją napięcia (pola elektrycznego) — w odróżnieniu od charakterystyki z rys. 7, gdzie napięcie było wielowartościową funkcją prądu.

Podczas jonizacji lawinowej czy przebicia różne obszary półprzewodnika przenoszą różne prądy, tworzą się jakby włókna prądowe. Krańcowym tego przykładem jest wyładowanie elektryczne (iskrzenie) w gazie lub uderzenie pioruna, stanowiące wąskie włókna przewodzące niemal cały prąd, podczas gdy pozostała objętość gazu przewodzi prąd bardzo mały. Iskra elektryczna jest w pewnym sensie domeną prądową w odróżnieniu od domeny pola elektrycznego występującej w oscylatorze Gunna.

Opisana zasada działania generatora Gunna jest dziś powszechnie akceptowana, potwierdziło ją wiele dodatkowych doświadczeń. Oscylacje są mało zależne od zmian temperatury kryształu, zewnętrznego pola magnetycznego czy sposobu zamocowania kontaktów. Oscylator ma w pewnym sensie własną częstość rezonansową, związaną z czasem przejścia domeny przez kryształ, a tym samym — z wymiarami płytki półprzewodnika. W grubych próbkach GaAs obserwowano generację równocześnie na kilku częstościach, co jest niewątpliwie związane z jednoczesnym istnieniem kilku domen w różnych miejscach kryształu. Częstość generacji jest oczywiście zależna od natężenia pola elektrycznego, a więc od napięcia przyłożonego do kryształu. Stwarza to możliwość łatwego przestrajania czy modulacji generatora.

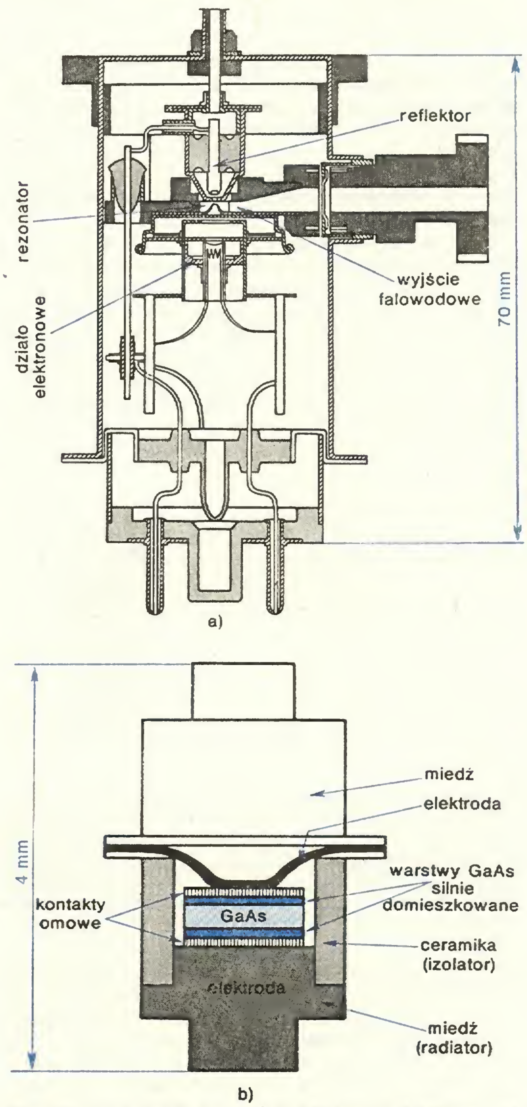
W używanych w praktyce oscylatorach Gunna stosuje się zewnętrzny rezonator mikrofalowy o znacznej dobroci; polepsza on znakomicie czystość spektralną generowanych mikrofal i stwarza możliwość przestrajania mechanicznego. Jest to związane z wektorowym dodawaniem się pól zmiennych o znacznej amplitudzie do stałego pola elektrycznego, a tym samym — z wpływaniem na czas i miejsce powstawania i zanikania domen. Powstawanie i wzrost domen można kontrolować przez odpowiednie kształtowanie rozkładu pola mikrofalowego wewnątrz kryształu. W krańcowym wypadku doprowadzić można w ten sposób do ograniczenia natężenia pola elektrycznego (wysokości domen) do określonej wielkości. Taki rodzaj pracy nazywa się rodzajem LSA (*Limited Space-Charge Accumulation*) i odznacza się znacznie wyższą sprawnością. Tak np. przy częstości rzędu 5 GHz dochodzi ona do 15%, podczas gdy zwykłe generatory Gunna przetwarzają energię ze sprawnością kilku procent.

rodzaj pracy LSA

możliwości generatorów Gunna

Jeśli chodzi o fluktuacje amplitudy i częstości, to już dziś oscylator Gunna konkuruje z powodzeniem z klistronem refleksowym. Produkowane seryjnie generatory Gunna mogą dostarczać mocy rzędu setek miliwatów przy pracy ciągłej i rzędu pojedynczych kilowatów przy pracy impulsowej i częstości kilku GHz. Górna granica częstości generowanych mikrofal jest, jak dotychczas, jeszcze niższa niż klistronów, lecz uzyskano już moc ok. 20 mW na częstości 100 GHz. Moc otrzymywaną z generatorów Gunna ogranicza przede wszystkim konieczność odprowadzania ciepła od elementu półprzewodnikowego. Generator

klistron refleksowy a generator Gunna



Rys. 10. Porównanie konstrukcji klistronu refleksowego na pasmo 8 mm (36 GHz) i generatora Gunna na ten sam zakres częstości. Generator Gunna jest niemal 20-krotnie mniejszy niż klistron refleksowy. Jego małe wymiary powodują duże trudności w odprowadzaniu ciepła. Lepsze odprowadzanie ciepła umożliwia zastosowanie dużych elektrod miedzianych (radiatorów)

Gunna jest tzw. dwójnikiem, tj. ma tylko dwie końcówki. Prostota strukturalna dwójników implikuje prostotę zasilania generatora Gunna (tylko jeden zasilacz niskonapięciowy) i prostotę współpracujących z nim obwodów elektrycznych.

Porównanie konstrukcji generatora Gunna z klistronem obrazuje rys. 10. Należy podkreślić, że mało złożona konstrukcja generatora półprzewodnikowego kryje w sobie wyrafinowaną technologię wytwarzania. Sam arsenek galu musi być niezwykle czysty i jednorodny, z minimalną ilością defektów strukturalnych. Kontakty omowe do próbki kryształu czynnego muszą mieć bardzo mały opór i są naparowywane z reguły na cienką warstwę powierzchniową silnie domieszkowaną tzw. płytkimi domieszkami. Technologia wytwarzania generatorów Gunna może być w znacznym stopniu zautomatyzowana (wzrost monokryształu GaAs z fazy gazowej lub ciekłej, implantacja domieszek w okolicach kontaktów, naparowywanie kontaktów i mocowanie elektrod). Cały proces produkcyjny odbywa się przy zachowaniu warunków niezwykłej czystości.

technologia wytwarzania generatorów Gunna

Wskutek kosztownej technologii wytwarzania, generatory półprzewodnikowe nie są — wbrew przewidywaniom — wielokrotnie tańsze od klistronów. Sytuacja jednak może się radykalnie zmienić wobec ciągle rozszerzającego się kręgu zastosowania mikrofal i potanienia technologii wytwarzania generatorów Gunna w związku z ich masową produkcją. Dalszy rozwój generatorów półprzewodnikowych doprowadzi niewątpliwie w najbliższych latach do rewolucji w technice mikrofalowej podobnej do tej, która się już dokonała w elektronice niższych częstości przez kolejne wprowadzenie tranzystorów i układów scalonych.

B. G. Bosch, R. W. H. Engelmann *Przyrządy półprzewodnikowe z efektem Gunna*, Warszawa 1980; R. Litwin, M. Suski *Technika mikrofalowa*, Warszawa 1972; A. Smoliński *Mikrofalowa elektronika ciała stałego*, Wrocław 1973.

Komputer jako narzędzie pracy fizyka

Wojciech Wójcik

Elektroniczne maszyny matematyczne są jednym z symboli obecnej epoki rozwoju techniki. Rozwój ich był i jest nadal niezwykle szybki, nieomal każdy miesiąc przynosi nowe, ciekawe rozwiązania techniczne i technologiczne w tej dziedzinie.

Elektroniczne maszyny matematyczne możemy podzielić na cyfrowe, analogowe i hybrydowe. Maszyny cyfrowe pozwalają na operacje na liczbach zapisanych w postaci skończonych ciągów cyfr, w odróżnieniu od maszyn analogowych, które operują na wielkościach mających charakter ciągły. Najczęściej przykłada się na wejście maszyny analogowej napięcie elektryczne. Maszyny hybrydowe posiadają cechy obu wymienionych typów maszyn. W dalszej części zajmiemy się wyłącznie omówieniem maszyn cyfrowych, jako maszyn matematycznych najczęściej w fizyce stosowanych. Coraz szerzej przyjmuje się obecnie nazwa „komputer", która oznacza elektroniczną maszynę cyfrową; w takim znaczeniu używać będziemy dalej tego terminu.

komputery

Szybki rozwój informatyki sprawił, że komputer stał się dziś powszechnie stosowanym narzędziem pracy w wielu gałęziach gospodarki narodowej, nauki i techniki. Fizyk wykorzystuje wszystkie najistotniejsze własności komputera, poczynając od jego zdolności rozwiązywania zadań numerycznych, a kończąc na możliwości przetwarzania danych, modelowania czy statystycznej analizy wyników pomiarów. Bardzo często doświadczenie wymaga użycia komputera jako urządzenia sterującego i jednocześnie analizującego wyniki pośrednie. Jest to tzw. technika pracy w czasie rzeczywistym lub na „linii" (*on-line*). Tak szeroki zakres zastosowania komputerów w fizyce stwarza konieczność posługiwania się różnymi typami komputerów, o różnych parametrach technicznych. Skonstruowano całą nową klasę tzw. mini-komputerów, głównie z myślą o obsłudze procesów sterowania w czasie rzeczywistym do zadań wymagających czasochłonnych obliczeń numerycznych stosuje się duże komputery, dysponujące odpowiednio pojemną pamięcią wewnętrzną i umożliwiające wykonywanie 1–2 milionów operacji arytmetycznych w ciągu sekundy. Rozwiązanie układu 100 równań liniowych ze 100 niewiadomymi trwa wówczas kilka sekund. To samo zadanie zajęłoby rachmistrzowi posługującemu się zwykłym arytmometrem kilka lat. Współczesne systemy komputerowe umożliwiają rozwiązywanie zadań, których ze względu na dużą liczbę parametrów nie można rozwiązać w inny sposób. Warto jednak pamiętać, że komputer jest jedynie zautomatyzowanym arytmometrem, umożliwiającym wykonywanie operacji arytmetycznych i logicznych z olbrzymią szybkością. O tym, czy przy użyciu komputera rozwiązane zostanie zadanie obliczeniowe, decyduje wiedza i doświadczenie programisty, który musi opracować procedurę prowadzącą do rozwiązania problemu (tzw. algorytm), a następnie przetłumaczyć ją na ciąg instrukcji (rozkazów) zrozumiałych dla komputera. Umiejętność programowania znacznie ułatwia korzystanie z komputera i przyspiesza uzyskiwanie wyników, nie trzeba bowiem korzystać z pośrednictwa zawodowych programistów. Wprowadzenie do powszechnego użytku tzw. języków algorytmicznych znacznie ułatwiło opanowanie sztuki programowania. Powszechnie używa się również gotowych algorytmów i programów standardowych przygotowanych z myślą o szerokim kręgu użytkowników. O przydatności komputera decydują nie tylko jego parametry techniczne, ale również zestaw standardowych programów (biblioteka programów).

mini-komputery

algorytm

Rozwój techniki obliczeniowej — od liczydła do komputera

Najstarszym i najwcześniej używanym przez ludzkość liczydłem są palce (fakt, że się obecnie posługujemy układem dziesiętnym jest potwierdzeniem tej tezy). Na tej samej zasadzie odliczania oparta jest konstrukcja liczydeł, których wiele odmian przetrwało do dnia dzisiejszego. Istotnym krokiem naprzód w konstrukcji maszyn liczących była maszyna matematyczna (il. 103, tabl. 26), zbudowana w 1652 r. przez B. Pascala. Przy jej konstrukcji Pascal wykorzystał fakt, że system dziesiętny jest systemem pozycyjnym. Podobnie jak wszystkim późniejszym arytmometrom mechanicznym, maszynie Pascala brakło pamięci wewnętrznej dla przechowywania wyników pośrednich. Aby wykonać ciąg działań:

arytmometr mechaniczny

$$2{,}5 \cdot 3{,}55 + 5{,}2 \cdot 3{,}5 + 1{,}5 \cdot 2{,}1 = 30{,}225,$$

nieodzowne było użycie ołówka i papieru dla zapamiętania wyników pośrednich, czyli wyników poszczególnych mnożeń, a dopiero potem można było wykonać dodawanie. Cechą charakterystyczną omawianej techniki obliczeniowej jest duża strata czasu, związana z ręcznym wprowadzaniem liczb do maszyny oraz zapisywaniem wyników pośrednich i końcowych. Prawdziwą rewolucją w dziedzinie automatyzacji obliczeń było wykorzystanie elektroniki do konstrukcji arytmometru. Pierwsze elektroniczne arytmometry

arytmometr elektroniczny

(„Mark I" i „Eniac") zbudowano w Stanach Zjednoczonych w latach 1944–45. Posiadały one urządzenie wczytujące instrukcje przygotowane uprzednio na perforowanych taśmach papierowych. Wykonywanie instrukcji odbywało się sekwencyjnie, w miarę ich wczytywania. Istotnym krokiem naprzód w automatyzacji obliczeń było wprowadzenie przez amerykańskiego matematyka, J. von Neumanna, cyfrowego sposobu przedstawiania rozkazów i umieszczenie ich bezpośrednio w pamięci maszyny. W ten sposób powstało pojęcie programowanej elektronicznej maszyny cyfrowej — komputera. Wzbogacenie zestawu rozkazów przez wprowadzenie instrukcji skoku (bezwarunkowego i warunkowego) stworzyło możliwość korzystania z komputera jako uniwersalnego narzędzia do prowadzenia obliczeń numerycznych. Wprowadzenie operacji logicznych umożliwiło jego zastosowanie do przetwarzania danych.

elektroniczna maszyna cyfrowa

Elementy składowe współczesnego komputera

Na il. 104 (tabl. 26) przedstawiono typowy komputer średniej klasy. Składa się on z wielu elementów, z których najważniejszym jest tzw. jednostka centralna. Zawiera ona arytmometr, pamięć wewnętrzną, układ sterowania, rejestry pamięciowe, adresowe i modyfikacyjne. W arytmometrze wykonywane są operacje arytmetyczne (dodawanie, odejmowanie, mnożenie i dzielenie), operacje logiczne (negacja, alternatywa, koniunkcja) oraz tzw. operacje przesuwania.

jednostka centralna

W pamięci wewnętrznej przechowywane są rozkazy (program) oraz liczby (dane). Pamięć zbudowana jest z układu elementów ferromagnetycznych, z których każdy może się znajdować w jednym z dwu stabilnych stanów namagnesowania. Przypisując stanom tym odpowiednio wartości 0 i 1, można stosować dwójkowy (binarny) zapis liczb. Zaletą tego typu pamięci jest krótki czas odczytu i zapisu informacji (rzędu 1 μs) oraz możliwość dosyć ścisłego upakowania rdzeni ferrytowych. Warto wspomnieć, że obecnie pracuje się intensywnie nad wprowadzeniem do masowej produkcji nowych typów pamięci, np. pamięci półprzewodnikowych lub laserowych. Zaletą pamięci półprzewodnikowych jest możliwość uzyskania krótszych czasów odczytu i zapisu informacji oraz zwiększona odporność na uszkodzenia mechaniczne. Jest to ważne w dziedzinie mini-komputerów, szczególnie — instalowanych w samolotach, satelitach i innych urządzeniach narażonych na wstrząsy. Pamięci laserowe nie wyszły obecnie poza stadium badań laboratoryjnych.

pamięć wewnętrzna

Układ sterowania koordynuje działanie jednostki centralnej przez realizację rozkazów pobieranych z pamięci wewnętrznej komputera. Rejestry pamięciowe służą do przekazywania informacji z pamięci wewnętrznej do arytmometru i z arytmometru do pamięci wewnętrznej, rejestry adresowe i modyfikacyjne służą do wyznaczania adresu informacji zawartej w pamięci wewnętrznej.

Oprócz jednostki centralnej komputer musi być wyposażony w urządzenia peryferyjne i pamięci masowe, czyli zewnętrzne pamięci o dużej pojemności. Są to urządzenia elektroniczno-mechaniczne, umożliwiające wprowadzenie danych, wprowadzenie wyników oraz przechowywanie dużych zbiorów danych. Na il. 104 (tabl. 26) widzimy również konsolę operatorską z monitorem ekranowym, na którym operator obserwuje pracę całego systemu i w razie konieczności wprowadza zmiany i poprawki.

pamięć zewnętrzna

Programowanie, czyli wykorzystanie możliwości komputera

Korzystanie z komputera wymaga przygotowania zestawu danych oraz programu, czyli zbioru instrukcji zrozumiałych dla układu sterowania komputera. Na ogół programy przygotowuje się w jednym z języków algorytmicznych, takich jak fortran, algol, cobol, apl/I. Ułatwiają one posługiwanie się komputerem niezawodowym programistom, niepotrzebna jest bowiem wówczas szczegółowa znajomość budowy i zasady działania konkretnego typu komputera. W językach algorytmicznych zapis operacji arytmetycznych nie różni się praktycznie od zapisu algebraicznego; ułatwia to nie tylko programowanie, ale również publikowanie algorytmów w postaci gotowej do wprowadzenia do komputera, a jednocześnie bardzo przejrzystej i czytelnej dla użytkownika. Program, wprowadzony do komputera w języku algorytmicznym, jest następnie tłumaczony przez specjalny program, zwany translatorem, na ciąg instrukcji podstawowych, czyli na język wewnętrzny komputera. W języku tym każda instrukcja składa się z części operacyjnej i adresowej. W części operacyjnej zawarta jest informacja o tym, jaka operacja i jakiego typu modyfikacje mają być wykonane, natomiast część adresowa określa komórki, na których zawartości ma być dokonana operacja. Formą pośrednią pomiędzy językiem wewnętrznym a językami algorytmicznymi jest tzw. język symboliczny. Pozwala on na wygodny zapis instrukcji podstawowych komputera przez użycie skrótów mnemotechnicznych na oznaczenie nazw operacji oraz symboli alfanumerycznych na określenie adresów. Wyznaczanie wartości adresów dokonywane jest automatycznie w czasie wprowadzania programu do pamięci komputera, przez co unika się wielu niedogodności programowania w języku wewnętrznym, a korzysta się ze wszystkich zalet tego języka.

języki algorytmiczne

translator

język wewnętrzny

język symboliczny

Zastosowanie komputerów w fizyce

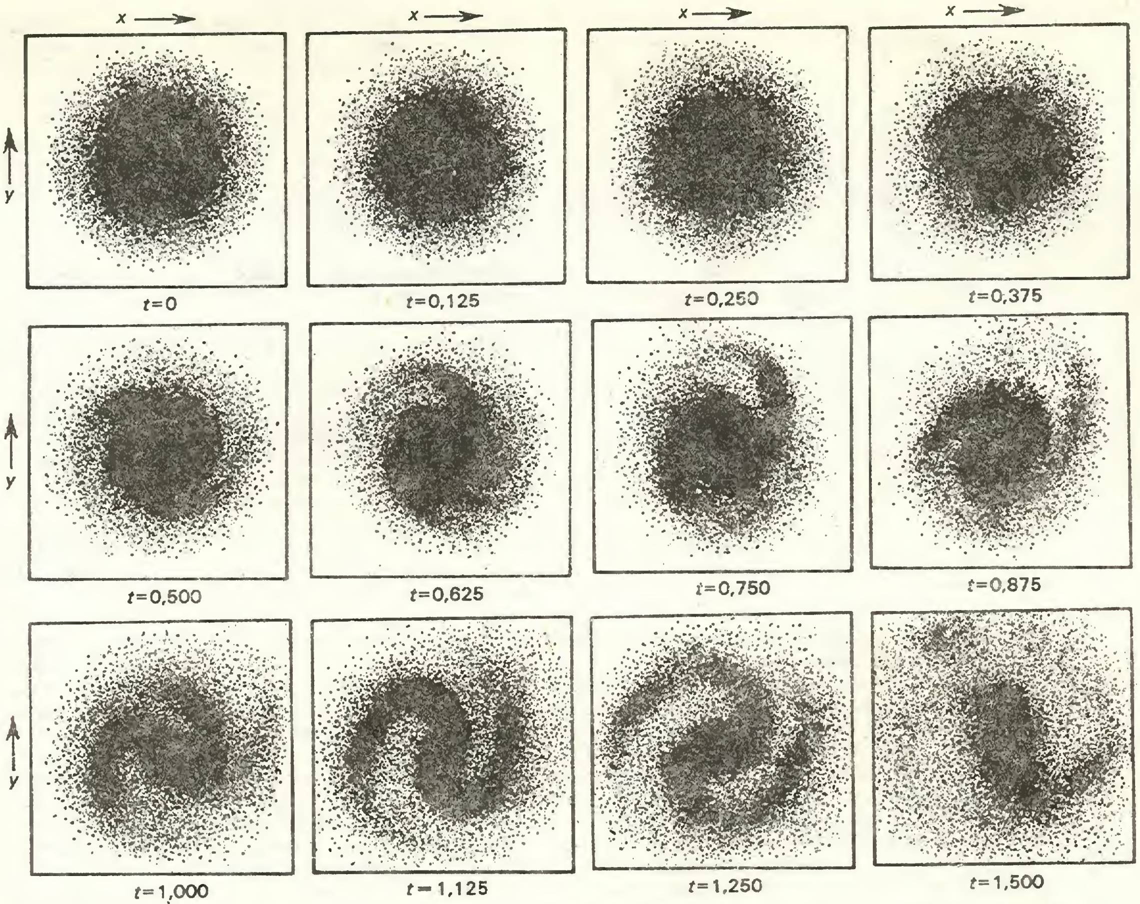
Numeryczne rozwiązywanie równań fizyki umożliwia szybkie uzyskanie odpowiedzi na podstawowe dla fizyka pytanie, czy rozwiązanie zgodne jest z obserwacją. Do rozwiązywania tego typu zagadnień stosuje się klasyczne metody numeryczne, a bardzo często łączy się je z zagadnieniami modelowania i symulacji.

Na rys. 1 przedstawiono przykład modelowania procesu ewolucji galaktyki. Problem polegał na rozwiązaniu równań ruchu zespołu 50 000 gwiazd, rozmieszczonych w postaci wirującego dysku o stałej gęstości gwiazd. Początkowo układ ten znajdował się w równowadze radialnej, lecz z biegiem czasu pojawiała się wyraźna anizotropia układu (rys. 1). Porównanie wyników tego modelowania z obserwacjami galaktyk przedstawiono na il. 106 i 107 (tabl. 26). Przykład ten stanowi doskonałą ilustrację zastosowania komputera do rozwiązania problemu, którego nie można było rozwiązać w inny sposób. Podobne zagadnienia spotykamy w hydrodynamice (rys. 2), aerodynamice czy mechanice ośrodków ciągłych. Rysunek 3 przedstawia modelowanie dyfrakcji fali uderzeniowej i wytwarzanie się stanu ustalonego. Zagadnienie to ma swoje praktyczne zastosowanie m.in. w badaniach z dziedziny astronautyki, pozwala bowiem przewidzieć jak się zachowa kabina statku kosmicznego przy wejściu w górną warstwę atmosfery.

modelowanie

Metoda Monte Carlo (metoda prób losowych) służy do analizy procesów przypadkowych (stochastycznych). Pozwala ona symulować zachowanie się układu zależnego od zespołu parametrów, których wartości podlegają postulowanym funkcjom rozkładu. Metoda Monte Carlo stosowana była od dawna i szczególnie często cytowany jest przykład jej zastosowania do wyznaczania wartości liczby π. Procedura polegała na rzucaniu igły na odpowiednio poliniowany stół. Jeśli odległości pomiędzy kolejnymi liniami równe były długości igły, to stosunek liczby rzutów, przy których igła przecinała którąkolwiek linię, do liczby wszystkich rzutów dążył do wartości π w miarę wzrostu liczby rzutów. Wprowadzenie kom-

metoda Monte Carlo

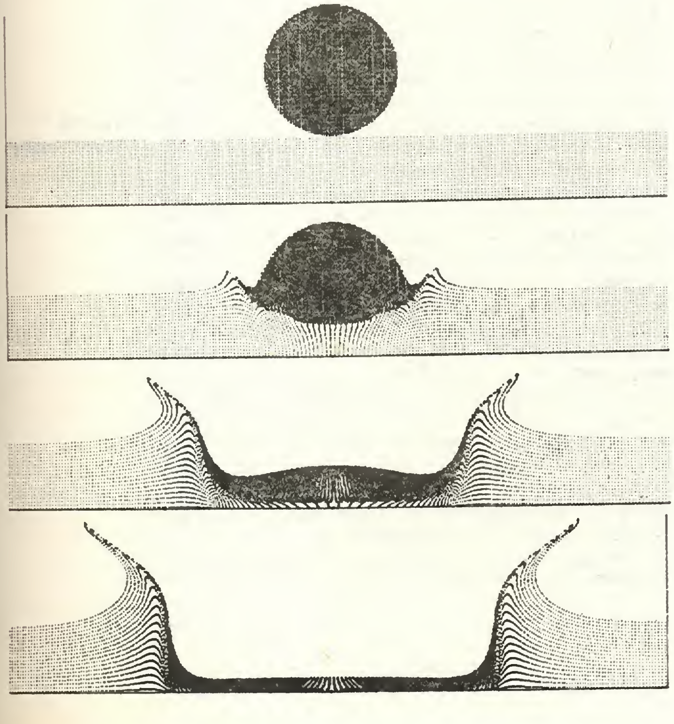


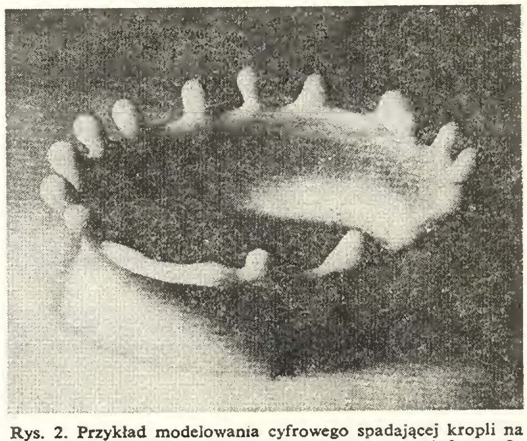
Rys. 1. Modelowanie procesu ewolucji galaktyki. Symulowany układ zawiera 50 000 gwiazd, rozmieszczonych w postaci wirującego dysku o stałej gęstości gwiazd. Jednostką czasu jest okres obrotu

puterów przekształciło metodę Monte Carlo z ciekawego zastosowania praw statystyki matematycznej w potężne narzędzie badań naukowych w wielu dyscyplinach wiedzy. Komputer bowiem pozwolił na zautomatyzowanie procesu otrzymywania liczb losowych oraz umożliwił wykonanie skomplikowanych obliczeń wymaganych przez tę metodę. Olbrzymi wzrost zainteresowania metodą Monte Carlo datuje się od czasu zastosowania jej do badań reakcji rozszczepienia ciężkich jąder. Wymienić w tym miejscu należy prace nad konstrukcją pierwszej bomby atomowej w USA oraz nad konstrukcją pierwszego reaktora jądrowego. Wprowadzone wówczas przez von Neumanna i Ulama metody są obecnie uważane za podstawowe w tej dziedzinie.

Obecnie zakres zastosowania metody Monte Carlo w fizyce jest niezwykle szeroki, od wyznaczania war-

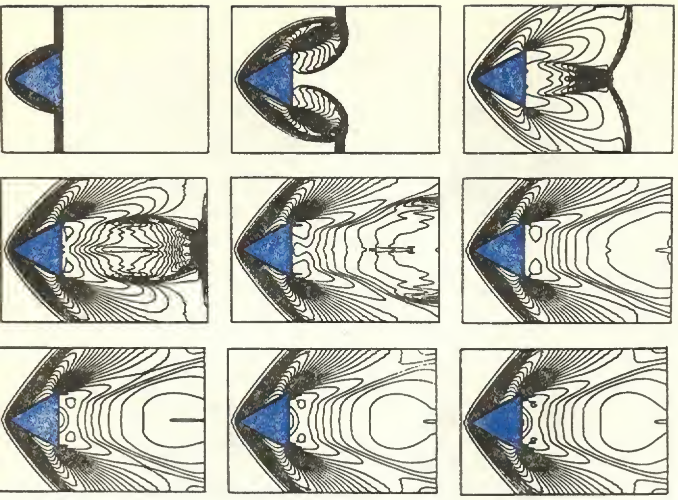
a)





b)

Rys. 2. Przykład modelowania cyfrowego spadającej kropli na powierzchnię cieczy: a) kolejne fazy tego procesu, b) fotografia rzeczywistego procesu

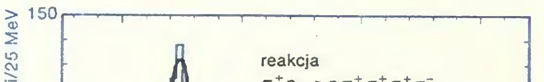


Rys. 3. Modelowanie procesu dyfrakcji fali uderzeniowej i wytworzenia się stanu stacjonarnego przy wejściu stożka o kącie rozwarcia 30° do jednorodnego ośrodka gazowego z prędkością ponaddźwiękową

tości całek wielokrotnych — po symulacyjne metody badania przebiegu procesów fizycznych o charakterze stochastycznym. Dziś nie sposób sobie wyobrazić projektu układu doświadczalnego, którego podstawowe parametry nie byłyby przebadane metodą symulacji rzeczywistych procesów fizycznych. Dzięki temu odpada konieczność przeprowadzania prób i konstruowania modeli i prototypów, a koszty konstrukcji aparatury ulegają znacznej obniżce.

statystyczna analiza danych doświadczalnych

W fizyce doświadczalnej często występuje konieczność przeprowadzania statystycznej analizy wyników doświadczenia. Wyniki pomiarów (kompletny materiał statystyczny) zapisywane są w postaci dużych zbiorów informacji przechowywanych w pamięciach masowych. Analiza polega na obliczaniu dla każdego indywidualnego pomiaru interesującej nas wielkości, a następnie — na sporządzeniu dla niej rozkładu gęstości prawdopodobieństwa. Parametry określające ten rozkład (wartość średnia, dyspersja itp.) związane są na ogół z wielkościami fizycznymi lub poszukiwanymi parametrami. Porównując uzyskane wyniki z przewidywanymi teoretycznymi, możemy weryfikować hipotezy stosując znane testy statystyczne.



Rys. 4. Przykład dopasowania metodą najmniejszych kwadratów krzywej rezonansowej (typu Breita–Wiegnera) do danych doświadczalnych, przy założeniu istnienia tła nierezonansowego

weryfikacja hipotez

Weryfikacja hipotez może polegać również na dobraniu takich wartości parametrów, przy których wyniki teoretyczne najlepiej opisują istniejącą sytuację doświadczalną. Do rozwiązania tego typu problemów stosuje się metody optymalizacji nieliniowej. Rozwiązanie problemu sprowadza się wówczas do znalezienia minimum pewnego funkcjonału. Przykład zagadnienia dopasowania krzywej (hipoteza) do danych doświadczalnych przedstawiono na rys. 4.

Automatyzacja pomiarów i praca w czasie rzeczywistym

Zastosowanie komputerów umożliwiło automatyzację pomiarów parametrów torów cząstek elementarnych ze zdjęć z komór śladowych. Na il. 105 (tabl. 26) przedstawiono automat pomiarowy „Polly", zbudowany w Stanach Zjednoczonych. Proces pomiaru sterowany jest przez komputer i tylko wówczas, gdy komputer nie może przeprowadzić analizy kształtu toru i dokonać obliczenia, żąda on dodatkowych informacji od operatora. Istnieje wiele aparatów pomiarowych, a zasady ich działania są bardzo podobne. Na ogół są to urządzenia prototypowe lub — ze względu na swoje specyficzne zastosowanie — produkowane w małych seriach.

W fizyce jądrowej i fizyce cząstek elementarnych łączy się obecnie coraz częściej technikę licznikową z techniką cyfrową, przy czym komputer wprowadza się jako element sterujący całością doświadczenia. Sprawdza on wówczas działanie wszystkich elementów układu liczników, steruje pracą tego układu, dobierając optymalne parametry, oraz przeprowadza analizę danych na bieżąco, w trakcie trwania doświadczenia. Stosowanie komputerów do sterowania układami licznikowymi stało się tak powszechne, że powstała cała rodzina wyspecjalizowanych komputerów (tzw. mini-komputerów), szczególnie nadających się do tego celu.

Sieci abonenckie

Coraz powszechniej stosuje się obecnie duże zestawy komputerowe, które wraz z końcówkami abonenckimi i siecią łączności tworzą tzw. sieci abonenckie. Końcówki abonenckie służą do wprowadzania programów i danych oraz do drukowania wyników obliczeń. Sieci abonenckie pozwalają najefektywniej wykorzystać zespół komputerów i mogą obsługiwać wielu użytkowników. Stosuje się różnego typu końcówki abonenckie, ale najwygodniejszą formą prowadzenia obliczeń jest korzystanie z systemów interakcyjnych. Pozwalają one na natychmiastową realizację programu (dialog: człowiek–maszyna) i uzyskanie odpowiedzi w postaci wykresu lub tabeli na monitorze ekranowym (tabl. 26, il. 108). Jeśli w programie jest błąd, programista ma możliwość zrobienia poprawki i posłania ponownie programu do wykonania. Na ogół połączenia komputerów z końcówkami abonenckimi dokonywane są przy użyciu łączy telekomunikacyjnych. Istnieją systemy umożliwiające wykorzystanie do tych celów normalnej sieci telefonicznej. Aby się połączyć z komputerem, wystarczy wykręcić odpowiedni numer i po uzyskaniu połączenia rozpocząć pracę w trybie interakcyjnym. Być może w niedalekiej przyszłości spełni się marzenie niektórych fizyków, by przy pomocy miniaturowych końcówek abonenckich można było każdej chwili korzystać z systemów komputerowych.

system interakcyjny

N. P. BUSLENKO i in. *Metoda Monte Carlo*, Warszawa 1967; D. POTTER Metody obliczeniowe fizyki, Warszawa 1977.

AKUSTYKA

Przedmiot i zakres akustyki · Badanie ośrodków za pomocą ultradźwięków · Akustyczne fale powierzchniowe i ich zastosowanie · Holografia akustyczna · Akustyczne zjawiska kwantowe · Modelowanie obiektów akustycznych · Fale uderzeniowe · Hałas

Przedmiot i zakres akustyki

Dziedzina zjawisk fizycznych, którą obejmuje współczesna akustyka jest niewspółmiernie szeroka w porównaniu z tym, co znane było jako dział akustyki klasycznej fizyki w końcu XIX i początku XX w. a ograniczało się do fal sprężystych słyszalnych (gr. *akustikos* — dotyczący słuchu). Obecnie przedmiotem akustyki są wszystkie zjawiska związane z rozchodzeniem się fal sprężystych w różnych ośrodkach, we wszystkich stanach skupienia materii, w pełnym zakresie częstości drgań możliwych w przyrodzie. Zakres ten obejmuje infradźwięki (poniżej słyszalności od ułamków herca do 16 Hz), dźwięki (słyszalne — od 16 Hz do 16 kHz), ultradźwięki (powyżej słyszalności — od 16 kHz–10^9 Hz) oraz hiperdźwięki (powyżej 10^9 Hz aż do częstości granicznej wyznaczonej przez nieciągłą strukturę atomową materii rzędu 10^{13}–10^{14} Hz).

Opis zjawisk akustycznych w tak szerokim zakresie częstości nie może być jednolity; mimo tej samej natury fal sprężystych w całym zakresie, różnią się one jednak długością fali. Przy opisie wielu procesów akustycznych, ośrodek, w którym rozchodzą się fale sprężyste może być traktowany jako ciągły. Opis taki nazywa się klasycznym. W wielu jednak wypadkach założenie ciągłości ośrodka nie może być spełnione, np. wtedy, gdy własności molekularne (nieciągłe) ośrodka mają istotne znaczenie (→ Akustyczne zjawiska kwantowe).

Klasyczny opis wielu zagadnień akustyki można jedynie traktować jako dostateczne przybliżenie, szczególnie w zakresie dźwięków słyszalnych, jest on jednak niewystarczający dla ultradźwięków o dużych częstościach i hiperdźwięków, gdzie musi być brana pod uwagę nieciągła struktura ośrodka (→ Wzbudzenia elementarne w ciałach stałych). Mówiąc dokładniej, taki przybliżony opis jest niewystarczający wówczas, gdy droga swobodna drobin staje się porównywalna z długością fali (np. w bardzo rozrzedzonych gazach), albo gdy odstępy międzymolekularne czy międzyatomowe stają się porównywalne z długością fali, jak np. przy rozchodzeniu się hiperdźwięków w ciałach stałych (→ Akustyczne zjawiska kwantowe).

W akustyce klasycznej opartej na teorii ośrodków ciągłych wiele zagadnień można rozwiązać za pomocą liniowych równań różniczkowych o pochodnych cząstkowych, niektóre jednak problemy wymagają bardziej skomplikowanego opisu matematycznego. Ścisłe równania mechaniki ośrodków ciągłych (w tym akustyki) są nieliniowe a ich rozwiązanie przedstawia tak duże trudności matematyczne, że mimo wielu wysiłków nie udało się, jak dotąd, uzyskać pełnych wyników. Na ogół trzeba zadowolić się teoriami przybliżonymi. Z tego powodu rozwój współczesnych zagadnień akustyki możliwy jest w ścisłym związku z doświadczeniem, które w ostatnich latach, szczególnie w odniesieniu do akustycznych zjawisk o dużej energii (jak np. fale uderzeniowe) pozwala na konfrontację, uzasadnienie i uzupełnienie wyników tych przybliżonych obliczeń.

W wielu współczesnych laboratoriach fizycznych na świecie prowadzi się intensywne badania akustycznych zjawisk o dużej energii (zjawisk nieliniowych) i wyznacza się wielkości charakteryzujące nieliniowe własności ośrodków gazowych, ciekłych i stałych.

Warto zauważyć, że wiele rozwiązań teoretycznych, szczególnie w zakresie nieliniowym, wykorzystywanych od lat na potrzeby akustyki, przenosi się ostatnio do zagadnień optyki nieliniowej (→ Optyka nieliniowa), które pojawiły się w ostatnim 10-leciu dzięki wynalezieniu źródeł światła o bardzo dużej (gigantycznej) mocy i użyciu ich do badania materii.

Relatywistyczne ujęcie teorii ośrodków ciągłych znajduje zastosowanie w zagadnieniach astrofizycznych, np. w badaniach atmosfer gwiazdowych, w szczególności Słońca, i wiąże się z zagadnieniem promieniowania elektromagnetycznego tych obiektów. W warunkach ziemskich ujęcie takie stosuje się do opisu niektórych zagadnień plazmy.

Teoria ośrodków ciągłych w zastosowaniu do hydrodynamiki tzw. „dwuprędkościowej" (pierwsza i druga prędkość dźwięku) daje makroskopowy opis ruchu cieczy nadciekłej, jaką jest np. ciekły hel w temperaturach bliskich 0 K (→ Nadpłynność).

Przy rozpatrywaniu fal sprężystych rozróżnia się ośrodek idealny (bezstratny czyli niedysypatywny), w którym fale sprężyste nie ulegają tłumieniu, oraz ośrodek rzeczywisty (stratny czyli dysypatywny), w którym fale zanikają (ulegają tłumieniu) w miarę rozchodzenia się. Ośrodki rzeczywiste można traktować jako idealne (szczególnie w wypadku fal sprężystych o małych częstościach) z dobrym dla praktycznych celów przybliżeniem (np. dźwięki słyszalne w powietrzu na odległościach kilkunastu metrów).

Teoria ośrodków ciągłych i fale sprężyste

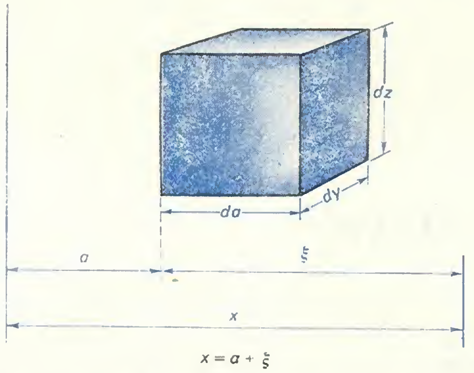
Antoni Śliwiński

Tradycyjnie do teorii ośrodków ciągłych zalicza się mechanikę płynów (cieczy i gazów) czyli hydrodynamikę, teorię sprężystości i plastyczności ciał stałych

oraz inne dyscypliny szczegółowe przyjmujące za punkt wyjścia założenie, że mikroskopowe nieciągłości materii nie mają wpływu na rozpatrywane zjawiska fizyczne w skali makroskopowej. Do takich szczegółowych dyscyplin należy również tzw. akustyka klasyczna, która w sposób fenomenologiczny opisuje mechanizm rozchodzenia się fal sprężystych w ośrodkach ciągłych. U podstaw tej teorii leży traktowanie materii jako środowiska ciągłego bez wnikania w jej rzeczywistą budowę nieciągłą (atomy, cząsteczki).

cząstka ośrodka

Wyrazem ciągłości ośrodka jest to, że możemy go w myśli podzielić na dowolnie małe elementy objętości o takich samych własnościach, jak cała rozważana jego objętość. Taki mały element objętości ośrodka stanowi elementarny obiekt podlegający prawom mechaniki i termodynamiki, musi więc być na tyle duży, aby mieściło się w nim wiele molekuł czy atomów, dla których wewnątrz niego spełnione byłyby prawa statystyczne (→ Fizyka statystyczna), aby słuszne były prawa termodynamiki, oraz na tyle mały, aby można go potraktować jako nieskończenie mały z punktu widzenia opisywanych zjawisk. W rozważaniach zjawisk fizycznych taki element objętości nazywamy cząstką ośrodka (rys. 1), a przy rozpatrywaniu fal sprężystych — cząstką akustyczną. Analiza ruchu cząstki ośrodka pod wpływem działających na nią sił pozwala opisać zjawiska odkształceń dynamicznych, w szczególności przepływów ośrodków (płynnych) i rozchodzenia się fal sprężystych.

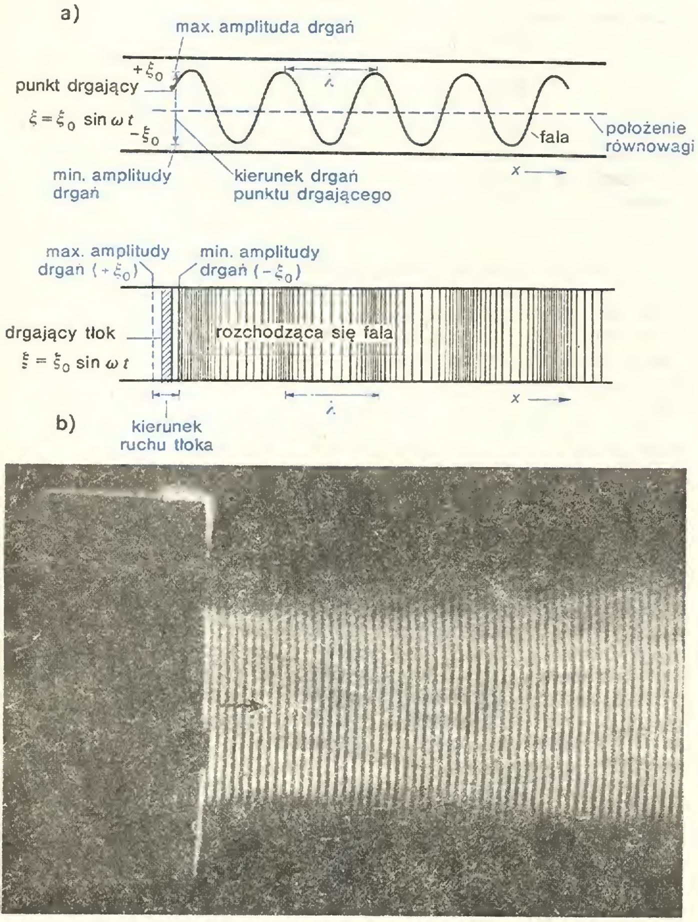


Rys. 1. Taki element objętości nazywamy cząstką ośrodka lub cząstką akustyczną. Jest to jej położenie równowagi. Przy zaburzeniu cząstka przesuwa się w miejsce $x = a+\xi$

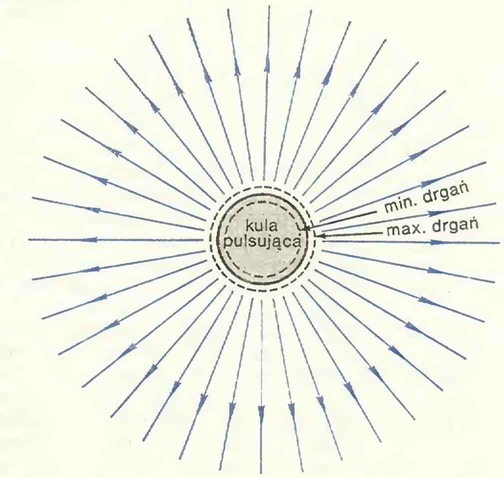
Istotą ruchu falowego jest zależność czasowo-przestrzenna przemieszczania się zaburzenia przez ośrodek. Zaburzenie rozchodzi się tak, że każda z cząsteczek ośrodka wykonuje ruch dookoła położenia równowagi, przekazując energię sąsiednim cząstkom zajmującym inne położenia w przestrzeni i drgającym w odpowiednio przesuniętych w czasie chwilach następnych. Stany ruchu powtarzają się okresowo w przestrzeni (co długość fali λ) oraz w czasie (co okres T) — rys. 2, przy czym $\lambda = cT$, gdzie c — prędkość rozchodzenia się zaburzenia czyli prędkość fazowa.

Źródłem fal sprężystych w ośrodkach ciągłych są mechaniczne układy drgające umieszczone w nich. Energia drgań tych układów zostaje zamieniona na energię fali akustycznej. Przykładem drgającego układu generującego falę sprężystą może być wspomniana już wyżej kula pulsująca (rys. 3) otoczona ciągłym ośrodkiem sprężystym. Układ taki jest równoważny z układem hipotetycznym, tzw. źródłem punktowym, które można sobie wyobrazić jako granicznie małą kulę pulsującą dla $r \to 0$, gdzie r — promień kuli.

kula pulsująca



Rys. 2. Sprężysta fala podłużna: a) wykres zmian ciśnienia fali rozchodzącej się w rurze, b) schematyczny rozkład ciśnienia zgęszczeń i rozrzedzeń lub sprężeń i rozprężeń ośrodka w rurze, c) fotografia cieniowa fali ultradźwiękowej w dużej objętości cieczy



Rys. 3. Poglądowe przedstawienie kuli pulsującej jako źródła kulistej fali sprężystej

Drgania kuli pulsującej odbywają się w ten sposób, że jej promień zmienia się okresowo z częstością ν, przy czym środek kuli pozostaje nieruchomy. Poglądowo można sobie wyobrazić kulę pulsującą jako gumowy pęcherzyk napełniony powietrzem, który na skutek np. periodycznych wahań temperatury w pierwszym półokresie rozszerza się, równomiernie pęcznieje, a w drugim półokresie kurczy się.

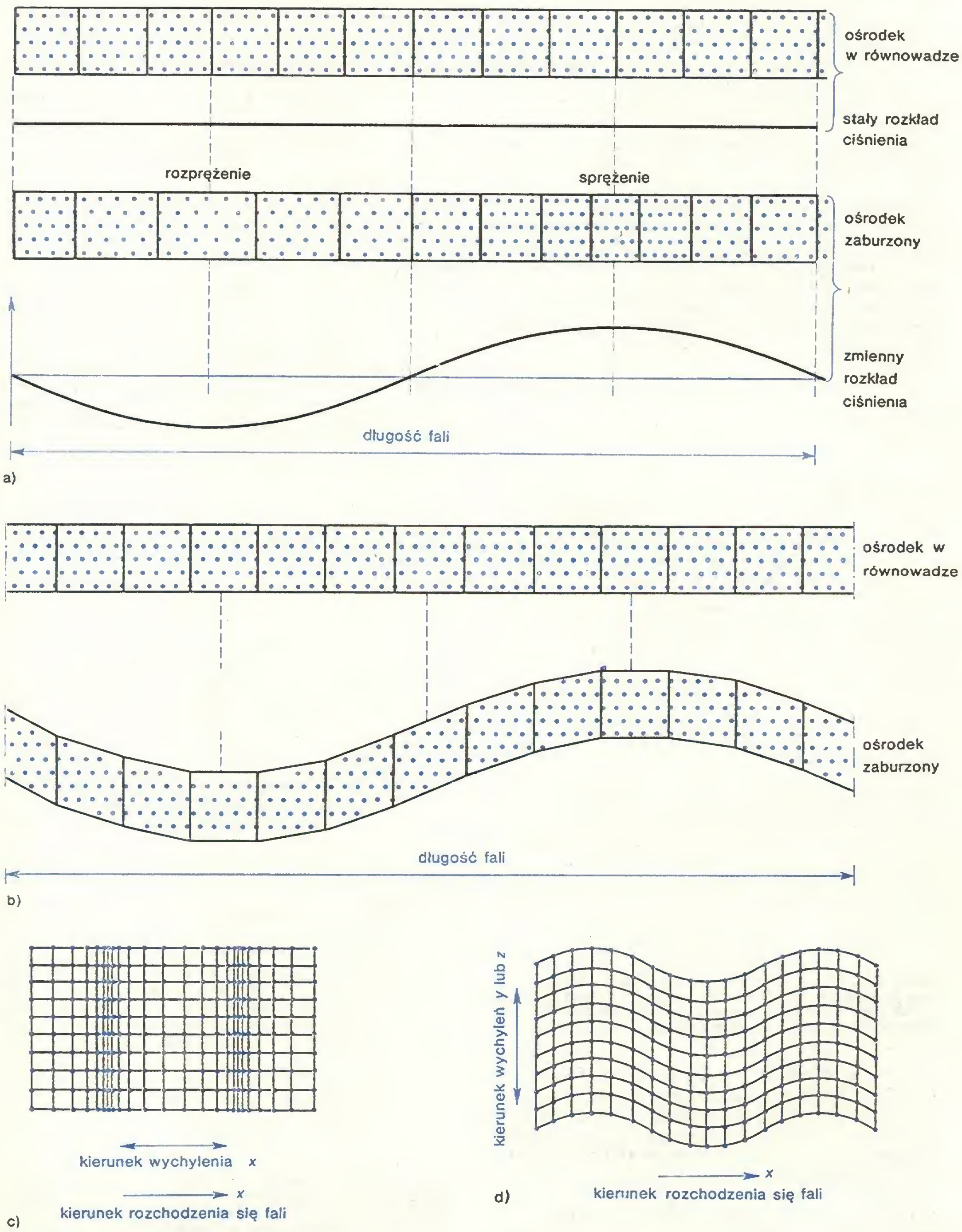
Fale sprężyste w ośrodku opisuje równanie falowe, które wyprowadza się z podstawowych równań teorii ośrodków ciągłych. Są to: równania ruchu (we współrzędnych Eulera lub Lagrange'a) dla cząstki akustycznej pobudzonej do drgań przez siły wywołane zaburzeniem ośrodka, równanie ciągłości ośrodka

(lub inaczej równanie zachowania masy) oraz równanie stanu termodynamicznego (najczęściej równanie adiabaty).

Najogólniejszą formę tych równań wyprowadza się dla ośrodka stałego anizotropowego, przyjmując za podstawę ogólne prawo Hooke'a, które wyraża liniową zależność odkształceń od naprężeń. W ciałach stałych anizotropowych własności sprężyste zależą od kierunku. Jest to charakterystyczne szczególnie dla monokryształów, w których decydującą rolę odgrywa stopień regularności sieci krystalicznej. Im bardziej uboga symetria danego kryształu, tym więcej wyróżnionych kierunków, dla których własności sprężyste (stałe sprężyste, a co za tym idzie i prędkości dźwięku) są różne. Własności sprężyste izotropowych ciał stałych są jednakowe dla różnych kierunków. Ciała izotropowe mogą być albo amorficzne, których struktura sieci przestrzennej jest przypadkowa — podobnie jak w cieczy, albo polikrystaliczne, w których istnieje duża ilość przypadkowo zorientowanych ziaren krystalicznych, czyli krystalitów. W teorii ośrodków ciągłych zakładamy, że rozmiary tych krystalitów, jak również komórek elementarnych w monokryształach są małe w porównaniu z cząstką akustyczną.

rozchodzenie się fal sprężystych

W najprostszym przypadku zagadnienie rozchodzenia się fal sprężystych w ośrodku rozpatruje się jednowymiarowo zakładając, że czoło fali jest nieograniczone. Zaburzenie ośrodka przedstawia wtedy falę



Rys. 4. Poglądowe przedstawienie przesunięcia cząstek akustycznych (a, b) oraz atomów w nieograniczonym ośrodku stałym (c, d) w przypadku fali podłużnej (a, c) i fali poprzecznej (b, d) o odległość równą jednej długości fali

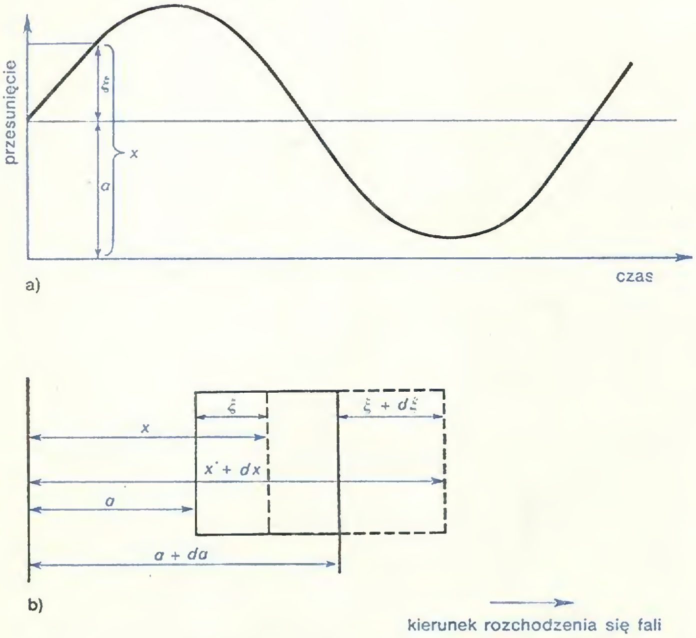
płaską. Czoło fali przesuwa się w jednym kierunku, np. x (rys. 4) tak, że dla każdej wartości współrzędnej x wszystkie ruchy cząstek akustycznych względem pozostałych współrzędnych y i z są te same. W nieograniczonej przestrzeni trójwymiarowej rozpatruje się jako najprostsze fale kuliste (czołem fali jest powierzchnia kulista) (rys. 3).

W czasie rozchodzenia się fali sprężystej cząstki ośrodka wykonują ruch drgający względem swych położeń równowagi. Jeżeli ruch ten odbywa się w tym samym kierunku, w jakim rozchodzi się fala, to falę nazywamy podłużną (rys. 4a), a jeśli odbywa się prostopadle do kierunku fali, to nazywamy ją poprzeczną (rys. 4b). Fale sprężyste w płynach są najczęściej falami podłużnymi. Fale poprzeczne występują głównie w ciałach stałych, choć mogą istnieć w bardzo lepkich cieczach (są jednak silnie tłumione).

fale podłużne

współrzędne Lagrange'a i Eulera

Do opisu ruchu fali podłużnej rozpatrywanej w ośrodku płynnym można podejść dwojako. Albo rozpatrujemy przesunięcie określonej cząstki z jej położenia równowagi badając dalej jej prędkość ı przyspieszenie, albo obserwujemy własności ośrodka w wybranym punkcie określając przesunięcie, prędkość i przyspieszenie płynu w tym punkcie, bez względu na to, które cząstki przechodzą przez punkt rozpatrywany w różnych czasach obserwacji. W zależności od tego którą koncepcję przyjmiemy za punkt wyjścia, w pierwszym przypadku mamy do czynienia ze współrzędnymi Lagrange'a lub tzw. współrzędnymi materialnymi (współrzędne związane z daną cząstką), w drugim przypadku (rys. 5) ze współrzędnymi Eulera lub tzw. przestrzennymi (współrzędne związane z punktem przestrzeni).



Rys. 5. Porównanie współrzędnych Lagrange'a i Eulera: a) wykres przesunięcia cząstki ośrodka we współrzędnych Lagrange'a a oraz współrzędnych Eulera x, b) jednowymiarowa deformacja elementu objętości we współrzędnych Lagrange'a i Eulera

Różnica wynikająca z tych dwóch sformułowań odgrywa istotną rolę dopiero przy bardzo dużych natężeniach fal (np. fal o amplitudzie skończonej omawianych w akustyce nieliniowej) nie ma natomiast znaczenia przy małych natężeniach (fal o małej amplitudzie omawianych w akustyce liniowej).

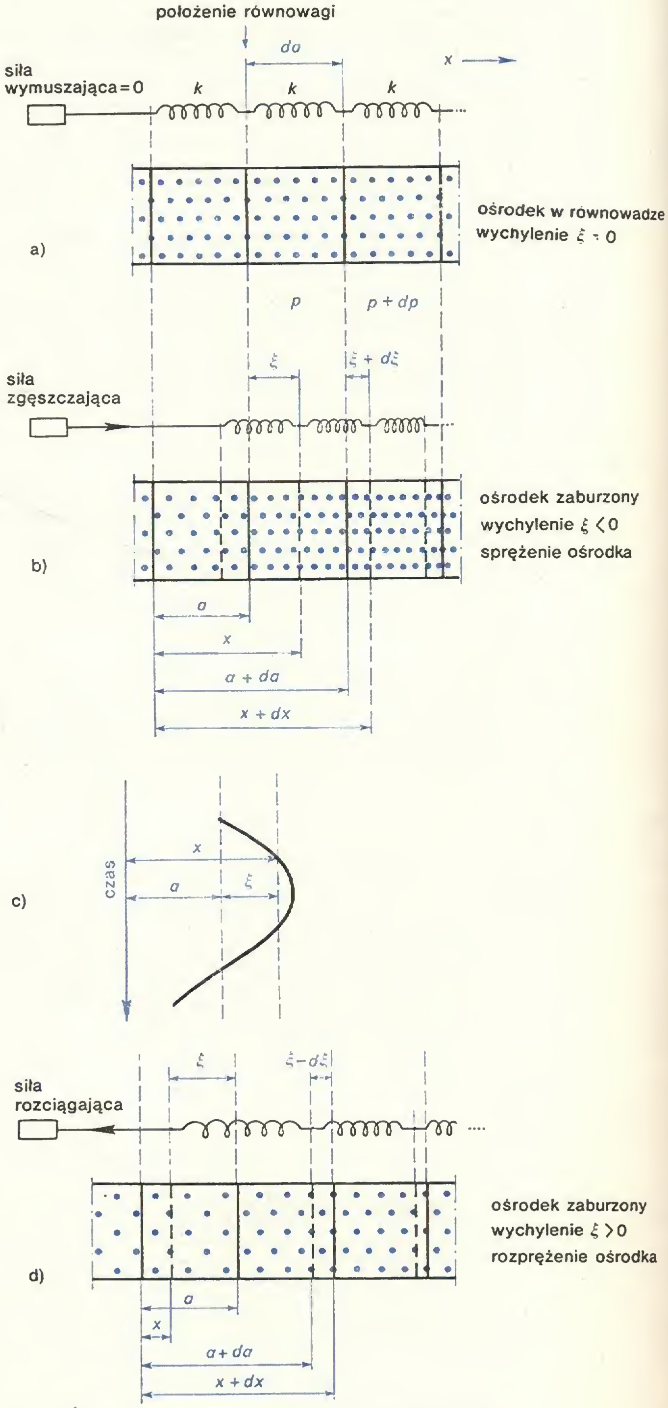
Rozważmy element ośrodka (cząstkę akustyczną) w spoczynku — położenie równowagi określa współrzędna a (rys. 1, 5, 6). Gdy pojawi się w ośrodku zaburzenie (fala sprężysta), cząstka zmieni swoje położenie na x, przy czym $x = a+\xi$, gdzie ξ oznacza przesunięcie cząstki.

Prędkość cząstki we współrzędnych Lagrange'a $u^L(a, t)$ możemy więc wyrazić następująco:

$$u^L(a, t) = \frac{\partial x}{\partial t} = \frac{\partial \xi}{\partial t},$$

a z kolei przyspieszenie jako

$$\frac{\partial u^L}{\partial t} = \frac{\partial^2 \xi}{\partial t^2}.$$



Rys. 6. Schemat przesunięć cząstek akustycznych w fali podłużnej i przedstawienie tego za pomocą modelu składającego się z jednakowych mas związanych jednakowymi sprężynkami o współczynnikach sprężystości k: a) w równowadze, b) w fazie zgęszczenia, c) wykres zmian przesunięcia w czasie, d) w fazie rozrzedzenia

Przyjmując, że ξ jest wielkością małą, dowolną współrzędną Lagrange'a q^L można rozwinać w szereg i wyrazić przez współrzędne Eulera q^E

$$q^L(a,t) = q^E(x,t)\Big|_{x=a+\xi(x,t)} =$$

$$= q^E(x,t)\Big|_{x=a} + \frac{\partial q^E(x,t)}{\partial x}\Big|_{x=a} \xi(x,t)+\ldots$$

W szeregu tym napisaliśmy tylko dwa pierwsze wyrazy (kreska z podaną u dołu wartością x oznacza, że chodzi o wartość q^E przy x równym podanej wartości).

Podobnie można wyrazić współrzędne Eulera q^E przez współrzędne Lagrange'a q^L:

$$q^E(x,t) = q^L(a,t)\Big|_{a=x-\xi(a,t)} =$$

$$= q^L(a,t)\Big|_{a=x} - \frac{\partial q^L(a,t)}{\partial a}\Big|_{a=x} \xi(a,t)+\ldots$$

Widzimy, że przesunięcie ξ występuje w obydwu układach współrzędnych. We współrzędnych Lagrange'a jest ono rozumiane jako przesunięcie cząstki początkowo umieszczonej w a i dlatego jest funkcją a i t. We współrzędnych Eulera ξ jest przesunięciem chwilowym dowolnej cząstki umieszczonej w miejscu x; tutaj ξ jest funkcją x i t.

równanie ciągłości

Rozważmy sytuację dynamiczną. Rys. 6a przedstawia element objętości ośrodka $da\,dy\,dz$ w spoczynku. Załóżmy, że fala płaska przesuwa się w prawo poprzez ośrodek tak, że w danej chwili cząstki, znajdujące się początkowo w spoczynku w miejscu a przesuwają się na odległość ξ, a cząstki znajdujące się początkowo w spoczynku w miejscu $a+da$ przesuwają się o odległość $\xi+d\xi$. Ponieważ nowe położenia ścianek elementu ośrodka mogą być zapisane jako x^L i x^L+dx^L, więc objętość elementu dV można wyrazić jako $dV = dx^L dy\, dz$ (ponieważ nie ma ruchu w kierunku y i z, to nie ma potrzeby rozróżnienia dla tych kierunków współrzędnych Lagrange'a od Eulera). Całkowita masa płynu w objętości dV jest ta sama w obydwu wypadkach, czyli

$$\varrho_0\, da\, dy\, dz = \varrho^L dx^L dy\, dz,$$

gdzie ϱ_0 oznacza gęstość ośrodka w spoczynku, a ϱ^L — gęstość płynu przesuniętego. Stąd wynika równanie ciągłości w postaci następującej:

$$\varrho^L = \varrho_0 \frac{da}{dx^L}$$

lub — uwzględniając związek $x^L = a+\xi$ —

$$\varrho^L = \varrho_0\left(1+\frac{\partial\xi}{\partial a}\right)^{-1}. \qquad (1)$$

równanie ruchu

Sformułowanie Lagrange'a pozwala bardzo łatwo napisać równanie ruchu. Jeśli po lewej stronie elementu działa ciśnienie p^L, a po prawej stronie $p^L+(\partial p^L/\partial x^L)dx^L$, wówczas siła wypadkowa działająca na cząstkę akustyczną (element płynu) wynosi $(-\partial p^L/\partial x^L)\,dx^L dy\, dz$ (jest skierowana w prawo) i równanie ruchu ma postać:

$$-\frac{\partial p^L}{\partial x^L} dx^L dy\,dz = \varrho_0 da\,dy\,dz\frac{\partial^2\xi}{\partial t^2}$$

$$\text{lub } -\frac{\partial p^L}{\partial x^L}\cdot\frac{\partial x^L}{\partial a} = -\frac{\partial p^L}{\partial a} = \varrho_0\frac{\partial^2\xi}{\partial t^2}, \qquad (2)$$

gdzie $\varrho_0\, da\, dy\, dz$ jest masą cząstki akustycznej.

Z doświadczenia wynika, że proces zmian ciśnienia i gęstości w fali akustycznej jest z punktu widzenia termodynamiki procesem adiabatycznym, jak to wykazał P.S. Laplace w 1816 r.; do tego czasu, począwszy od I. Newtona (1687 r.), uważano falę dźwiękową za proces izotermiczny. Prędkość dźwięku dla procesu adiabatycznego $c^2 = (\partial p^L/\partial\varrho^L)_{ad}$.

Ponieważ dla gazu równanie adiabaty ma postać $p = p_0(\varrho/\varrho_0)^\gamma$, gdzie $\gamma = C_p/C_V$, C_V — ciepło właściwe przy stałej objętości, C_p — ciepło właściwe przy stałym ciśnieniu, więc

$$c^2 = \left(\frac{\partial p^L}{\partial\varrho^L}\right)_{ad} = \frac{\gamma p_0}{\varrho_0}\left(\frac{\varrho}{\varrho_0}\right)^{\gamma-1} = \frac{\gamma p_0}{\varrho_0}\frac{1}{\left(1+\frac{\partial\xi}{\partial a}\right)^{\gamma-1}} =$$

$$= c_0^2\frac{1}{\left(1+\frac{\partial\xi}{\partial a}\right)^{\gamma-1}}.$$

Równanie ruchu przekształcimy następująco:

$$\frac{\partial^2\xi}{\partial t^2} = -\left(\frac{1}{\varrho_0}\right)\frac{\partial p}{\partial a} = -\left(\frac{1}{\varrho_0}\right)\frac{\partial p^L}{\partial\varrho^L}\cdot\frac{\partial\varrho^L}{\partial a} =$$

$$= \frac{c^2}{\left(1+\frac{\partial\xi}{\partial a}\right)^2}\frac{\partial^2\xi}{\partial a^2} = \frac{c_0^2}{\left(1+\frac{\partial\xi}{\partial a}\right)^{\gamma+1}}\frac{\partial^2\xi}{\partial a^2}, \qquad (3)$$

gdzie $c_0 = \sqrt{\gamma p_0/\varrho_0}$ — prędkość rozchodzenia się fali o małej amplitudzie (w przypadku liniowym).

przypadek $|\partial\xi/\partial a| \ll 1$

Jeżeli w mianowniku pominiemy pochodną $\partial\xi/\partial a$ jako małą w porównaniu z 1, np. w powietrzu przy głośnej rozmowie $\partial\xi/\partial a$ jest rzędu 10^{-3}, to równanie (3) staje się zwykłym równaniem falowym

$$\frac{\partial^2\xi}{\partial t^2} = c_0^2\frac{\partial^2\xi}{\partial a^2}. \qquad (4)$$

Rozwiązaniem tego równania jest funkcja $\xi = \xi_0\sin(\omega t-ka)$, gdzie ω — częstość kołowa, k — stała propagacji, przy czym $\omega = 2\pi\nu$ oraz $k = 2\pi/\lambda$, gdzie ν — częstość drgań fali, λ — długość fali.

Wykresem tego wzoru na wychylenie jest sinusoida przedstawiona na rys. 6d oraz 2a.

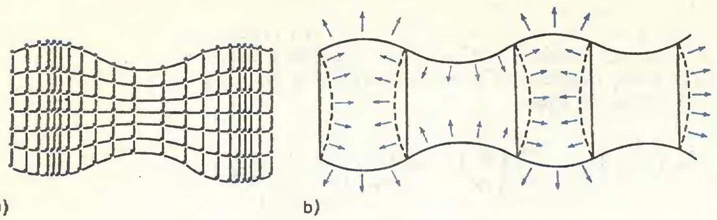
Wyraz $\partial\xi/\partial a$ można pominąć, gdy $|\partial\xi/\partial a| \ll 1$, co oznacza że $u_0/c_0 \ll 1$, gdzie $u_0 = \omega\xi_0$ — amplituda prędkości cząstki akustycznej i to jest równoważne warunkowi $|\xi| \ll \lambda$ (por. akustyczne zjawiska liniowe). Te ostatnie związki wynikają z przeliczeń, jeśli $\xi = \xi_0\sin(\omega t-ka)$.

Jeżeli ten rozważany warunek jest spełniony, to nie ma potrzeby odróżniać współrzędnych Lagrange'a od współrzędnych Eulera i można zmienną a zastąpić przez x. W zakresie tego przybliżenia zjawiska akustyczne są liniowe, a wówczas gdy wyrazu $\partial\xi/\partial a$ nie można pominąć w porównaniu z 1, zjawiska przebiegają nieliniowo.

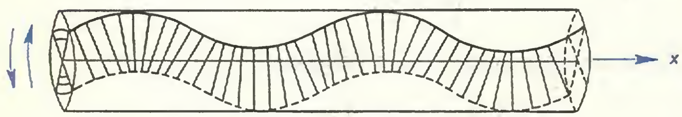
wpływ ograniczenia ośrodka

Zjawiska falowe w ośrodkach ciągłych ograniczonych przebiegają inaczej niż w ośrodkach nieograniczonych, zwłaszcza gdy rozmiary ograniczonego ośrodka stają się porównywalne z długością fali. Szczególnie wyraźnie występuje to w odniesieniu do ciał stałych, kiedy istnienie ograniczenia prowadzi do nowych typów fal sprężystych. Na przykład w ośrodku, którym jest długi pręt, powstają fale dylatacyjne (rys. 7), fale skrętne (rys. 8) i fale giętne (rys. 9). Innym przykładem fal w ośrodku ograniczonym są fale powierzchniowe (→ Akustyczne fale powierzchniowe i ich zastosowania), z których najbardziej znanymi są tzw. fale Rayleigha (na powierzchni granicznej ośrodka stałego lub ciekłego i powietrza). Fale Stonleya (na powierzchni rozdzielającej dwa ośrodki stałe), fale Lamba zwane również płytowymi (powstają w blachach, niegrubych płytach, powłokach itp. o grubościach porównywalnych z długością fali). Fale powierzchniowe, szczególnie w zakresie mikrofalowym, w ostatnich latach często są stosowane w takich urzą-

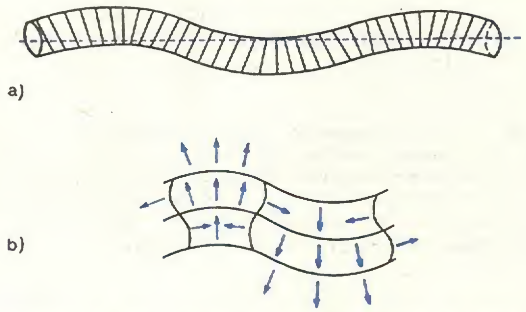
dzeniach, jak ultradźwiękowe linie opóźniające, deflektory i modulatory optyczne (→ Akustyczne zjawiska kwantowe).



Rys. 7. Fale dylatacyjne w pręcie: a) rozkład chwilowy deformacji, b) schemat rozkładu naprężeń



Rys. 8. Fale skrętne w pręcie



Rys. 9. Fale giętne w pręcie: a) chwilowy rozkład deformacji b) schematyczny rozkład naprężeń

Ograniczenia ośrodków powodują poza tym powstawanie odbić lub załamań fal sprężystych na tych granicach. W płynach zjawiska te przebiegają w sposób dość prosty, w ciałach stałych są bardziej złożone.

Ilustracja 151 (tabl. 41) przedstawia fotografię fali ultradźwiękowej ulegającej załamaniu przy przejściu z jednego ośrodka ciekłego do drugiego, il. 154 (tabl. 41) natomiast falę ulegającą załamaniu i odbiciu przy przejściu przez płytkę mosiężną zanurzoną w wodzie. W pierwszym przypadku mamy pojedynczą falę padającą i pojedynczą falę załamaną. W drugim przypadku podłużna fala padająca na płytkę pod wybranym kątem wzbudza w niej falę Lamba, która z kolei wypromieniowuje falę wtórną na swej drodze.

dyfrakcja na przeszkodach

Wystąpienie w ośrodku ciągłym obszarów nieciągłych porównywalnych z długością fali prowadzi do zjawiska dyfrakcji (ugięcia) fal sprężystych. Przykład dyfrakcji fali ultradźwiękowej na przeszkodzie walcowej przedstawia il. 152 (tabl. 41).

powstawanie dźwięku w wirach

Bardzo często fale akustyczne w ośrodkach płynnych powstają przy przepływie. Szczególnie silnym źródłem fal dźwiękowych są zawirowania strumienia ośrodka (turbulencja). W zakresie słyszalnym wiele hałasów jest tego pochodzenia. Na il. 153 (tabl. 41) pokazano przykładowo powstawanie fal akustycznych w wirach strumienia powietrza wypływającego z dyszy.

Akustyczne zjawiska liniowe

Antoni Śliwiński

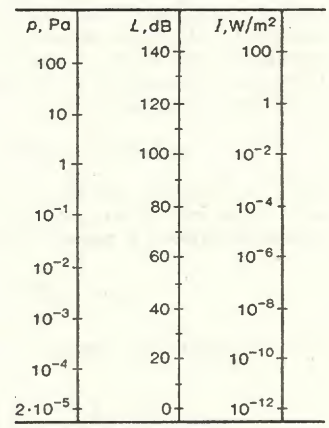
Zjawiska związane z rozchodzeniem się fal sprężystych w ośrodkach ciągłych, gdy zaburzenia (zmiany ciśnienia, gęstości) są małe, w porównaniu z wielkościami określającymi stany równowagi, rozpatruje się w liniowym przybliżeniu uzyskując zadowalającą zgodność z doświadczeniem. Zakłada się, że wielkości charakteryzujące pole akustyczne są wzajemnie do siebie proporcjonalne, a własności ośrodka opisane są przez współczynniki stałe niezależnie od wielkości zaburzenia. Takie założenie jest tym dokładniej spełnione, im mniejsze są zaburzenia sprężyste ośrodka, mówi się więc często o akustycznych zjawiskach liniowych jako o falach o małej amplitudzie (lub amplitudzie nieskończenie małej) w przeciwieństwie do fal o dużej amplitudzie (lub amplitudzie skończonej) w akustycznych zjawiskach nieliniowych.

fale o małej i o dużej amplitudzie

Amplitudy przesunięcia cząstki akustycznej w powietrzu w zakresie dźwięków słyszalnych są stosunkowo małe i wahają się od rzędu dziesiętnych części milimetra dla subiektywnie dużych natężeń, do rzędu miliardowych części milimetra dla natężeń ledwie dostrzegalnych przez ucho. Na przykład gwizd lokomotywy wywołujący uczucie bólu w uszach (próg bólu, częstość ok. 1000 Hz) powoduje przesunięcie cząstki akustycznej w powietrzu o ok. 10^{-3} mm. Fala o tej samej częstości, która jest zaledwie słyszalna (próg słyszalności) ma amplitudę przesunięcia rzędu 10^{-9} mm. Na rys. 10 podane są przykładowo wartości ciśnień akustycznych p oraz natężeń dźwięku I w powietrzu w całym zakresie słyszalnym (od progu słyszalności do progu bólu) dla częstości 1000 Hz. Wielkości te wyrażone są również (środkowy słupek) w powszechnie używanej w akustyce skali logarytmicznej (decybelowej). Za pomocą tej skali określa się względne wartości ciśnień, czy natężeń jako tzw. poziom L(dB) ciśnienia lub natężenia dźwięku, zgodnie z relacjami;

$$L(\mathrm{dB}) = 10 \lg I/I_0 = 10 \lg (p/p_0)^2 = 20 \lg p/p_0,$$

gdzie dla fali płaskiej $I = p^2/\varrho c$, ϱ — gęstość, c — prędkość dźwięku, $I_0 = 10^{-12}$ W/m² oraz $p_0 = 2 \cdot 10^{-5}$ N/m² odpowiadają wartościom progu słyszalności. Dla poziomów powyżej 100 dB efekty nieliniowe stają się już mierzalne.



Rys. 10. Porównanie wartości ciśnienia akustycznego i natężenia dźwięku I wyrażonych w skali liniowej i w skali decybelowej logarytmicznej (L)

Pojęcie wielkości amplitudy określamy względem długości fali akustycznej i zjawiska akustyczne możemy uważać za liniowe, jeżeli amplituda przesunięcia ξ jest znacznie mniejsza od długości fali $|\xi| \ll \lambda$. Wtedy procesy rozchodzenia się fal sprężystych w ośrodku opisuje równanie falowe:

równanie falowe

$$\frac{\partial^2 \xi}{\partial t^2} = c^2 \frac{\partial^2 \xi}{\partial y^2},$$

którego rozwiązaniem dla przypadku jednowymiarowego jest fala płaska (tabl. 155, il. 41):

$$\xi = \xi_0 \sin \omega \left(t - \frac{x}{c}\right).$$

Natomiast dla przypadku trójwymiarowego rozwiązaniem jest fala kulista

$$\xi' = \frac{\xi_0'}{r}\sin\omega\left(t - \frac{r}{c}\right),$$

gdzie r — promień fali, ξ' — przesunięcie cząstki dla fali kulistej. Amplituda fali kulistej ξ_0'/r zależy odwrotnie proporcjonalnie od odległości od źródła, gdzie ξ_0' — amplituda na powierzchni źródła (kuli pulsującej) dla $r = r_0$.

pole fali płaskiej

Istotne dla zjawisk liniowych w polu akustycznym fali płaskiej jest to, że pomiędzy wielkościami charakteryzującymi pole akustyczne: przesunięciem cząstki ξ, prędkością cząstki $d\xi/dt$, względnymi zmianami gęstości $\delta = \Delta\varrho/\varrho$ i ciśnieniem akustycznym p zachodzą związki:

$$\frac{d\varrho}{dt} = \xi_0\omega\cos\omega\left(t - \frac{x}{c}\right) = \frac{d\xi_0}{dt}\cos\omega\left(t - \frac{x}{c}\right),$$

$$\delta = \delta_0\cos\omega\left(t - \frac{x}{c}\right), \quad p = p_0\cos\omega\left(t - \frac{x}{c}\right),$$

gdzie $\xi_0\omega = d\xi_0/dt$ — amplituda prędkości cząstki akustycznej, δ_0 — amplituda względnych zmian gęstości, p_0 — amplituda zmian ciśnienia akustycznego.

Stosunek ciśnienia akustycznego p do prędkości cząstki $d\xi/dt$ jest wielkością stałą dla fali płaskiej i charakterystyczną dla danego ośrodka; nazywa się oporem akustycznym właściwym ośrodka i wynosi

$$p\Big/\left(\frac{d\xi}{dt}\right) = \varrho_0 c,$$

gdzie ϱ_0 — gęstość ośrodka w równowadze (niezaburzonego).

akustyczny ośrodek liniowy

Charakterystyczne dla zjawisk liniowych jest to, że kształt fali nie zależy od odległości od źródła. Ośrodek, w którym to zachodzi, nazywa się akustycznym ośrodkiem liniowym. Zakłada się, że taki ośrodek jest ośrodkiem idealnym, tzn. z jednej strony jest zdolny odkształcać się, aby przenieść zaburzenie, zaś z drugiej strony jego własności przy deformacjach nie ulegają zmianie, co oznacza, że np.

$$\varrho_0 \approx \varrho_0 + \delta\varrho, \quad p \approx p + \delta p,$$

czyli zmieniona wielkość prawie nie różni się od wielkości przy równowadze, przed zaburzeniem. Konsekwencją tych założeń jest to, że matematyczny opis zjawisk jest taki sam zarówno we współrzędnych Lagrange'a jak i we współrzędnych Eulera.

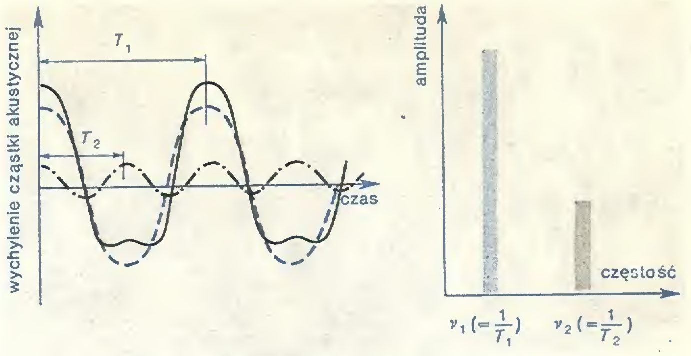
zasada liniowej superpozycji

W stosunku do akustycznych zjawisk liniowych obowiązuje zasada tzw. superpozycji liniowej, co oznacza, że przy spotkaniu się kilku fal, fala wypadkowa jest sumą fal składowych. Spotykające się fale interferują między sobą (nakładają się) w obszarze wzajemnego przenikania się, jednak nie powstają żadne fale dodatkowe w wyniku tego wzajemnego oddziaływania (oddziaływania fonon–fonon). Interferencja fal jest typowym zjawiskiem liniowym (il. 156, tabl. 41) (→ Holografia akustyczna).

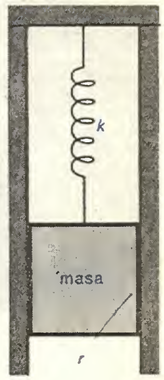
Przy analizie zjawisk dyfrakcyjnych (ugięcia fal) zwykle zakłada się, że spełniona jest zasada superpozycji liniowej (zasada Huygensa) i pole dyfrakcyjne jest wynikiem nałożenia na siebie fal, które na skutek ugięcia zmieniły kierunek rozchodzenia się (il. 157, tabl. 41). Falę złożoną (wypadkową) można wyrazić jako sumę składowych o różnych częstościach niezależnych od siebie (rys. 11).

liniowe źródła akustyczne

W odniesieniu do źródeł akustycznych (układów drgających) jak również przetworników akustycznych (układów przetwarzających energię akustyczną na inne rodzaje energii) mówimy, że są liniowe, jeśli procesy w nich zachodzące nie zależą od amplitudy drgań lub amplitudy sygnałów przenoszonych. Własności takich układów można opisać za pomocą współczynników stałych (rys. 12).



Rys. 11. Przykład fali złożonej z dwóch składowych o różnych częstościach: a) schemat superpozycji, b) widmo prążkowe



Rys. 12. Liniowy mechaniczny układ drgający o jednym stopniu swobody; k — współczynnik sprężystości — sprężyny, r opór mechaniczny

W układach liniowych:

— drgania okresowe są stabilne (tzn., że małe działania na układ powodują małe zmiany jego ruchu),

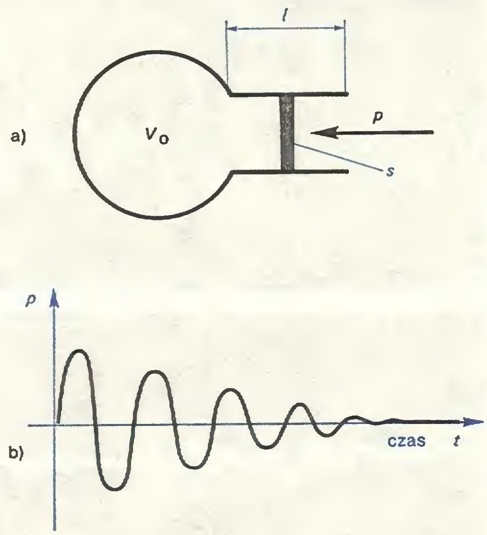
— drgania swobodne układów tłumionych gasną,

— w układzie o n stopniach swobody istnieje tylko n częstości własnych (rezonansowych),

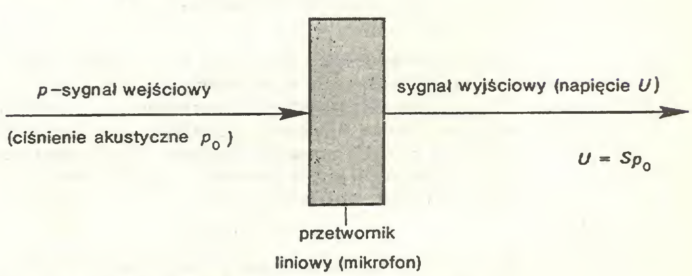
— częstości własne układu zależą tylko od współczynników charakteryzujących układ.

rezonator Helmholtza

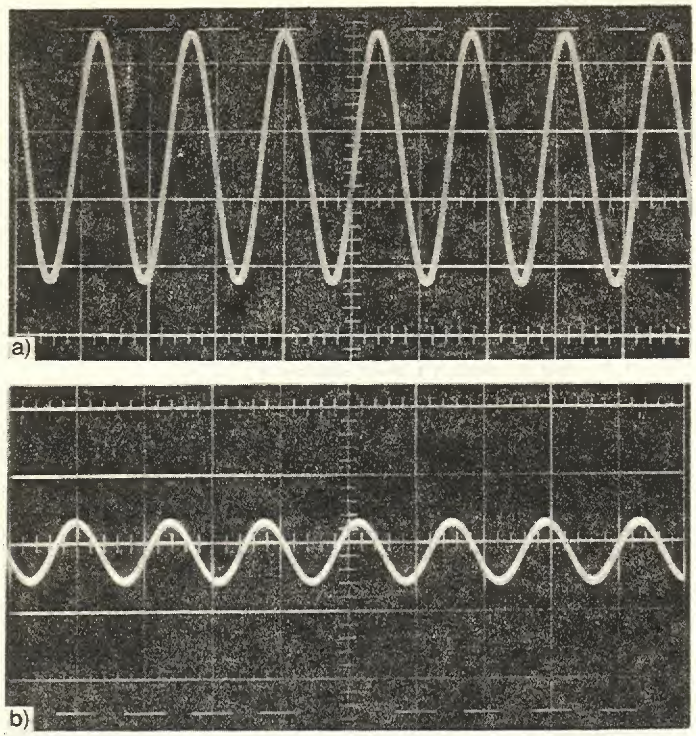
Przykładem liniowego układu akustycznego jest tzw. rezonator Helmholtza (rys. 13); drgania powietrza w szyjce rezonatora zależą tylko od masy akustycznej $m_a = \varrho_0 l/s$, podatności akustycznej $c_a = V_0/\gamma p_0$ i oporu akustycznego $r_a = r/s^2$, gdzie ϱ_0 — gęstość, l — długość szyjki rezonatora, s — powierzchnia przekroju szyjki, r — opór mechaniczny



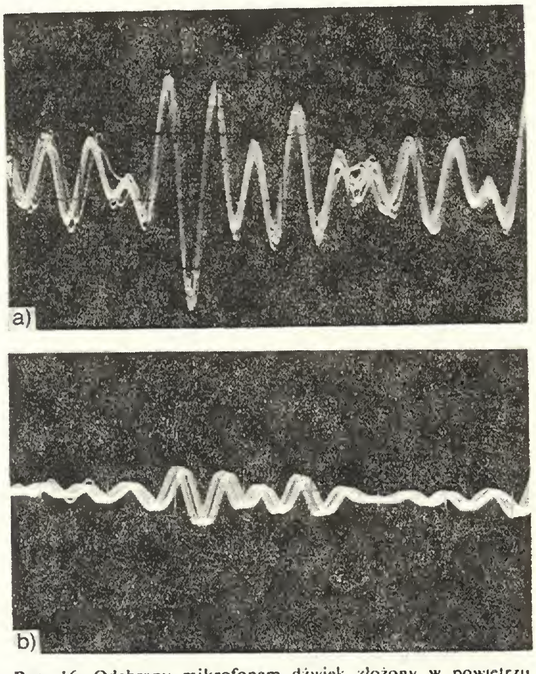
Rys. 13. Liniowy akustyczny układ drgający (rezonator Helmholtza): a) schemat budowy; V_0 objętość, l długość szyjki, s powierzchnia przekroju szyjki, p ciśnienie; b) drgania tłumione takiego układu



Rys. 14. Przykład mikrofonu jako przetwornika liniowego



Rys. 15. Odebrany mikrofonem czysty ton w powietrzu w dwóch odległościach od źródła: a) 0,5 m, b) 5 m



Rys. 16. Odebrany mikrofonem dźwięk złożony w powietrzu w dwóch odległościach od źródła: a) 0,5 m, b) 5 m

tarcia wprowadzający tłumienie, V_0 — objętość rezonatora, p_0 — ciśnienie w równowadze.

mikrofony głośniki

Przykładem liniowych przetworników akustycznych mogą być dobre (nie zniekształcające) mikrofony lub głośniki, których stosunek sygnału odbieranego do przekazywanego jest wielkością stałą niezależną od wielkości tego sygnału (rys. 14).

$$U/p = S = \text{const};$$

S — nazywa się skutecznością (lub czułością) mikrofonu i nie zależy od wielkości sygnału.

Na rys. 15 przedstawiono przykład pomiaru pojedynczego tonu o częstości 1000 Hz w powietrzu za pomocą mikrofonu i oscyloskopu (układ liniowy) przy dwóch różnych odległościach od źródła. Jak widać, na skutek zwiększenia odległości od źródła dźwięku zmniejszyła się tylko amplituda dźwięku, a sinusowa postać sygnału nie uległa zniekształceniu. W przypadku tego słabego sygnału powietrze można uważać za ośrodek liniowy. Na rys. 16 przestawiony jest podobny pomiar dla dźwięku złożonego, którego charakter widma również nie ulega zmianie przy zmianie odległości.

Akustyczne zjawiska nieliniowe

Antoni Śliwiński

Zjawiska akustyczne nie przebiegają liniowo, gdy zaburzenie sprężyste ośrodka jest duże w porównaniu z wielkościami określającymi jego stan równowagi. W zjawiskach nieliniowych nie można przyjąć założenia, że amplituda wychylenia cząstki akustycznej w fali sprężystej jest bardzo mała w porównaniu z długością fali ($|\xi| \approx \lambda$). W zjawiskach nieliniowych mamy do czynienia z tzw. falami o amplitudzie skończonej, w przeciwieństwie do fal o amplitudzie nieskończenie małej, w zjawiskach liniowych, a opis matematyczny ich rozchodzenia się uzyskuje się jako rozwiązanie nieliniowych równań różniczkowych. Istotą akustycznych zjawisk nieliniowych jest to, że stosunki pomiędzy wielkościami charakteryzującymi pole akustyczne nie są stałe, lecz zależą od wielkości zaburzenia. Im większe zaburzenie, tym większe są efekty nieliniowe.

Typowymi akustycznymi zjawiskami nieliniowymi są: zniekształcenia frontu falowego w miarę przebiegu przez ośrodek (rys. 17) oraz występowanie ciśnienia promieniowania fali (ciśnienia wywieranego przez czoło fali na ośrodek) powodującego stały przepływ ośrodka w kierunku rozchodzenia się fali sprężystej tzw. „wiatr akustyczny" (il. 163, tabl. 42). Do akustycznych zjawisk nieliniowych należą również takie zjawiska jak kawitacja ultradźwiękowa, fale uderzeniowe, oddziaływania dźwięku z dźwiękiem (oddziaływania fonon–fonon), przy których zasada superpozycji nie jest spełniona.

W zakresie dźwięków słyszalnych efekty nieliniowe są bardzo rzadko brane pod uwagę. Są one mierzalne dopiero przy poziomach natężeń dźwięku powyżej 100 dB i występują np. przy przelotach samolotów odrzutowych, szczególnie przy przekroczeniu bariery dźwięku, gdy powstaje fala uderzeniowa (ang. *boom* lub franc. *bang*, rys. 18), przy wybuchach i innych silnych a krótkotrwałych (impulsowych lub udarowych) zjawiskach akustycznych.

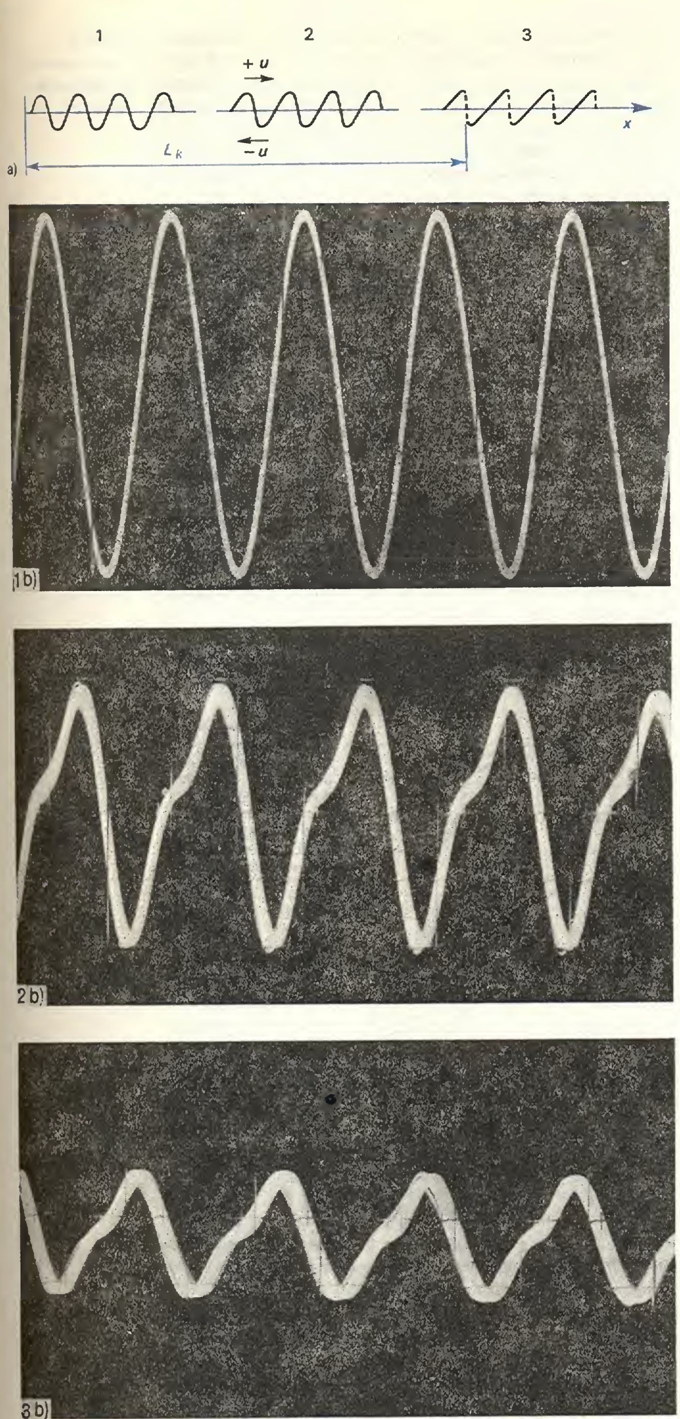
Zjawiska nieliniowe stają się szczególnie wyraźne w zakresie ultradźwięków, ponieważ w miarę wzrostu częstości, a więc zmniejszenia się długości fali, uzyskanie dużych amplitud w porówaniu z długością fali jest stosunkowo coraz łatwiejsze.

Uwzględnienie efektów nieliniowych w matematycznym opisie fal sprężystych wymaga wyrazów wyższych rzędów i opis wyraźnie zależy od tego czy używamy podejścia Lagrange'a, czy też Eulera. W odniesieniu do gazów równanie falowe z uwzględnieniem efektów nieliniowych we współrzędnych Lagrange'a ma postać (równanie 3):

równanie fal sprężystych w gazach

$$\frac{\partial^2 \xi}{\partial t^2} = \frac{c_0^2}{\left(1 + \frac{\partial \xi}{\partial a}\right)^{\gamma+1}} \frac{\partial^2 \xi}{\partial a^2}.$$

W przypadku cieczy równanie stanu $p = p(\varrho, S)$, gdzie p — ciśnienie, ϱ — gęstość, S — entropia, przyjmuje się w postaci rozwinięcia w szereg Taylora:



Rys. 17. Postępujące zniekształcenie fali w przypadku dużej amplitudy. a) Schemat: *1* — blisko źródła — kształt sinusoidalny; *2* — w pewnej odległości od źródła — kształt piłowaty ale jeszcze ciągły; *3* — w odległości krytycznej L_k — fala stała się nieciągła. b) Odebrany mikrofonem sygnał L = 130 dB czystego tonu (1000 Hz) w powietrzu i jego zniekształcenie: *1b* w odległości 0,5 m od źródła; *2b* w odległości 2,5 m od źródła; *3b* w odległości 5 m od źródła

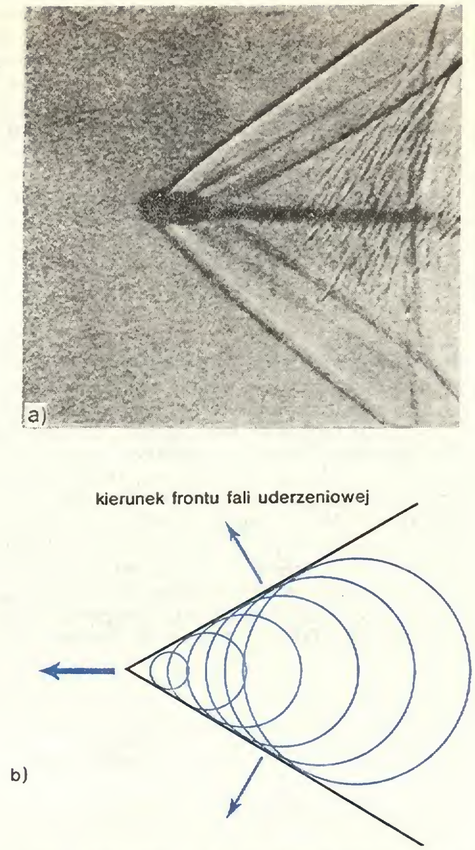
wirialne równanie stanu

$$p = p_0 + \left(\frac{\partial p}{\partial \varrho}\right)_{S,\rho=\rho_0} (\varrho - \varrho_0) + $$

$$+ \frac{1}{2}\left(\frac{\partial^2 p}{\partial \varrho^2}\right)_{S,\rho=\rho_0} (\varrho - \varrho_0)^2 + \dots ,$$

co można przy pominięciu wyrazów wyższego rzędu niż drugi zapisać jako

fala uderzeniowa



Rys. 18. Fala uderzeniowa. Źródło dźwięku (zaburzenia) porusza się szybciej niż dźwięk: a) fotografia cieniowa poruszającej się kuli karabinowej: b) schemat rozchodzenia się fal dźwiękowych (kulistych) i tworzenie się frontów fali uderzeniowej, gdy źródło porusza się z prędkością większą od prędkości dźwięku

$$p - p_0 = A\left(\frac{\varrho - \varrho_0}{\varrho_0}\right) + \frac{B}{2}\left(\frac{\varrho - \varrho_0}{\varrho_0}\right)^2 .$$

A, B — nazywają się pierwszym i drugim współczynnikiem wirialnym i zależą od temperatury, p_0 i ϱ_0 odpowiadają równowadze ośrodka. $A = \varrho_0 c_0^2$, natomiast

$B/A = 2\varrho_0 c_0 (\partial c/\partial p)_{S,\rho=\rho_0}$, gdzie $c = \left(\frac{\partial p}{\partial \varrho}\right)^{1/2}_{S,\rho=\rho_0}$ — jest

prędkością dźwięku. Stosunek B/A jest miarą efektów nieliniowych w cieczach. Wprowadzenie go do równania falowego ($B/A = \gamma - 1$) prowadzi do równania

równanie fal sprężystych w cieczach

$$\frac{\partial^2 \xi}{\partial t^2} = \frac{c_0^2}{\left(1 + \frac{\partial \xi}{\partial a}\right)^{\frac{B}{A}+2}} \frac{\partial^2 \xi}{\partial a^2} .$$

Rozwiązanie tego równania daje wzór na prędkość rozchodzenia się fali sprężystej o skończonej amplitudzie

$$c = c_0[1 + (B/2A)(u/c_0)],$$

co oznacza, że zależy ona od prędkości cząstki akustycznej $u = \partial\xi/\partial t$. Prędkość cząstki akustycznej u jest wielkością okresowo zmienną i zmienia fazę co pół okresu fali (rys. 17a (2)) przyjmując odpowiednio wartości $+u$ w fazie zagęszczenia i $-u$ w fazie rozrzedzenia (dla fali podłużnej). Z powyższego wzoru wynika, że fala porusza się w fazie zgęszczenia z prędkością $c = c_0(1 + bu)$, a więc szybciej, natomiast w fazie rozrzedzenia z prędkością $c = c_0(1 - bu)$, a więc wolniej (w tych wzorach przyjęto $b = B/2Ac_0$). Prowadzi to do deformacji kształtu fali wraz z odległością przebywaną przez falę, przy czym deformacja

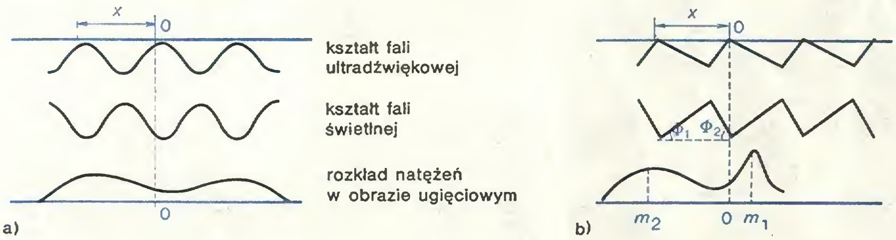
zniekształcenia frontu falowego

ta narasta do pewnej odległości krytycznej, przy której fala przyjmuje kształt przedstawiony na rys. 17a(3), odpowiadający doścignięciu rozrzedzenia przez zgęszczenie. W rzeczywistości efekt wyprzedzania nie może zachodzić dalej, gdyż istnieje skok od zgęszczenia do rozrzedzenia i równowaga możliwa jest tylko jako nieciągłość ośrodka w tym miejscu — tworzy się fala uderzeniowa (udarowa). Odległość krytyczna dla utworzenia takiego frontu udarowego wynosi

$$L_{kr} = \frac{\lambda \varrho_0 c_0^2}{\left(\frac{B}{A}+2\right)\pi p_a},$$

gdzie λ — długość fali, p_a — amplituda ciśnienia akustycznego przy źródle. Zniekształcenie frontu falowego jest wyrazem powstawania wyższych harmonicznych rozchodzącej się fali. Istnienie składowych harmonicznych można wykazać przez ich odfiltrowanie z fali głównej. Michajłow i Szutiłow (wg których jest il. 160 z tabl. 42) byli pierwszymi, którzy zastosowali metodę optyczną do analizy nieliniowych zniekształceń fal ultradźwiękowych. Zasadę takiej analizy wyjaśnia poglądowo rys. 19. W przypadku fali akustycznej o małym natężeniu (rys. 19a) jej kształt jest sinusoidalny i wtedy rozkład natężeń światła ugiętego jest symetryczny względem wiązki nieugiętej, natomiast w przypadku fali akustycznej o dużym natężeniu

analiza zniekształceń



Rys. 19. Schematyczne przedstawienie kształtu fal przy dyfrakcji światła na fali ultradźwiękowej: a) przy małym natężeniu fali ultradźwiękowej, b) przy dużym natężeniu fali ultradźwikowej

(rys. 19b) jej kształt jest zdeformowany, piłokształtny i rozkład natężeń światła ugiętego jest niesymetryczny (por. il. 160, tabl. 42).

ciśnienie promieniowania

Działanie ciśnienia promieniowania, które powoduje przepływ ośrodka (il. 163, tabl. 42) w kierunku fali, ma specyficzny charakter na granicy dwóch ośrodków. W szczególności na granicy ciecz–powietrze, gdy falę ultradźwiękową dużej mocy promieniuje się pionowo do góry (il. 162, tabl. 42), powstaje nad swobodną powierzchnią cieczy charakterystyczna fontanna. Wytrysk fontanny następuje wtedy, gdy siły pochodzące od ciśnienia promieniowania przewyższą siły napięcia powierzchniowego. Wysokość fontanny jest tym większa, im większe jest natężenie fali ultradźwiękowej. Kierunek działania siły na granicy dwóch ośrodków zależy od stosunku wartości gęstości energii akustycznej w tych ośrodkach. Gdy $W_1 > W_2$ (W_1 — gęstość energii w ośrodku, w którym znajduje się źródło fali, W_2 — gęstość energii w drugim ośrodku), to siła wywołująca fontannę działa zgodnie z biegiem fali, natomiast gdy $W_2 < W_1$, to siła ma zwrot przeciwny do biegu fali. W przypadku szczególnym, gdy fala ultradźwiękowa przechodzi przez granicę ośrodków bez odbicia, to natężenia fali w obydwu ośrodkach są sobie równe i zachodzi związek

$$W_1 c_1 = W_2 c_2,$$

gdzie c_1 i c_2 — prędkość rozchodzenia się dźwięku odpowiednio w ośrodku pierwszym i drugim. Wynika stąd, że dla $c_1 > c_2$ będzie $W_2 > W_1$ (siła ciśnienia promieniowania przeciwna do biegu fali), a dla $c_1 < c_2$ będzie $W_2 < W_1$ (siła zgodna z biegiem fali). Gdy graniczą ze sobą dwa ośrodki ciekłe nie mieszające się ze sobą, to w zależności od rodzaju ośrodka mogą zachodzić obydwa przypadki (il. 161, tabl. 42).

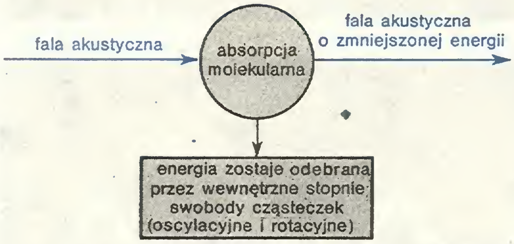
efekty nieliniowe w ciałach stałych

Efektom nieliniowym w ciałach stałych poświęca się w ostatnich latach bardzo wiele uwagi. Większość prac dotyczy wyznaczania stałych elastycznych wyższych rzędów oraz analizy niedoskonałości siatki krystalicznej (→ Wzbudzenia elementarne w ciałach stałych). Zjawisko powstawania wyższych harmonicznych przy rozchodzeniu się fali ultradźwiękowej w ciałach stałych zostało wykorzystane do parametrycznego wzmacniania fal ultradźwiękowych.

Akustyczne procesy molekularne

Antoni Śliwiński

Akustyczne procesy molekularne występują jako oddziaływanie fali sprężystej przechodzącej przez ośrodek z jego drobinami. Oddziaływanie to polega na wymianie energii pomiędzy falą a poszczególnymi stopniami swobody drobin (rys. 20). W zależności od częstości fali sprężystej oddziaływanie może być procesem bardziej lub mniej nieodwracalnym z punktu widzenia przenoszenia energii przez falę, dlatego w różnych zakresach częstości procesy molekularne w różnym stopniu wpływają na tłumienie (absorpcję) fali akustycznej. W niektórych zakresach częstości wpływ ten jest bardzo duży i jego wynikiem jest występowanie charakterystycznego dla danego ośrodka



Rys. 20. Schemat procesu tłumienia fali akustycznej przez procesy molekularne w ośrodku

maksimum tłumienia. Zwykle towarzyszy temu pojawienie się dyspersji prędkości rozchodzenia się fali, czyli jej zależności od częstości.

W klasycznym opisie zjawisk akustycznych nie bierze się pod uwagę procesów molekularnych, które zachodzą przy zmianach stanu ośrodka, gdy przechodzi przez niego fala sprężysta.

amplitudowy i energetyczny współczynnik tłumienia

Tłumienie fal dźwiękowych w ośrodku opisuje się za pomocą tzw. amplitudowego współczynnika tłumienia α, który wyraża względny zanik amplitudy na jednostkę przebytej przez falę drogi:

$$\frac{dA}{A} = -\alpha dx,$$

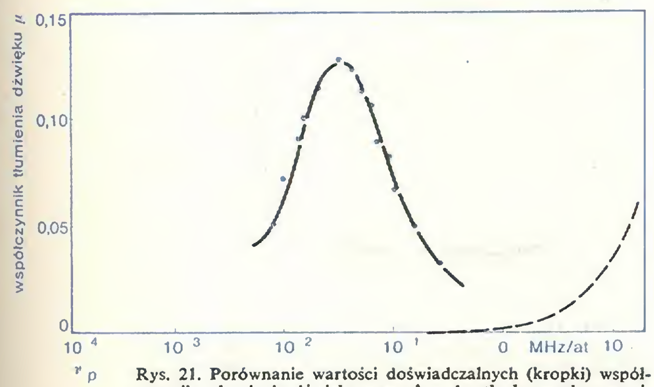
gdzie: A — amplituda fali sprężystej, dA — zmiana amplitudy na drodze dx, minus oznacza, że amplituda maleje, lub za pomocą tzw. energetycznego współczynnika tłumienia γ, przy czym $\gamma = 2\alpha$, gdyż przyjmuje się (w zakresie akustycznych zjawisk liniowych), że natężenie fali, równe liczbowo energii przepływającej przez jednostkę przekroju w jednostce czasu, jest proporcjonalne do kwadratu amplitudy.

W ujęciu klasycznym przyjmuje się, że jedynymi czynnikami wpływającymi na tłumienie dźwięku są lepkość (tarcie wewnętrzne) ośrodka oraz przewodnictwo cieplne (w mniejszym stopniu promieniowanie cieplne). Zgodnie z obliczeniami przeprowadzonymi przez Stokesa oraz Kirchhoffa tzw. klasyczny współczynnik tłumienia dźwięku powinien być wprost proporcjonalny do kwadratu częstości fali akustycznej zgodnie z wzorem

klasyczny współczynnik tłumienia

$$\alpha = \frac{2\pi\nu^2}{\varrho c^3}\left[\frac{4}{3}\eta+\varkappa\left(\frac{1}{c_V}-\frac{1}{c_p}\right)\right],$$

gdzie: ν jest częstością fali dźwiękowej, η — współczynnikiem lepkości stycznej, ϱ — gęstością, c — prędkością dźwięku, $\varkappa = c_p/c_V$, c_p jest ciepłem właściwym przy stałym ciśnieniu, c_V — ciepłem właściwym przy stałej objętości. Zatem $\alpha/\nu^2 = \alpha'$ powinno być niezależne od częstości fali, ale już stosunkowo dawno stwierdzono (Nieklapajew i Lebiediew w 1911 r.), że współczynnik tłumienia α' w niektórych gazach a także i cieczach wykazuje odstępstwa od wzoru. Rysunek 21 pokazuje takie odstępstwo w CO_2 w temperaturze pokojowej. Absorpcję dźwięku inną niż przewiduje to akustyka klasyczna można w zadowalający sposób wyjaśnić za pomocą molekularnej (relaksacyjnej) teorii absorpcji dźwięku.



Rys. 21. Porównanie wartości doświadczalnych (kropki) współczynnika tłumienia dźwięku $\mu = \alpha\lambda$ w dwutlenku węgla z wynikami obliczeń na podstawie wzoru klasycznego (krzywa przerywana) i wzoru relaksacyjnego (krzywa ciągła)

Współczynnik tłumienia fal sprężystych α oraz prędkość rozchodzenia się dźwięku c w danym ośrodku są wielkościami dla niego charakterystycznymi. W ujęciu klasycznym nie tylko $\alpha' = \alpha/\nu^2$, ale również c nie powinno zależeć od częstości. Jeżeli w ośrodkach rzeczywistych pojawia się zależność od częstości, to możemy wyciągnąć wniosek, że jest to wynik molekularnych oddziaływań ośrodka z falą sprężystą. Akustyczne procesy molekularne są różne w zależności od reakcji ośrodka na pobudzenie energią fali akustycznej. W cieczach i gazach spotykamy głównie relaksację termiczną i strukturalną, zwaną inaczej objętościową.

relaksacja termiczna

W pierwszym przypadku są to procesy związane z wymianą energii pomiędzy falą a poszczególnymi stopniami swobody ruchu cząsteczek (zewnętrznymi — translacyjnymi, wewnętrznymi — rotacyjnymi i oscylacyjnymi). Są one analogiczne do procesów aktywacji termicznej tych stopni swobody, polegających na uzyskiwaniu przyrostów energii kinetycznej poszczególnych cząsteczek na skutek zmian temperatury.

relaksacja strukturalna

W drugim przypadku są to procesy związane z wymianą energii pomiędzy falą a cząsteczkami lub zespołami cząsteczek (asocjatami) w taki sposób, że następuje zmiana struktury molekularnej, np. jest to przechodzenie jednych struktur w drugie (przegrupowania atomów), albo dysocjacja czyli rozpad cząsteczek lub ich zespołów, a czasami jest to proces odwrotny, czyli asocjacja (łączenie).

Procesy relaksacji polegają na opóźnieniu reakcji zaburzonego ośrodka (odchylonego od stanu równowagi) względem samego zaburzenia (fali sprężystej), o pewien czas τ zwany czasem relaksacji potrzebnym na powrót do równowagi. Wynikiem tego opóźnienia jest przesunięcie fazowe powstające między falą (ciśnieniem akustycznym) a reakcją ośrodka, a więc zmianami temperatury, zmianami ruchu wewnętrznych stopni swobody, zmianami struktury itp. Konsekwencją tego opóźnienia fazowego jest dodatkowe tłumienie fali akustycznej; energia, która wzbudziła układ molekularny i jest z opóźnieniem przez niego oddawana (w innej fazie) nie może być z powrotem przez falę przejęta.

dyspersja prędkości dźwięku

Tej absorpcji relaksacyjnej równolegle towarzyszy dyspersja prędkości dźwięku, tj. jej zależność od częstości fali. Jest ona wynikiem tego, że dla różnych częstości omówione wyżej przesunięcia fazowe są różne.

W przypadku relaksacji strukturalnej (przy tworzeniu się i rozpadaniu asocjatów np. za pośrednictwem wiązania wodorowego w wodzie, alkoholach i wielu innych cieczach, w polimerach, przy przejściach fazowych itd.) obserwuje się stosunkowo dużą dyspersję dźwięku, natomiast przy relaksacji termicznej (np. przy wzbudzaniu rotacyjnych stopni swobody cząsteczek związków organicznych w różnych stanach izomerycznych) dyspersja prędkości dźwięku jest mała, niekiedy zupełnie niemierzalna.

współczynnik tłumienia relaksacyjnego

Wyrażenie na współczynnik tłumienia α' z uwzględnieniem określonego procesu molekularnego zależnego od częstości można przedstawić następująco:

$$\alpha' = \frac{\alpha}{\nu^2} = B+\frac{A}{1+\left(\frac{\nu}{\nu_m}\right)^2}, \tag{5}$$

gdzie stała B — odpowiada klasycznej części tłumienia, natomiast drugi składnik zdaje sprawę z dodatkowego tłumienia molekularnego (relaksacyjnego) $\alpha_{\text{rel}}/\nu^2$, A — stała, ν — częstość fali akustycznej, ν_m — tzw. częstość relaksacji, przy której występuje maksimum tłumienia relaksacyjnego, przy czym

$$\nu_m = \frac{1}{2\pi\tau}\frac{c_V}{c_a},$$

τ — czas relaksacji, który określa czas powrotu stanu zaburzonego do stanu równowagi, c_V — ciepło właściwe przy stałej objętości, c_a — część ciepła właściwego przy stałej objętości związana z wewnętrznymi stopniami swobody cząsteczki, która absorbuje i oddaje (wypromieniowuje z pewnym opóźnieniem) energię fali akustycznej.

Często wprowadza się współczynnik tłumienia (bezwymiarowy) odniesiony do długości fali λ:

$$\mu = \alpha\lambda = \alpha\frac{c}{\nu},$$

(c — prędkość dźwięku). Wówczas w obszarze relaksacji

$$\mu_{\text{rel}} = \alpha_{\text{rel}}\lambda = Ac\nu_m\frac{\nu/\nu_m}{1+(\nu/\nu_m)^2} =$$

$$= 2\mu_{\max}\frac{\nu/\nu_m}{1+(\nu/\nu_m)^2}; \tag{6}$$

$\mu_{\max} = Ac\nu_m/2$ dla $\nu = \nu_m$, czyli dla $2\pi\nu\tau \approx 1$.

Prędkość dźwięku w obszarze relaksacji określa następujący wzór dyspersyjny

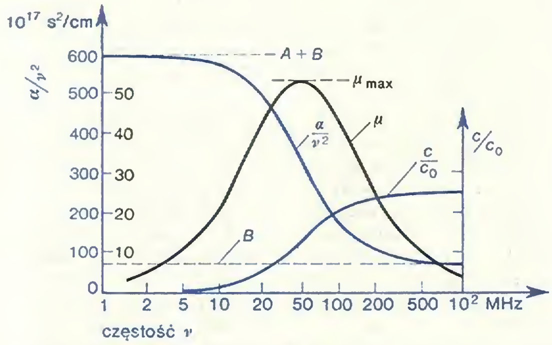
wzór dyspersyjny

$$\frac{c}{c_0} = \left[1+\frac{\varepsilon}{1-\varepsilon}\,\frac{\omega^2\tau^2}{1+\omega^2\tau^2}\right]^{-1/2}, \tag{7}$$

gdzie $\omega = 2\pi\nu$, $\varepsilon = 1-c_0^2/c_\infty^2$, c_0 jest prędkością dźwięku dla małych częstości ($\omega \to 0$) poniżej obszaru dyspersji, a c_∞ jest prędkościądźwięku dla bardzo wysokich częstości ($c \to \infty$) powyżej obszaru dyspersji.

Na rys. 22 pokazane są przykładowe zależności $\alpha' = \alpha/\nu^2$, $\mu_{rel} = \alpha_{rel}\,\lambda$ i c/c_0 jako funkcje częstości ν w obszarze relaksacji zgodnie z równaniami (2–7).

Dla pojedynczych procesów molekularnych (o jednym czasie relaksacji), np. w układzie z dwoma stanami konfiguracji izomerowej znajomość μ_{max} oraz ν_m wystarcza do scharakteryzowania procesu. μ_{max} zależy tylko od wielkości termodynamicznych, natomiast

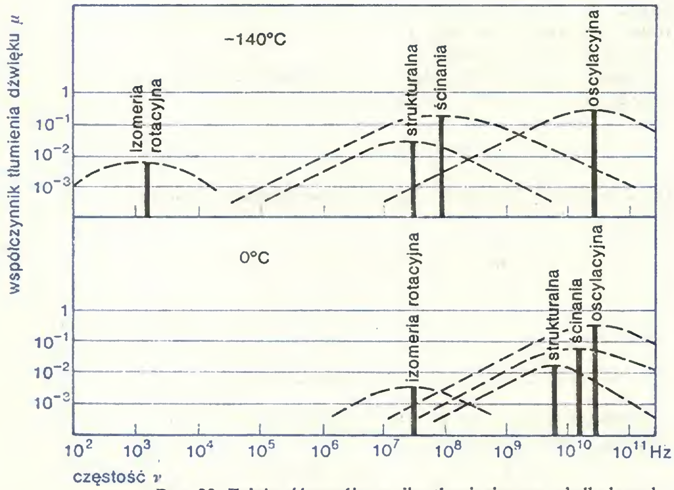


Rys. 22. Zależność wielkości α/ν^2, $\mu = \alpha\lambda$ oraz c/c_0 od częstości określone przez równania (5), (6) i (7) dla pojedynczego procesu relaksacji; $\nu_{max} = 50$ MHz, $c_0 = 1{,}2\cdot 10^5$ cm s^{-1}, $A = 524\cdot 10^{-17}$ s^2/cm, $B = 70\cdot 10^{-17}$ s^2/cm (wg J. Lamba)

ν_m określone jest przez kinetykę procesu. Na przykład w układzie złożonym z dwóch rodzajów izomerów μ_{max} związane jest z różnicą energii pomiędzy dwoma stanami odpowiadającymi równowadze konfiguracyjnej, zaś ν_m określa wysokość bariery potencjału dla wewnętrznej rotacji drobiny. μ_{max} oraz ν_m są przedmiotem wielu eksperymentalnych pomiarów, w szczególności bada się ich zależność od temperatury i ciśnienia. Zależności doświadczalne pozwalają wyznaczać energię aktywacji oraz parametry kinetyczne procesów molekularnych. Takie postępowanie leży u podstaw spektroskopii ultradźwiękowej. Powyższy opis procesu o jednym czasie relaksacji można rozszerzyć na przypadek, gdy istnieje więcej niż jeden obszar relaksacji, tzn. odnieść również do procesów o wielu czasach relaksacji.

podstawy spektroskopii ultradźwiękowej

Na rys. 23 podany jest schematyczny spektrogram dla bromku izobutylowego, który pokazuje różne procesy relaksacyjne w odpowiednich obszarach

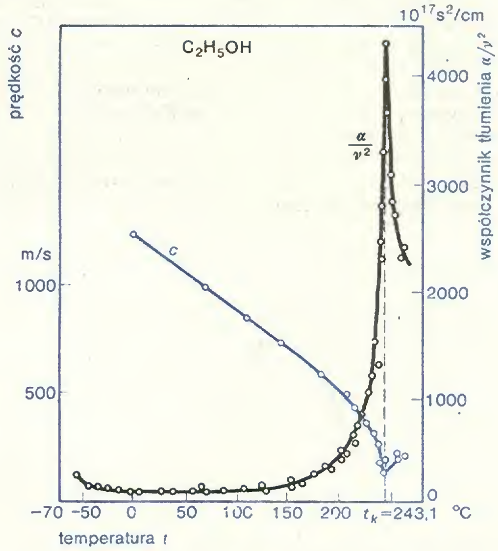


Rys. 23. Zależność współczynnika tłumienia $\mu = \alpha\lambda$ dla bromku izobutylowego od częstości ν dla dwóch temperatur: −140°C i 0°C. Pionowe linie odpowiadają częstościom relaksacyjnym ν_m i tłumieniom μ_{max} dla każdego z procesów relaksacyjnych. Linie przerywane przedstawiają schematycznie przebiegi tłumienia μ w różnych zakresach częstości, w których dominują procesy relaksacyjne związane z wymianą energii fali z różnymi stopniami swobody cząsteczki. Niektóre obszary zachodzą na siebie, dają się jednak rozdzielić przy obniżeniu temperatury (wg Litovitza i Davisa)

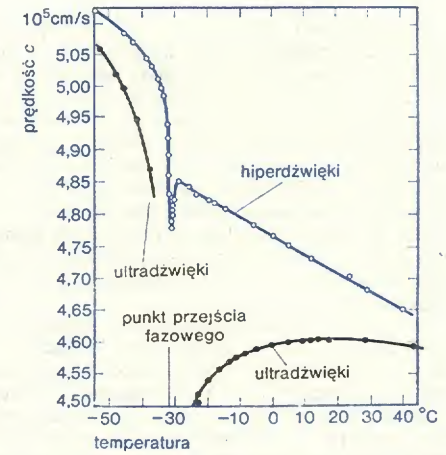
częstości. Porównanie górnego (dla temperatury −140°C) i dolnego (dla temperatury 0°C) rysunku pozwala zauważyć, jak przesuwają się zakresy relaksacyjne w zależności od temperatury. W praktyce zakres pomiarowy jest w skali częstości ograniczony; zmiany temperaturowe pozwalają więc odpowiednio przesuwać położenie na skali częstości. Zmiany ciśnienia w gazach wywołują również przesunięcie obszarów relaksacyjnych. Bada się wtedy μ w zależności od stosunku ν/p, gdzie p — ciśnienie gazu.

relaksacja w pobliżu przejść fazowych

Zjawiska relaksacji strukturalnej odgrywają dużą rolę przy rozchodzeniu się ultradźwięków w układach niejednorodnych, np. w pobliżu przejść fazowych, gdzie obserwuje się występowanie ostrego maksimum tłumienia dźwięku oraz charakterystycznego minimum dla prędkości rozchodzenia się dźwięku (rys. 24 i 25).



Rys. 24. Zależność od temperatury współczynnika tłumienia α/ν^2 i prędkości rozchodzenia się c fali ultradźwiękowej o częstości 2 MHz w alkoholu etylowym C_2H_5OH w pobliżu punktu krytycznego (t_k = 243,1°C) (przejścia ciecz–para); wg N. F. Nozdriewa



Rys. 25. Zależność prędkości rozchodzenia się podłużnej fali sprężystej rozchodzącej się w kierunku [100] w NH_4Cl w pobliżu punktu przejścia fazowego „porządek”–„nieporządek” (wg P.D. Lazaya)

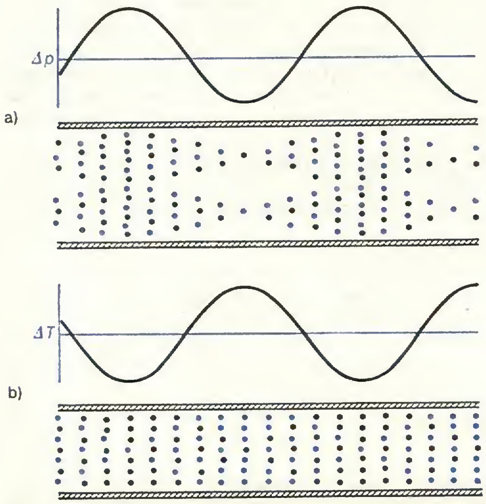
akustyczne procesy molekularne w ciekłym helu

Bardzo interesująco przebiegają akustyczne procesy molekularne w ciekłym helu szczególnie w pobliżu punktu λ, poniżej którego (w helu II) występuje zjawisko nadpłynności (→ Nadpłynność). W ostatnich doniesieniach w literaturze spotyka się prace dotyczące dalszych faz helu — helu III i helu IV. Tłuma-

czy się je zwykle za pomocą tzw. modelu dwucieczowego, w myśl którego hel II jest mieszaniną dwu cieczy nie oddziałujących na siebie: jednej normalnej a drugiej kwantowej. W myśl tego modelu w helu II powstają dwie fale akustyczne odpowiadające tym dwu cieczom. Jedna z nich o prędkości $c_1 = (\partial p/\partial \varrho)_s$ zwana dźwiękiem normalnyml ub „pierwszym dźwiękiem", oraz druga o prędkości $c_2 = (\varrho_s/\varrho_n) S^2 (\partial T/\partial S)_\varrho$, zwana „drugim dźwiękiem" (lub naddźwiękiem), gdzie p — ciśnienie, ϱ — gęstość, przy czym $\varrho = \varrho_n + \varrho_s$ (ϱ_n — gęstość składowej normalnej, ϱ_s — gęstość składowej nadciekłej), S — entropia, T — temperatura.

dźwięk „pierwszy"

„Pierwszy dźwięk" jest zwyczajną falą sprężystą, w której obydwa składniki helu II poruszają się razem i wykazują periodyczne zmiany ciśnienia i gęstości (rys. 26a). „Drugi dźwięk" co do swej natury nie jest falą dźwiękową w sensie zmian gęstości i ciśnienia, jest to natomiast fala zmian temperatury i entropii przy stałej gęstości, gdy obydwa składniki poruszają się względem siebie (rys. 26b).

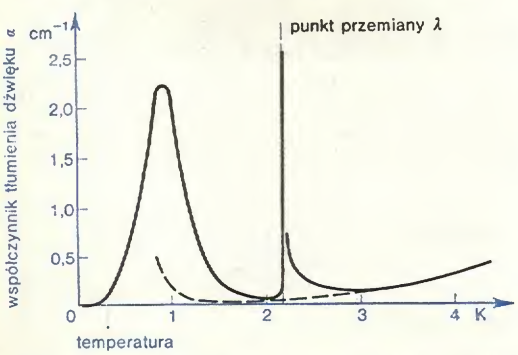


Rys. 26. Schemat ułożenia cząstek przy rozchodzeniu się w helu II dźwięku pierwszego (a) i drugiego (b). Punkty niebieskie i czarne oznaczają odpowiednio cząstki akustyczne cieczy normalnej i nadpłynnej

dźwięk „drugi"

Rozchodzenie się „drugiego dźwięku" w He II zostało po raz pierwszy zaobserwowane przez Pieszkowa w 1944 r. Wytwarzał on zmienną temperaturę za pomocą zanurzonego w He II drucika, przez który płynął prąd zmienny o częstości kilku kHz i odbierał falę „drugiego dźwięku" za pomocą termopary z fosforo-brązu.

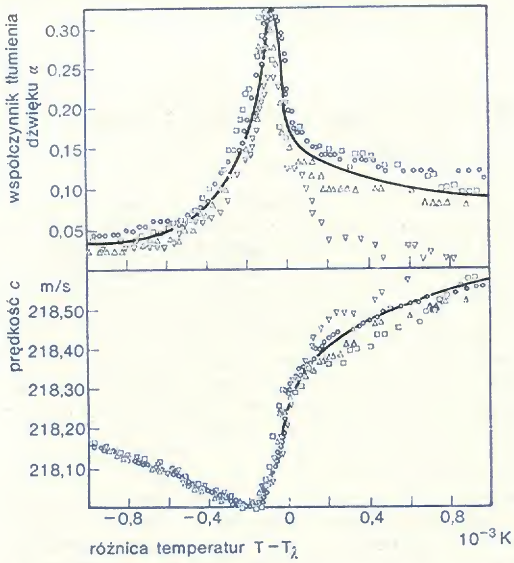
Rysunek 27 przedstawia wynik pomiaru absorpcji „pierwszego dźwięku" w ciekłym helu (He) fali ultradźwiękowej o częstości 12 MHz. Powyżej punktu λ w helu I tłumienie przebiega zgodnie z teorią klasyczną. W okolicy punktu λ tłumienie silnie i anormalnie wzrasta, w okolicy temperatury 1 K występuje bardzo regularne maximum tłumienia. Tego rodzaju tłumienie dźwięku w ciekłym helu zostało wyjaśnione przez Landaua i Kalatnikowa na gruncie kwantowym jako występowanie dwóch procesów rozpraszania fononów na fononach (proces pięciofononowy) oraz fononów z rotonami (kwanty energii rotacyjnej). Każdemu z tych procesów odpowiada określony czas relaksacji: τ_{ff} oraz τ_{fr}.



Rys. 27. Tłumienie dźwięku pierwszego o częstości 12 MHz w ciekłym helu w pobliżu punktu przemiany λ. Linia przerywana (pozioma) oznacza tłumienie obliczone w oparciu o teorię klasyczną (wg Atkinsona)

tłumienie i dyspersja w pobliżu punktu λ

Relaksacyjnemu tłumieniu w pobliżu punktu λ odpowiada również wyraźna dyspersja fal ultradźwiękowych (rys. 28).



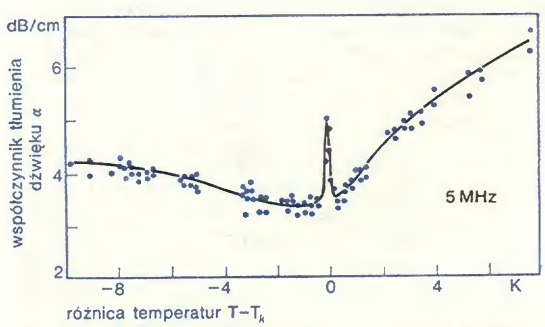
Rys. 28. Tłumienie i prędkość fali ultradźwiękowej o częstości 1 MHz w pobliżu punktu przemiany λ, wyniki pochodzą z różnych pomiarów (wg Atkinsona)

Z uwagi na duże analogie przejścia fazowego λ w ciekłym helu do przejścia ze stanu przewodnictwa do nadprzewodnictwa w metalach (→ Nadprzewodnictwo) przeprowadza się również w niektórych laboratoriach na świecie badania tych przejść metodami ultradźwiękowymi. Ostatnio analogię tę przeniesiono (de Genne) na przejście fazowe smektyk A-nematyk w ciekłych kryształach (→ Ciekłe kryształy).

tłumienie w ciekłym krysztale

Ultradźwiękowe badania tego przejścia w ciekłym krysztale CBOOA (N-*p*-cyjanobenzilideno-*p*-*n*-oktylooksyanilina) wykazały występowanie maksimum tłumienia fal ultradźwiękowych, którego charakter odpowiada przejściu typu λ (rys. 29).

Badania akustycznych procesów molekularnych są tym ciekawsze, im większe są częstości stosowanych



Rys. 29. Tłumienie fali ultradźwiękowej o częstości 5 MHz w pobliżu temperatury t_k przejścia smektyk A-nematyk, w ciekłym krysztale CBOOA (wg R. Bartolino, F. Scudieri, D. Sette i A. Śliwiński)

fal sprężystych, gdy długość fali staje się bardziej współmierna z rozmiarami molekularnymi, rozwijają się więc szczególnie w zakresie hiperdźwiękowym.

W ciałach stałych odpowiednikiem akustycznych procesów molekularnych są oddziaływania fali sprężystej z atomami sieci krystalicznej (→ Wzbudzenia elementarne w ciałach stałych) i jej defektami oraz swobodnymi nośnikami ładunku.

Akustyka molekularna jak i kwantowa bardzo się w ostatnich latach rozwinęły dzięki coraz to większym możliwościom eksperymentalnych badań ultradźwiękowych i hiperdźwiękowych w pełnym zakresie widma częstości fal sprężystych. W wielu laboratoriach na świecie (również w Polsce) prowadzi się obecnie badania akustycznych procesów molekularnych a wyniki wykorzystuje się np. w chemii, biologii, czy materiałoznawstwie.

Ultradźwięki i hiperdźwięki

Antoni Śliwiński

Fale sprężyste o częstościach powyżej słyszalnych to jest powyżej 16 kHz — częstość taką przyjmuje się jako granicę słyszalności, nazywa się ultradźwiękami (naddźwiękami), przy czym terminem tym obejmuje się zjawiska do częstości 10^9 Hz. Fale sprężyste, których częstość przewyższa 10^9 Hz, nazywa się hiperdźwiękami. Hiperdźwiękowy zakres zjawisk sprężystych kończy się od góry naturalną granicą wyznaczoną dla danego ośrodka przez częstość odpowiadającą długości fali porównywalnej z odstępami międzyatomowymi. Fale sprężyste o mniejszych długościach nie mogą powstać, gdy znikają warunki konieczne dla ich rozprzestrzeniania się (to jest znika możliwość przekazania zaburzenia sprężystego jako energii mechanicznej pomiędzy atomami w drodze bezpośredniej). W kryształach te graniczne częstości przypadają w zakresie 10^{12}–10^{13} Hz. W powietrzu pod normalnym ciśnieniem hiperdźwięki właściwie nie mogą się rozchodzić, gdyż już przy częstości 10^9 Hz długość fali staje się porównywalna ze średnią drogą swobodną cząsteczek. W cieczach hiperdźwięki wytwarzane są przez cieplne fluktuacje gęstości i teoretyczną górną granicą jest tutaj 10^{14} Hz.

granice częstości hiperdźwięków

Hiperdźwięki zostały wydzielone w osobną grupę, której dolną granicą jest 109 Hz, ponieważ dotychczasowe metody eksperymentalne nie pozwalały wytwarzać i następnie odbierać fal sprężystych w ośrodkach o częstościach większych niż 10^9 Hz. Granica ta jednak przesuwa się coraz wyżej (w monokryształach udało się wzbudzić i odbierać fale sprężyste, których częstość wynosiła $8 \cdot 10^{13}$ Hz). Z drugiej strony częstości fal sprężystych wzbudzonych przez cieplne drgania atomów w sieci krystalicznej, ruchy oscylacyjne cząsteczek, jak również cieplne fluktuacje gęstości w cieczach obejmują zakres częstości powyżej 10^9 Hz.

własności ultradźwięków i hiperdźwięków

W przeciwieństwie do ultradźwięków, które wytwarzamy i odbieramy jako koherentne (spójne) wiązki fal sprężystych, hiperdźwiękowe fale cieplne (tzw. debayowskie) stanowią zbiór fal niekoherentnych rozchodzących się we wszystkich możliwych kierunkach wewnątrz ciała, które rejestrujemy tylko pośrednimi metodami wykorzystując zjawiska akustooptyczne (→ Akustyczne zjawiska kwantowe).

Ultradźwięki, a tym bardziej hiperdźwięki, odznaczają się małymi długościami fal, np. przy częstości 16 kHz długość fali w powietrzu wypada ok. 2 cm, w cieczach ok. 8 cm, w ciałach stałych ok. 30 cm, natomiast przy 10^9 Hz długość fali w powietrzu byłaby ok. $3 \cdot 10^{-4}$ cm, w cieczy rzędu $1{,}2 \cdot 10^{-4}$ i w ciele stałym $4 \cdot 10^{-4}$ cm. W zakresie hiperdźwiękowym długości fal sprężystych stają się porównywalne z długościami światła widzialnego (400–800 nm), w granicznym przypadku w ciałach stałych przy częstości 10^{12}–10^{13} Hz długości fal sprężystych wynoszą od 5–0,5 nm.

Niewątpliwie małe długości fal ultra- i hiperdźwiękowych zadecydowały o specjalnym ich zastosowaniu. Dzięki małym długościom fal ultradźwięki można wizualizować za pomocą światła (zjawiska akustooptyczne), można je ogniskować i kształtować w wiązki o dobrej kierunkowości (→ Akustyka morza, → Holografia akustyczna) i można mówić z dobrym przybliżeniem o promieniach ultradźwiękowych.

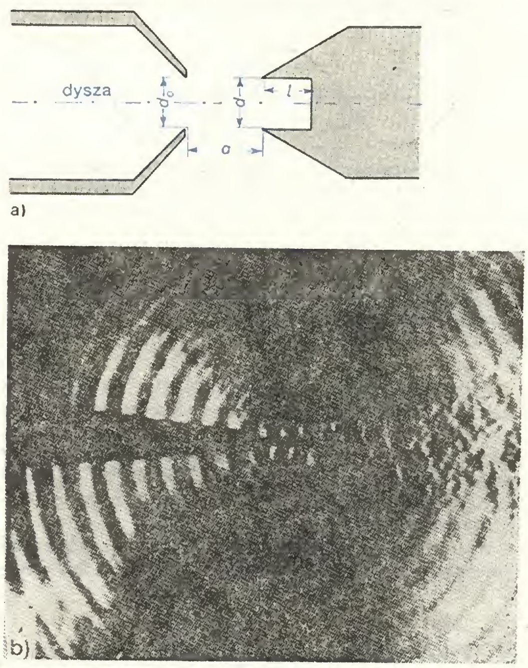
zastosowanie ultradźwięków

Dzięki małym długościom fal a wysokim częstociom (natężenie jest proporcjonalne do kwadratu częstości) można stosunkowo łatwo otrzymywać ultradźwięki o dużym natężeniu (dziesiątki W/cm^2), przy którym pojawiają się zjawiska nieliniowe nie występujące przy falach o małej amplitudzie.

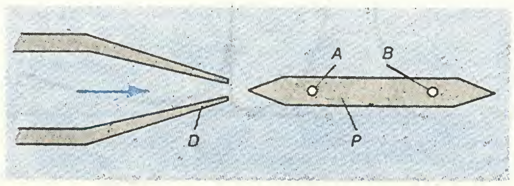
Jeśli długości fal ultradźwiękowych stają się porównywalne z wielkością niejednorodności lub ziarnistości ośrodka, w szczególności z rozmiarami określającymi jego strukturę molekularną, wtedy charakter rozchodzenia się tych fal zależy wyraźnie od własności ośrodka. Badając prędkość rozchodzenia się i tłumienia fal ultradźwiękowych możemy określać procesy molekularne w różnych ośrodkach (→ Badanie ośrodków za pomocą ultradźwięków). Oddziaływanie fal ultradźwiękowych o dużych częstościach ze strukturą ośrodka prowadzi do wystąpienia zjawisk, które wykazują kwantowy charakter tych procesów.

przetworniki ultradźwiękowe

Do wytwarzania i odbioru ultradźwięków stosuje się różne urządzenia zwane przetwornikami ultradźwiękowymi. Przetwarzają one energię określonego



Rys. 30. Generator Hartmanna — źródło periodycznych zmian ciśnienia z częstością ultradźwiękową: a) schemat, b) fotografia cieniowa wytwarzanej fali. Gaz wypływa z dyszy z prędkością naddźwiękową. Naprzeciwko dyszy ustawiony jest rezonator, który powoduje wzmocnienie ciśnienia akustycznego: a) schemat, b) fotografia wytworzonej przez niego fali



Rys. 31. Piszczałka Pohlmana-Jankowskiego. Wypływająca z dyszy *D* ciecz wzbudza drgania rezonansowe płytki *P* o częstości ultradźwiękowej; *A*, *B* — węzły drgań płytki

układu nieakustycznego, drgającego z częstością ultradźwiękową, w energię akustyczną (przetworniki nadawcze — generatory ultradźwiękowe) lub odwrotnie — energię akustyczną w energię innego rodzaju (przetworniki odbiorcze). W zależności od rodzaju energii, która jest przetwarzana na akustyczną lub odwrotnie, rozróżniamy generatory lub odbiorniki ultradźwiękowe mechaniczne, elektryczne, magnetyczne, cieplne, chemiczne i optyczne. Przetwornikami odwracalnymi nazywają się takie, które działają w obydwu kierunkach z równymi sprawnościami. Różne rodzaje przetworników ultradźwiękowych dzieli się na grupy biorąc za podstawę zasadę działania wykorzystującą określone zjawisko fizyczne, w którym zachodzi przetwarzanie jednej energii w drugą.

przetworniki mechaniczne

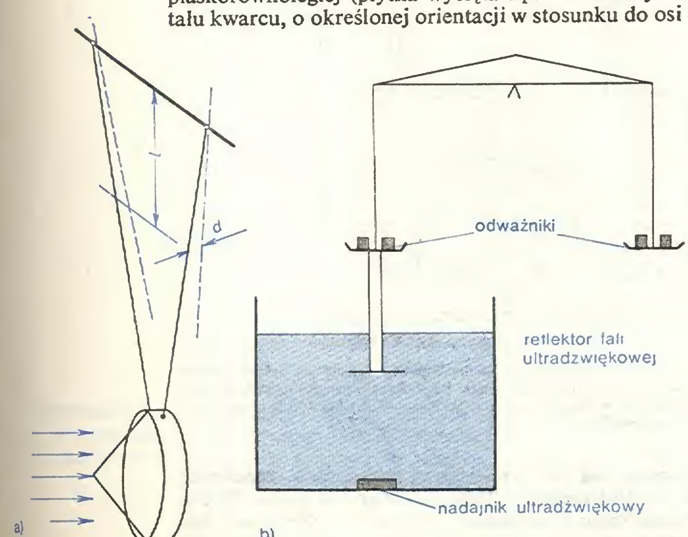
Przetworniki mechaniczne stanowiące czyste oscylatory mechaniczne jak piszczałki, syreny (generatory) i radiometry (odbiorniki) są w większości przypadków nieodwracalne, czyli używane albo jako nadajnik albo jako odbiornik. Układy takie służą najczęściej do wytwarzania i odbioru ultradźwięków dużej mocy w zakresie częstości rzadko kiedy przekraczającym 50 kHz. Generatory tego typu mają duże zastosowanie w ośrodkach gazowych (np. syrena, generator Hartmanna widoczny na rys. 30 do koagulacji dymów) i rzadziej ciekłych (np. piszczałka Pohlmana–Jankowskiego widoczny na rys. 31 do tworzenia emulsji, np. homogenizacji śmietanki).

Szczególną rolę spełniają przetworniki umożliwiające wizualizację pola ultradźwiękowego stosowane w optosonice (→ Akustyczne zjawiska kwantowe) i w holografii akustycznej (→ Holografia akustyczna).

Radiometry ultradźwiękowe służą do pomiaru natężenia fal ultradźwiękowych wykorzystując ciśnienie promieniowania wywierane przez falę ultradźwiękową na powierzchnię prostopadłą do kierunku rozchodzenia się (rys. 32).

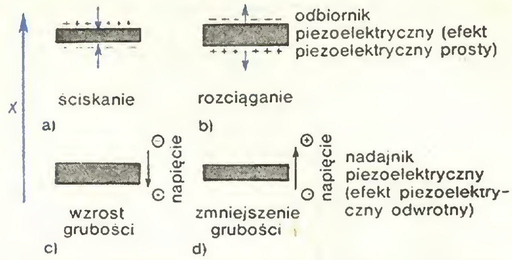
przetworniki piezoelektryczne

Przetworniki piezoelektryczne i elektrostrykcyjne działają na zasadzie zjawiska piezoelektrycznego (występującego w monokryształach piezoelektrycznych) i elektrostrykcyjnego (występującego w polikryształach piezoelektrycznych), które polegają na powstawaniu deformacji sprężystych w wyniku przyłożenia pola elektrycznego. Efekty te są odwracalne i wykorzystuje się je w nadajnikach i odbiornikach ultradźwiękowych w zakresie częstości od 20 kHz do 10 GHz. Tego rodzaju przetworniki zadecydowały w ciągu ostatnich 50 lat o olbrzymim rozwoju ultradźwięków. Rysunki 33 i 34 ilustrują zastosowanie płytki piezoelektrycznej płaskorównoległej (płytka wycięta np. z monokryształu kwarcu, o określonej orientacji w stosunku do osi

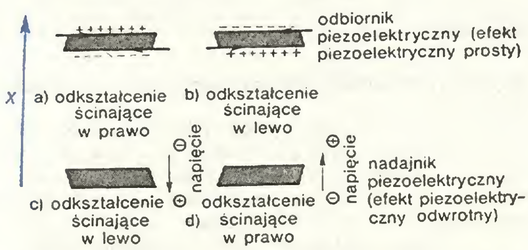


Rys. 32. Radiometry ultradźwiękowe: a) w postaci wahadła (miarą ciśnienia promieniowania jest kąt wychylenia określony przez stosunek d/l); b) w postaci wagi hydrostatycznej (miarą ciśnienia promieniowania jest zmniejszenie siły wyporu ciężarka)

piezoelektrycznej i posrebrzona po obu płaskich stronach) jako nadajnik i jako odbiornik. Zwykle tak dobiera się grubość płytki, aby wzbudzane drgania mechaniczne odpowiadały jej drganiom rezonansowym (gdy $d = \lambda/2$). Istnieje więc odwrotna proporcjonalność pomiędzy grubością płytki d a częstością

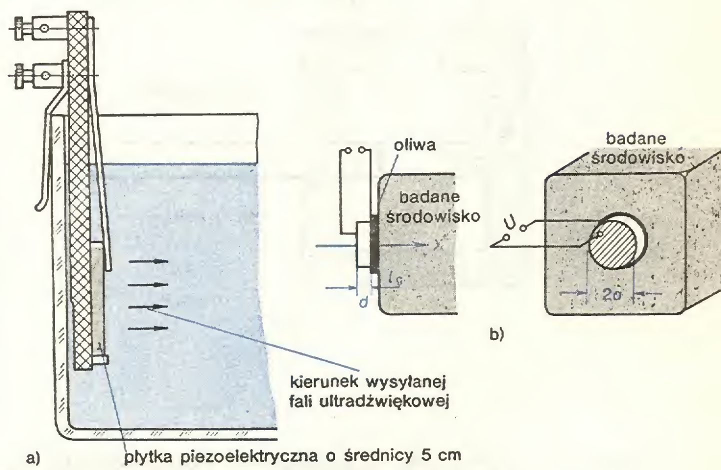


Rys. 33. Wykorzystanie efektu piezoelektrycznego w ultradźwiękowym przetworniku fal podłużnych (w przypadku kwarcu o cięciu X): jako odbiornika (a i b) oraz nadajnika (c i d). Rozkład ładunku powstającego na elektrodach przy ściskaniu (a) oraz przy rozciąganiu (b). Zwiększenie grubości płytki (c) oraz zmniejszenie grubości płytki (d) wywołane przez przykładane napięcie



Rys. 34. Wykorzystanie efektu piezoelektrycznego w ultradźwiękowym przetworniku fal poprzecznych (w przypadku kwarcu o cięciu Y): jako odbiornika (a) i (b) oraz nadajnika (c) i (d). Rozkład ładunku przy ścinającym odkształceniu: dodatnim (a) oraz ujemnym (b). Odkształcenie ścinające ujemne (c) oraz dodatnie (d) wywołane przez przykładane napięcie

drgań rezonansowych ν: $d = c/2\nu$, gdzie c — prędkość dźwięku w płytce. Rysunek 35 pokazuje przykład sprzężenia płytki piezoelektrycznej z ośrodkiem,



Rys. 35. Sprzężenie przetwornika piezoelektrycznego z ośrodkiem: a) przetwornik w kontakcie z cieczą; b) przetwornik w kontakcie z ciałem stałym — sprzężenie poprzez warstewkę oliwy celem lepszego dopasowania, u oznacza napięcie zmienne, d — grubość płytki równą $\lambda/2$ oraz $2a$ — średnica płytki (zwykle kilka centymetrów)

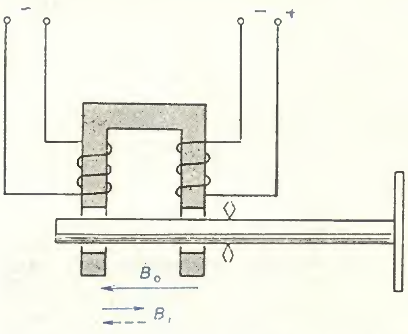
do którego wprowadzamy (lub z którego odbieramy) drgania ultradźwiękowe.

przetworniki magnetostrykcyjne

Przetworniki magnetostrykcyjne oparte są na zjawisku magnetostrykcji, czyli zmian objętości pod

wpływem pola magnetycznego; budowane są z materiałów ferromagnetycznych (metali ferromagnetycznych lub ferrytów). W większości wypadków są one używane w zakresie częstości do 40 kHz, chociaż można je także stosować w zakresie wyższych częstości. Niekiedy przetworniki magnetostrykcyjne stosuje się w zakresie megaherców, a nawet gigaherców. Na rys. 36 przedstawiony jest prętowy przetwornik magnetostrykcyjny pobudzany elektromagnesem w warunkach rezonansu.

prętowy przetwornik magnetostrykcyjny



Rys. 36. Schemat działania prętowego przetwornika magnetostrykcyjnego. Stałe pole magnetyczne o indukcji B_0 wprowadza magnetyczną polaryzację materiału, zmienne pole magnetyczne o indukcji B_i powoduje okresowe wydłużenie się i kurczenie pręta

Istnieje jeszcze wiele innych rodzajów przetworników ultradźwiękowych opartych na przykład na efektach termicznych, używanych zwłaszcza do pomiarów natężenia fal ultradźwiękowych, lub — na efektach chemicznych, służących do określania na przykład wydajności ultradźwiękowych procesów chemicznych.

W zakresie mniejszych częstości stosowane są również przetworniki elektrostatyczne i elektrodynamiczne działające na tej samej zasadzie co przetworniki dźwięków słyszalnych.

odbiorniki wiązek i sondy

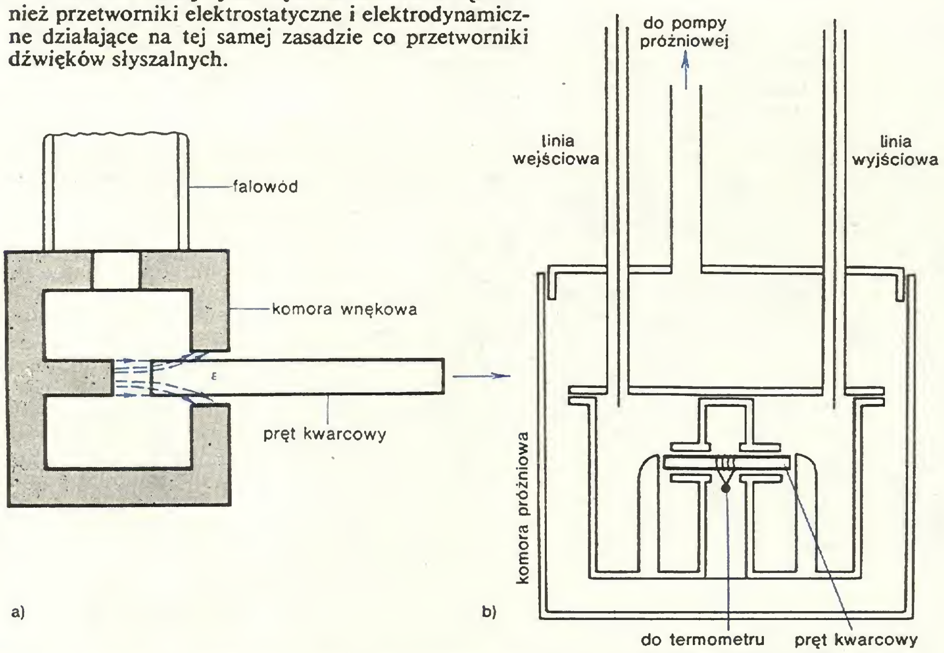
Odbiorniki ultradźwiękowe można podzielić na dwie grupy: na duże w porównaniu z długością fali, które służą zwykle do odbioru całej wiązki ultradźwiękowej (wtedy są większe od jej przekroju — rejestrują wartość sygnału uśrednioną na całą powierzchnię, rys. 32), i na bardzo małe w porównaniu z długością fali (np. mniejsze od $0{,}1\lambda$), i które działają jak sonda i służą do określania wartości pola ultradźwiękowego (ciśnienie akustyczne, prędkość cząstki lub natężenie) w danym punkcie. Rysunek 37 przedstawia piezoelektryczny mikrofon-sondę o średnicy milimetrowej do określania rozkładu pola ultradźwiękowego w cieczach (→ Fizyka morza) pracujący do częstości ok. 30 MHz.

Do wytwarzania i odbioru fal ultradźwiękowych i hiperdźwiękowych (o częstościach aż do 10^{10} Hz) stosuje się specjalne techniki, które rozwinęły się w ostatnich latach i ciągle są ulepszane. Osiągnięciem w tej dziedzinie było opracowanie w 1958 r. metody wytwarzania podłużnych i poprzecznych fal w pręcie kwarcowym przez umieszczenie go w rezonatorze mikrofalowym (rys. 38). Koniec pręta kwarcowego (lub innego kryształu piezoelektrycznego) umieszcza się we wnęce rezonatora mikrofalowego, tak jak to pokazuje rys. 38b. Energia elektromagnetyczna zostaje doprowadzona do rezonatora za pomocą falowodu mikrofalowego. Zmienne pole elektryczne wywołuje odkształcenie piezoelektryczne w objętym przez siebie obszarze. Fale akustyczne generowane są tylko w tym obszarze, a więc w powierzchniowej warstwie poprzecznej pręta na lewym jego końcu, rozchodzą się wzdłuż niego i odbijają od drugiego końca. W doświadczeniu przedstawionym na rys. 38b prawy koniec pręta umieszczony jest w bliźniaczej wnęce rezonansowej i fala akustyczna na końcu pręta wzbudza drgania mikrofalowe, czyli zostaje odebrana i zarejestrowana w układzie odbiorczym. Porównanie sygna-



Rys. 37. Ultradźwiękowy miniaturowy mikrofon tytaniano-barowy do pomiaru rozkładu pola ultradźwiękowego o rozmiarach kilku milimetrów

generacja we wnęce mikrofalowej



Rys. 38. Metoda wytwarzania hiperdźwięków we wnęce mikrofalowej: a) sprzężenie pręta piezoelektrycznego z polem elektrycznym w rezonatorze mikrofalowym, b) kompletny układ nadawczo-odbiorczy

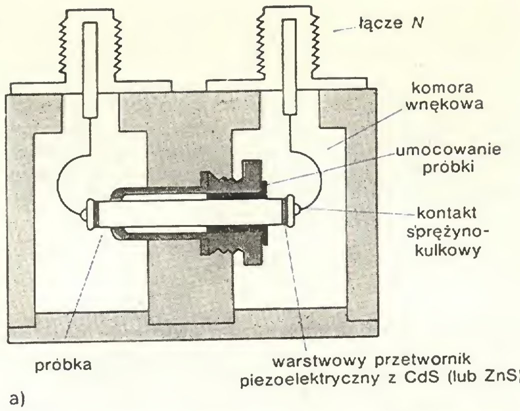
łu odebranego z sygnałem nadanym pozwala wyznaczyć tłumienie fali akustycznej i jej prędkość w próbce. W innych rozwiązaniach opartych na metodzie impulsowej rejestruje się sygnał odbity za pomocą tego samego falowodu.

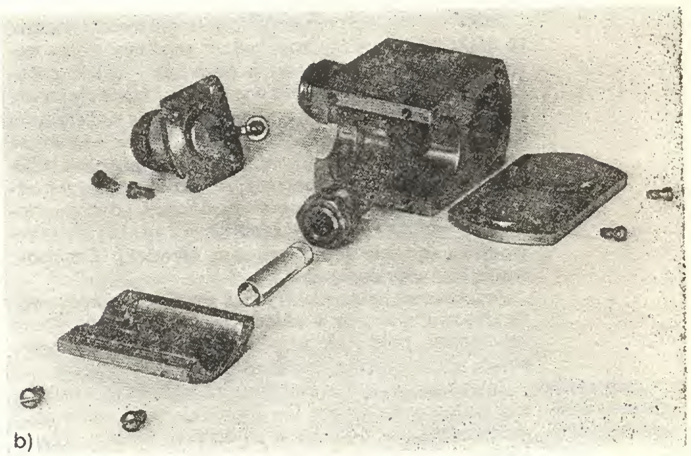
Metoda generacji fal hiperdźwiękowych w pręcie piezoelektrycznym umieszczonym w rezonatorze mikrofalowym ograniczona jest do badania rozchodzenia się ich w tym samym pręcie. Przekazanie fali do innej próbki wiąże się z dużymi stratami, które powstają na granicy zetknięcia pręta piezoelektrycznego z próbką, nawet przy bardzo dokładnym doszlifowaniu obydwu powierzchni, czy też stosowaniu odpowiednich warstewek sprzęgających. Problem generacji hiperdźwięków w próbkach z dowolnego materiału

przetworniki cienkowarstwowe

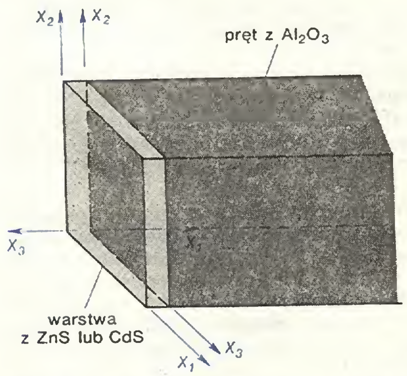
został rozwiązany głównie przez zastosowanie cienkich warstw magnetostrykcyjnych lub piezoelektrycznych, które osadza się bezpośrednio (metodą napylania katodowego, naparowania lub elektrolitycznie) na powierzchni próbki.

Aby materiał mógł stanowić cienkowarstwowy przetwornik piezoelektryczny powinien mieć następujące własności: duży opór elektryczny, duży stopień orientacji krystalograficznej przy osadzaniu, łatwość osadzania, dużą wartość współczynnika sprzężenia elektromechanicznego (kwadrat tego współczynnika jest równy stosunkowi otrzymanej energii mechanicznej drgań do wprowadzonej energii elektrycznej). Takim materiałem okazał się siarczek kadmu i chociaż jego współczynnik sprzężenia elektromechanicznego nie jest szczególnie duży w porównaniu z innymi materiałami, jednak dzięki stosunkowo dużej łatwości osadzania, zyskał największą popularność (rys. 39). Warstewki monokrystaliczne mają grubości rzędu 0,3–3 μm, co odpowiada podstawowym częstościom własnym (grubość równa $\lambda/2$) rzędu 10^9–10^{10} Hz.





Rys. 40. Podwójny rezonator wnękowy z próbką i warstewkami nadawczą i odbiorczą dla pobudzenia drgań podłużnych — pole E_3: a) schemat, b) fotografia rezonatora (wg J. de Klerka)

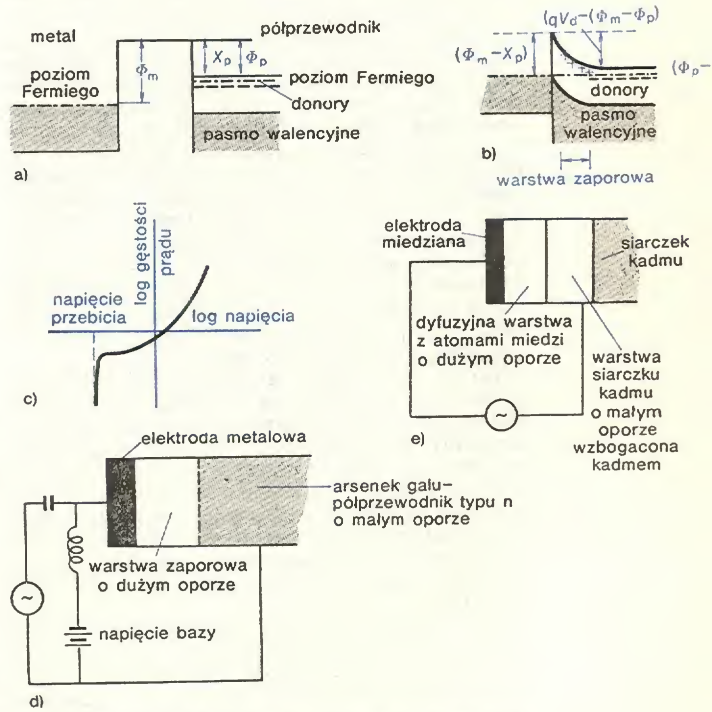


Rys. 39. Przykład wzajemnej orientacji kierunków krystalograficznych warstewki piezoelektrycznej CdS lub ZnS oraz jej podłoża — pręta Al_2O_3 (wg J. de Klerka)

Rysunek 40 przedstawia przykład pobudzenia cienkowarstwowego przetwornika piezoelektrycznego w mikrofalowej wnęce rezonansowej do drgań podłużnych (rys. 40a). Rysunek 40b przedstawia fotografie mikrofalowych rezonatorów wnękowych rozłożonych na części. Przedstawiony tu układ jest nadawczo-odbiorczy, składa się więc z dwóch identycznych wnęk połączonych ze sobą poprzez badaną próbkę; jest to metoda przepuszczania. Sygnał wejściowy doprowadza się jednym falowodem, a wyjściowy odprowadza się drugim.

przetworniki z warstwą zaporową

Specjalny rodzaj przetworników cienkowarstwowych stanowią układy piezoelektryczne półprzewodnikowe z warstwą zaporową. Warstwa zaporowa powstaje przy zetknięciu metalu i półprzewodnika (rys. 41) albo dwóch półprzewodników typu *p* i *n*. Ponieważ poziomy Fermiego (poziomy energetyczne, dla których prawdopodobieństwo obsadzeń elektronów w atomach siatki jest 50%) (→ Wzbudzenia elementarne w ciałach stałych) w metalu i w półprzewodniku



Rys. 41. Tworzenie się warstewki zaporowej na powierzchni zetknięcia metalu i półprzewodnika (wg Beechama): a) płytka metalu i płytka półprzewodnika o różnych poziomach Fermiego, b) wyrównanie poziomów Fermiego, c) przebieg charakterystyki prądowo-napięciowej, d) schemat przetwornika z warstwą zaporową, e) schemat przetwornika z warstwą dyfuzyjną

różnią się między sobą, powstaje po zetknięciu płytek napięcie kontaktowe, które jest wynikiem wyrównania tych poziomów. Warstwa zaporowa stanowi przestrzeń wolną od ładunku, który zostaje przesunięty poza nią, dzięki temu wytwarza się różnica potencjałów; jest ona rzędu 1 V dla przeciętnych układów metal–półprzewodnik (np. arsenek galu–złoto). Wielkość napięcia kontaktowego (V_d) jest wprost proporcjonalna do różnicy pracy wyjścia z metalu (Φ_m) i pracy wyjścia z półprzewodnika (Φ_p):

$$V_d = \frac{\Phi_m - \Phi_p}{q},$$

gdzie q — ładunek elementarny elektronu. Warstwa ta jest bardzo cienka i ma duży opór elektryczny, przy czym jej otoczenie jest przewodzące. Przyłożenie zewnętrznego pola elektrycznego do warstwy zaporowej zmienia jej grubość.

Jeżeli użyty półprzewodnik jest piezoelektrykiem, to efekt piezoelektryczny będzie możliwy tylko wewnątrz warstwy zaporowej, a poza nią — gdzie próbka przewodzi — nie będzie występował (zwarcie). Gdy do warstwy przyłożymy zmienne napięcie elektryczne, dzięki efektowi piezoelektrycznemu będzie ona drgać generując w próbce falę sprężystą. Fala sprężysta przechodząc z warstwy zaporowej do próbki półprzewodnika napotyka ten sam opór akustyczny, gdyż własności mechaniczne warstwy zaporowej i próbki półprzewodnika są takie same (ten sam ośrodek). Dopasowanie jest więc doskonałe.

Przetworniki piezoelektryczne z warstwą zaporową są stosowane często w układach impulsowych; wtedy zewnętrzne pole stałe przykłada się też tylko impulsowo, aby nie przegrzewać układu.

zastosowanie złącza nadprzewodzącego do generacji i detekcji

Interesujące są też możliwości wykorzystania złącza nadprzewodzącego (→ Zjawiska tunelowe w nadprzewodnikach) do generacji i detekcji fal hiperdźwiękowych. Pozytywne rezultaty w tym zakresie uzyskali po raz pierwszy Eisenmenger i Dayem w 1966 r.

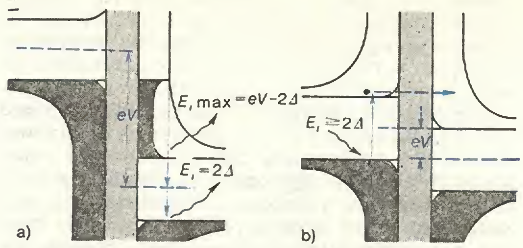
Zjawisko Josephsona polega na tym, że jeżeli dwa ciała nadprzewodzące zbliżyć do siebie zachowując między nimi przerwę izolacyjną rzędu kilku nanometrów, to przez tę przerwę może płynąć prąd w wyniku tunelowego przejścia elektronów (zjawiska tunelowania) przez przerwę złącza. Mogą wystąpić dwie składowe prądu tunelowego: stała i zmienna. Zjawisku tunelowego przejścia niesparowanych elektronów przez warstwę izolującą (przerwę energetyczną) między dwoma nadprzewodnikami towarzyszy dwustopniowa emisja fononów relaksacyjnych — w rezultacie zmiany poziomów energetycznych elektronów, które uległy tunelowaniu, i rekombinacyjnych — w rezultacie łączenia się elektronów w pary (pary Coopera). Emitowane fonony tworzą falę hiperdźwiękową. Używa się tutaj złączy z nadprzewodników o grubościach rzędu 100 nm oddzielonych przerwą rzędu 1 nm i powierzchni generującej ok. 1 mm^2.

Do uzyskania zjawiska tunelowania niesparowanych elektronów konieczna jest wstępna polaryzacja elektryczna złącza nadprzewodzącego takim napięciem, aby odpowiadało ono szerokości przerwy energetycznej (rys. 42a), to znaczy żeby był spełniony warunek $V \geqslant 2\Delta/e$, gdzie V — napięcie przyłożone, Δ — szerokość przerwy energetycznej, e — ładunek elektronu. Przy przerwach energetycznych rzędu 1 MeV występujących w znanych nadprzewodnikach uzyskuje się fale hiperdźwiękowe o częstościach od 100 GHz do 1 THz.

Jeżeli napięcie $V < 2\Delta/e$, to takie samo złącze może służyć do detekcji fal hiperdźwiękowych (rys. 42). Wtedy absorpcja fononów wywołuje wzbudzenie elektronów przechodzących (dzięki zjawisku tunelowania) przez przerwę tworząc prąd, który jest proporcjonalny do ilości fononów czyli natężenia fali hiperdźwiękowej.

Dzięki zastosowanym metodom superdużych częstości bardzo rozwinęły się w ostatnich latach badania ciała stałego metodami akustycznymi (→ Akustyczne zjawiska kwantowe).

Ultradźwięki w całym zakresie częstości mają olbrzymie zastosowanie techniczne, każdy jednak rodzaj



Rys. 42. Półprzewodnikowy model poziomów energetycznych dla złącza utworzonego z dwóch identycznych nadprzewodników: a) Generacja fononów, $eV \geqslant 2\Delta$; energia E_t tunelujących elektronów jest zawarta w przedziale $\Delta \leqslant E_t \leqslant eV-2\Delta$. Maksymalne energie emitowanych fononów relaksacyjnych ($E_{f\,max}$) i fononów rekombinacyjnych (2Δ) są zaznaczone przy odpowiednich wężykowatych strzałkach oznaczających poglądowo akty emisji fononów; b) Detekcja fononów, $0 < eV < 2\Delta$; energia absorbowanego fononu (strzałka wężykowata) przechodzi w energię wzbudzenia tunelowego elektronu

zastosowań związany jest z określonym obszarem częstości.

zastosowanie ultradźwięków

Zastosowanie ultradźwięków może być czynne i bierne. Zastosowanie czynne polega na stosowaniu średnich i dużych natężeń powodujących zmiany nieodwracalne lub częściowo nieodwracalne w nadźwiękawianym ośrodku. Zastosowanie bierne polega na używaniu tak słabych wiązek, że nie wpływają one destrukcyjnie na ośrodek, w którym się rozchodzą, natomiast pozwalają badać jego własności. W czynnych zastosowaniach zmiany w ośrodku związane są też ze zjawiskami nieliniowymi, z których najistotniejsze znaczenie mają kawitacja ultradźwiękowa i przepływy akustyczne.

Powstającej w silnym polu ultradźwiękowym kawitacji polegającej na tworzeniu się i zapadaniu pęcherzyków (il. 158, tabl. 42) towarzyszy wiele efektów wtórnych: procesy chemiczne, luminescencja ultradźwiękowa czyli tzw. sonoluminescencja (il. 159, tabl. 42), procesy biologiczne i inne.

zastosowanie technologiczne

Użycie ultradźwięków dużej mocy pozwala przeprowadzać procesy technologiczne, które innymi metodami są niewykonalne lub bardzo skomplikowane, jak np. obróbka twardych i kruchych materiałów, spawanie materiałów różniących się bardzo własnościami (metale, plastyki), otrzymywanie stopów ze składników różniących się znacznie temperaturami topnienia i gęstościami itp. Ultradźwięki przyspieszają i polepszają takie procesy, jak oczyszczanie, trawienie powierzchni, otrzymywanie zawiesin oraz otrzymywanie emulsji i areozoli.

zastosowanie w medycynie i biologii

Osobną grupę stanowi zastosowanie ultradźwięków w medycynie i biologii. W tej dziedzinie zastosowanie również może być czynne (działanie na tkanki i organizmy żywe, przy czym dzięki możliwościom ogniskowania ultradźwięków działanie takie można precyzyjnie lokalizować), jak i bierne — do celów diagnostycznych. Terapia i diagnostyka ultradźwiękowa uczyniła w ostatnich kilku latach bardzo duże postępy i bardzo rozpowszechnia się jej zastosowanie. Na ilustracji 180 (tabl. 46) podana jest fotografia ultradźwiękowego przyrządu do diagnostyki oka (oftalmografu) polskiej konstrukcji. Na ilustracji 181 (tabl. 46) pokazany jest obraz ultradźwiękowy płodu w łonie matki.

Stosunkowo niedawno (od kilku lat) w medycynie zaczęto stosować z dużym powodzeniem urządzenia oparte na zasadzie Dopplera do obserwacji elementów ruchomych w organizmie. Tą metodą określa się przepływ krwi w naczyniach krwionośnych, określa i wizualizuje ruchy zastawek serca itp.

W zastosowaniu biernym oprócz badań nieniszczących (→ Badania ośrodków za pomocą ultradźwię-

badanie struktury i własności lepkosprężystych

ków) mających na celu określenie struktury ośrodka lub jego defektów istotne jest określenie własności lepkosprężystych wielu materiałów. Do tego celu konieczne jest wykonanie precyzyjnych pomiarów współczynnika tłumienia α i prędkości rozchodzenia się dźwięku c. Na il. 164 (tabl. 43) mamy np. zestaw pomiarowy do wyznaczania prędkości i tłumienia ultradźwięków w cieczach w zakresie częstości 1–60 MHz. Pomiary tłumienia i prędkości dźwięku ważne są w pełnym zakresie częstości od najniższych do najwyższych. W zakresie hiperdźwiękowym np. (powyżej osiągalnej eksperymentalnie granicy 10^{10} Hz dla generacji i odbioru koherentnych fal sprężystych) tłumienie i prędkość wyznacza się metodami pośrednimi, np. przez zjawisko rozpraszania światła Mandelsztama–Brillouina (→ Akustyczne zjawiska kwantowe). Metody te m.in. pozwoliły zaobserwować dyspersje prędkości dźwięku w cieczach w zakresie częstości 10^{10}–10^{11} Hz i uzupełniły w decydujący sposób wyobrażenia fizyków o stanie ciekłym materii, który jak dotąd mniej jest poznany niż stan gazowy i ciało stałe (kryształy).

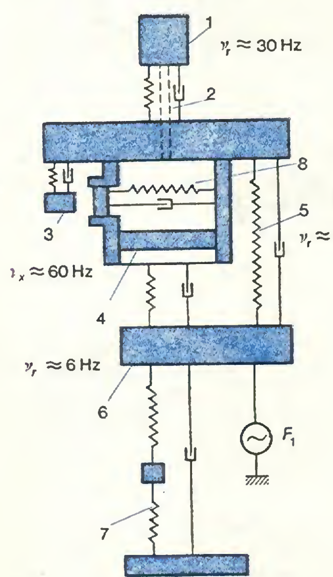
Infradźwięki

Ignacy Malecki

Infradźwiękami nazywamy drgania ośrodka gazowego lub ciekłego o częstościach poniżej słyszalnej. Zwykle przyjmuje się umownie jako zakres infradźwięków pasmo o częstości 0,1–20 Hz. W ostatnich latach zainteresowanie tego typu drganiami bardzo wzrosło, gdyż w środowisku współczesnego człowieka stanowią one ważny czynnik zakłócający.

szkodliwe działanie infradźwięków

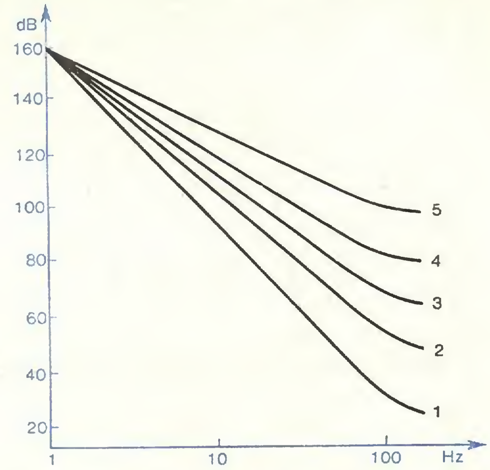
Fale infradźwiękowe działają na cały organizm ludzki. Wywołują one drgania rezonansowe klatki piersiowej, przepony brzusznej i organów trawienia. Schemat zastępczy ciała człowieka pokazuje rys. 43. Powoduje to zaburzenia systemu oddychania, a przy dłuższym działaniu prowadzi do chorób układu trawienia. Infradźwięki mogą też powodować zakłócenia organu równowagi i zmniejszenie ostrości widzenia. Istnieje pewna analogia i addytywność działania infradźwięków i alkoholu, objawiająca się zmniejszeniem szybkości reakcji nerwowych.



Rys. 43. Schemat zastępczy ciała człowieka jako organu drgającego: *1* odpowiada głowie, *2* tchawicy, *3* klatce piersiowej, *4* brzuchowi, *5* kręgosłupowi, *6* miednicy, *7* nogom; F_1F_2 oznaczają siły działające w pozycji siedzącej i stojącej

odczuwanie infradźwięków

Granicę bólu i próg odczuwania wrażeń pochodzących od infradźwięków określa się podobnie, jak dla dźwięków słyszalnych. Im niższa częstość, tym bardziej te dwie granice do siebie się zbliżają, widać to na rys. 44. Ogólnie można rozróżnić następujące zakresy oddziaływania infradźwięków na organizm ludzki.



Rys. 44. Krzywe równej intensywności odczuwania infradźwięków w skali względnej od 1 do 5 (wg Stevensa)

Poniżej 120 dB. W tym zakresie krótkie działanie infradźwięków nie wywołuje wrażeń przykrych i nie jest szkodliwe. Przy dłuższym działaniu wystąpić mogą jeszcze mało zbadane ujemne skutki infradźwięków.

Między 120–140 dB. Przebywanie w polu infradźwiękowym powodować może lekkie zakłócenie procesów fizjologicznych i uczucie nadmiernego zmęczenia.

Między 140–160 dB. Już przy krótkim (2 min) działaniu infradźwięki powodują nieprzyjemne objawy fizjologiczne (zakłócenie zmysłu równowagi, wymioty). Dłuższe działanie spowodować może trwałe uszkodzenie organiczne.

Powyżej 170 dB. Stwierdzono na zwierzętach śmiertelne działanie infradźwięków, spowodowane przeważnie przekrwieniem płuc.

źródła infradźwięków

Źródła infradźwięków podzielić możemy na naturalne i sztuczne. W naturze główną przyczyną powstawania infradźwięków są ruchy powietrza i wody. Falowanie powierzchni mórz i oceanów i prądy podwodne wytwarzają szumy o maksimach leżących w widmie dźwięków słyszalnych, ale wchodzących również w zakres infradźwiękowy. Prócz tego falująca powierzchnia morza jest źródłem fal infradźwiękowych o bardzo niskich częstotliwościach (rzędu 0,2 Hz) rozchodzących się w atmosferze. Inny jest mechanizm powstawania infradźwięków w wodospadach, gdzie rezonans obszaru między skałą a płaszczem wodnym daje czasem wyraźne maksima szumu w zakresie infradźwiękowym.

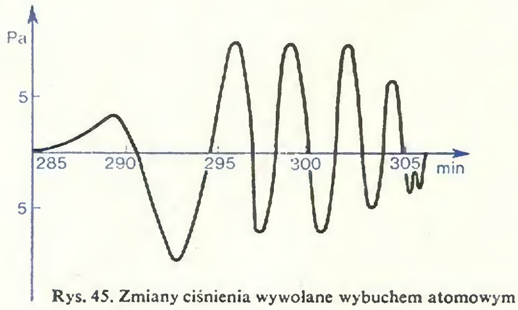
Ruchy górnych warstw atmosfery powodują odbicia fal powstających na powierzchni morza. Wyładowania atmosferyczne są źródłem fali infradźwiękowej towarzyszącej grzmotowi. Wiatr opływający wysokie budynki także generuje fale infradźwiękowe o natężeniu mogącym przekraczać 100 dB.

Wśród źródeł sztucznych najsilniejszymi są wybuchy atomowe lub termojądrowe. Typowy przebieg zmian ciśnienia wywołanych wybuchem atomowym pokazano na rys. 45. Z natężenia fali infradźwiękowej wywołanej wybuchem wnioskować można o wielkości ładunku wybuchowego.

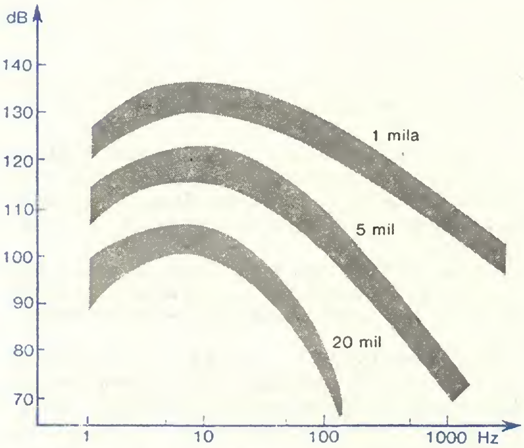
Najgroźniejszym w przyszłości źródłem infradźwięków może stać się transport lotniczy. W lotnictwie

infradźwięki w lotnictwie

poddźwiękowym najhałaśliwsze są helikoptery dające infradźwięki zawierające wyraźne maksima o częstościach odpowiadających liczbie obrotów śmigieł. Samoloty naddźwiękowe przy przekraczaniu bariery dźwięku wytwarzają falę uderzeniową o bardzo dużej amplitudzie. Maksimum przenoszonej energii zależy od wielkości samolotu — pościgowce dają fale o maksimum przy 20 Hz, gdy dla ciężkich samolotów typu Concorde maksimum to przypada przy ok. 2 Hz. Silna fala uderzeniowa powodowana jest również wystrzeleniem ciężkich rakiet. Obserwowano, że przy



Rys. 45. Zmiany ciśnienia wywołane wybuchem atomowym



Rys. 46. Widmo fali powstającej przy odpalaniu rakiety Saturn w różnych odległościach od miejsca wybuchu (odległość podana w milach, 1 mila = 1853 m)

nie sprzyjającym rozkładzie gradientu temperatur w atmosferze, fala podmuchu od rakiety uszkodzić może budynki. Na rys. 46 pokazano widmo fali powstałej przy odpaleniu rakiety Saturn V. Pociągi i ruch drogowy dają lokalne pole infradźwiękowe w pasie ok. 200 m wokół trasy przejazdów.

Na statkach, zwłaszcza na szybkich jednostkach, nieprzyjemne są drgania o częstości infradźwiękowej przenoszone od silników Diesla poprzez konstrukcję statku.

infradźwięki w przemyśle

W przemyśle jednym z głównych źródeł infradźwięków są szybkie przepływy gazów, np. w dmuchawach wielkopiecowych osiągają one poziom 120 dB. Szczególnie uciążliwe mogą być drgania rezonansowe kanałów wentylacyjnych.

Narzędzia udarowe, jak nitownice, młoty pneumatyczne i inne wytwarzają jednocześnie infradźwięki i drgania przenoszące się na ręce robotników. Stwierdzono, że przyspieszenia przekraczają w tych przypadkach wielokrotnie przyspieszenie ziemskie i działają szkodliwie na organizm.

Systematyczne badania źródeł fal infradźwiękowych i ich działania na żywe organizmy datują się zaledwie od kilku lat, toteż wielu zjawisk nie wyjaśniono do końca.

Akustyka molekularna i nieliniowa, Warszawa 1965; R. T. Beyer, S. V. Letcher *Physical Ultrasonics*, New York 1969; J. Blitz *Fundamentals of Ultrasonics*, London 1967; G. H. A. Cole *Dynamika płynów*, Warszawa 1964; *Colloque International sur Infrasons*, Paris 1973; I. E. Elpiner *Ultradźwięki, działanie fizykochemiczne i biologiczne*, Warszawa 1968; F. G. Grawford *Fale*, Warszawa 1973; W. A. Krasilnikow *Zwukowyje i ultrazwukowyje wolny*, Moskwa 1960; M. Kwiek *Akustyka laborytoryjna*, cz. 1, Poznań 1968; M. Kwiek i in. *Akustyka laboratoryjna*, cz. 2, Poznań 1971; L. Landau, E. Lifszic *Mechanika ośrodków ciągłych*, Warszawa 1958; I. Malecki *Podstawy teoretyczne akustyki kwantowej*, Warszawa 1972; I. Malecki *Teoria ośrodków ciągłych*, Warszawa 1958; I. Malecki *Teoria fal i układów akustycznych*, Warszawa 1964; W. P. Mason *Physical Acoustics*, vol. 1–12, New York 1964–1976 (ros. t. 1–7, Moskwa 1967–1974), I. G. Michajłow i in. *Osnowy molekularnoj akustiki*, Moskwa 1964; W. F. Nozdriew *Primienienije ultraakustiki k issledowaniju wieszczestwa*, Moskwa 1958; A. Piotrowska i in. *Ultradźwięki w chemii*, Warszawa 1968; *Proceeding of International Conference on Public Mealth Aspects of Noise*, Dubrownik 1973; C. Puzyna *Kryteria oceny i niektóre sposoby zmniejszania szkodliwego działania drgań na człowieka*; A. Śliwiński, K. Ozimek *Akustyka laboratoryjna*, cz. 3, Warszawa 1974; A. Śliwiński *Chemical Aspects of Ultrasonics* w *Acoustics and Vibration Progress*, vol. 1 (eds. R. W. B. Stephensa i H. G. Leventhall), London 1974; R. Truell i in. *Ultrasonics Methods on Solid State Physics*, New York 1969; J. Wehr *Pomiary prędkości i tłumienia fal ultradźwiękowych*, Warszawa 1972; R. Wyrzykowski *Liniowa teoria pola akustycznego ośrodków gazowych*, Rzeszów 1972; L. K. Zarembo, W. A. Krasilnikow *Wwiedienije w nieliniejnuiu akustiku*, Moskwa 1966.

Badanie ośrodków za pomocą ultradźwięków

Ultradźwięki umożliwiają nieniszczące badanie materiałów. Badaniem makrostruktury materiałów zajmuje się defektoskopia, mikrostrukturę bada się korzystając z metod spektroskopii ultradźwiękowej i mikrodefektoskopii.

Spektroskopia ultradźwiękowa

Jerzy Wehr

Termin spektroskopia przyjął się od dawna w akustyce w odniesieniu do widmowej analizy dźwięków (badanie zależności natężenia od częstości) i do badania z tego punktu widzenia ich źródeł lub kanałów przenoszenia. Termin spektroskopia ultradźwiękowa jest stosunkowo nowy w fizyce. Nadaje mu się przeważnie inne znaczenie, a mianowicie znaczenie analizy zależności współczynnika tłumienia i prędkości fal ultradźwiękowych od częstości drgań lub innej wielkości równoważnej. Tak rozumiana spektroskopia jest analogiczna do różnych rodzajów spektroskopii absorpcyjnej w dziedzinie promieniowania elektromagnetycznego.

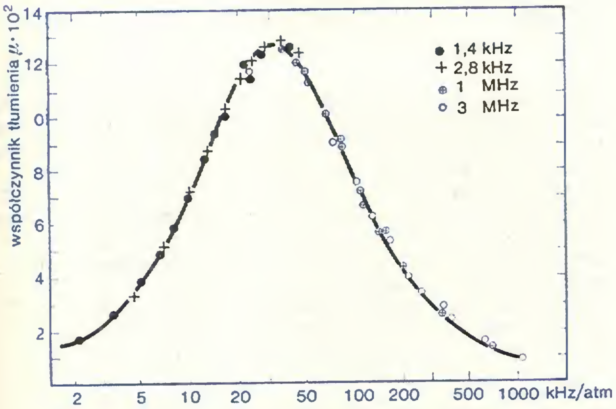
analogia ze spektroskopią absorpcyjną

Tłumienie i prędkość rozchodzenia się fal akustycznych o różnych częstościach, m.in. ultradźwiękowych, zależą od procesów zachodzących w skali atomowej i cząsteczkowej (→ Przedmiot i zakres akustyki, rozdział Akustyczne procesy molekularne) i tym samym mogą informować o ich przebiegu. Mianowicie wystąpienie maksimum tłumienia przy określonej częstości lub wystąpienie obszaru dyspersji prędkości dźwięku (zależności prędkości od częstości drgań) jest wynikiem procesów relaksacyjnych związanych ze

zmianami poziomów energetycznych układów molekularnych.

czas relaksacji Układ (np. zespół atomów w molekule) może zostać wytrącony z równowagi przez biegnącą falę ultradźwiękową w ten sposób, że odbiera on od fali energię (absorbuje) przechodząc w stan wzbudzony. Powrót układu do poprzedniego stanu równowagi wymaga określonego czasu (czas relaksacji), który charakteryzuje dany układ. Energia wzbudzenia oddawana jest przy powrocie do stanu równowagi w postaci ciepła.

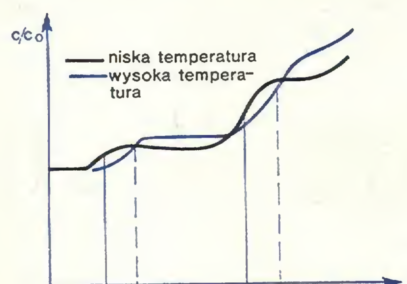
W ośrodku złożonym z wielu molekuł takich procesów pojawia się wiele, a jeśli przy tym są to molekuły jednakowe, to wszystkie te procesy mają ten sam czas relaksacji. Wypadkowy wynik oddziaływania z falą ultradźwiękową objawia się więc dużym selektywnym ubytkiem energii fali i wystąpieniem maksimum współczynnika tłumienia μ_{max} (μ jest bezwymiarowym współczynnikiem tłumienia) dla określonej częstości (rys. 1). Częstość odpowiadająca temu maksimum ν_r,



Rys. 1. Zależność współczynnika tłumienia μ (bezwymiarowego) dwutlenku węgla w temperaturze 21°C od stosunku częstości do ciśnienia fali akustycznej

zwana częstością relaksacji, wyznacza czas relaksacji zgodnie z wzorem $\nu_r \sim 1/\tau$, gdzie τ— czas relaksacji. Częstość ν_r pokrywa się zwykle z częstością odpowiadającą środkowi obszaru dyspersji prędkości dźwięku (rys. 2).

Przebieg zależności współczynnika tłumienia od częstości μ (ν) jest symetryczny w funkcji ln (ν/ν_r) i stosunek częstości, leżących powyżej i poniżej częstości relaksacji ν_r, dla których $\mu = 0{,}5\mu_{max}$, wynosi dla prostego procesu relaksacyjnego ok. 13,9.



Rys. 2. Dyspersja prędkości c (c_0 — prędkość dźwięku dla bardzo małych częstości) i tłumienie relaksacyjne w różnych temperaturach

Pomiary widm relaksacyjnych w funkcji częstości drgań wymagają więc pokrycia bardzo szerokich zakresów częstości, co związane jest z dużymi trudnościami aparaturowymi.

Częstość relaksacji ośrodka gazowego jest funkcją gęstości i temperatury. Gdy gaz stosuje się do prawa gazów doskonałych, zależność od gęstości można zastąpić zależnością od ciśnienia. W spektroskopii ultradźwiękowej stosuje się więc często metodę zmiennego ciśnienia przy ustalonej częstości i z kolei skalę ciśnienia przelicza się na skalę częstości (rys. 1).

Jeżeli procesy relaksacyjne są aktywowane cieplnie, tj. jeżeli drgania cieplne powodują fluktuacje poziomów energetycznych danych konfiguracji mikrostruktury, można zależność od częstości drgań zastąpić zależnością od temperatury T. Najczęściej przyjmuje się w tym celu, że częstość ν_r spełnia równanie Arrheniusa: **równanie Arrheniusa**

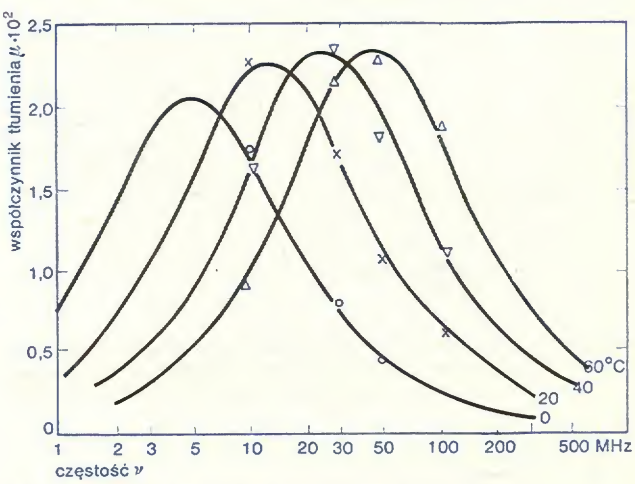
$$\nu_r = \nu_0 e^{-W/kT},$$

gdzie ν_0 jest częstością drgań cieplnych, W — energią aktywacji procesu relaksacyjnego, k — stałą Boltzmanna.

Pomiary współczynnika tłumienia μ_{max} w funkcji temperatury i częstości pozwalają na wyznaczenie energii aktywacji W.

W ciałach stałych realizuje się najczęściej pomiary współczynnika tłumienia μ w funkcji temperatury dla niewielu ustalonych częstości drgań. W cieczach pomiary μ w funkcji temperatury są ograniczone wąskim temperaturowym zakresem ciekłego stanu skupienia i typowe są pomiary μ w funkcji częstości przy ustalonych temperaturach, nie przekraczających temperatur krzepnięcia i parowania.

Rysunek 3 przedstawia przykład przesuwania się maksimum tłumienia na skutek zmian temperatury. Przy wzroście temperatury maksimum przesuwa się



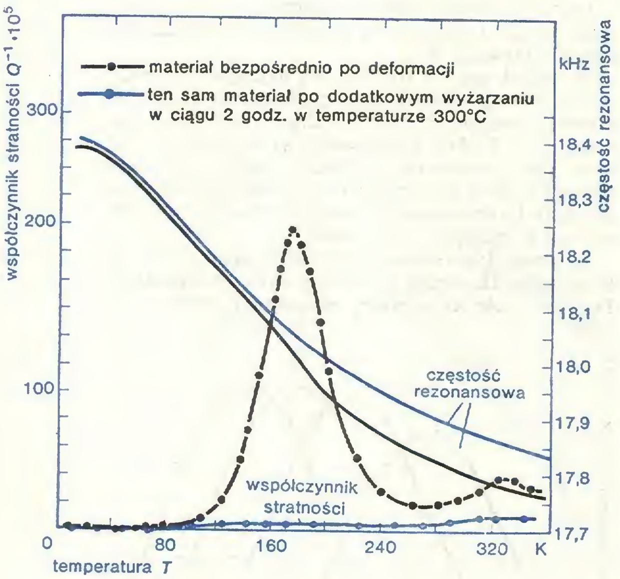
Rys. 3. Tłumienie relaksacyjne uwarunkowane izomerią obrotową w 1,1,2-trójchloroetanie

w stronę wyższych częstości. Obliczenie z tego przesunięcia energii aktywacji oparte na wzorze Arrheniusa pozwoliło ustalić, że wykryty tutaj eksperymentalnie proces relaksacji w 1,1,2-trójchloroetanie jest związany z przechodzeniem molekuł z jednego stanu izomerycznego (o niższej energii) w drugi (o wyższej energii).

Jeżeli bezwymiarowy współczynnik tłumienia fal akustycznych μ rośnie w funkcji częstości ν wolniej niż pierwsza potęga ν (szybszy wzrost μ (ν) zdarza się przeważnie w wypadku rozproszenia fal na wtrąceniach, granicach ziaren itp.), to tłumienie można interpretować jako skutek procesów relaksacyjnych w badanym ośrodku i można zaproponować dyskretny lub ciągły model (widmo) takich czasów relaksacji i takich stratności materiału, który zgadza się z wyni-

stratność materiału

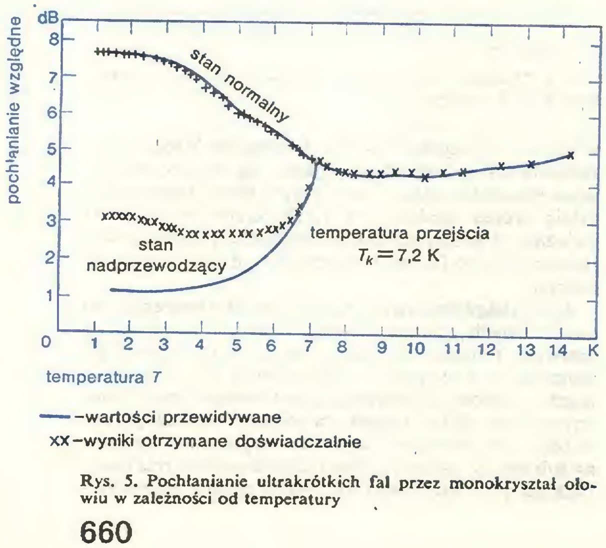
kami pomiarów (stratność materiału $\varepsilon = \mu/\pi = Q^{-1}$, gdzie Q jest współczynnikiem dobroci materiału). Należy jednak podkreślić, że — po pierwsze — taki model odnosi się tylko do zakresu częstości objętego pomiarami, a po drugie — może istnieć często wiele różnych widm relaksacyjnych odpowiadających tym samym wynikom pomiarów. Odtworzenie widma relaksacyjnego ośrodka z pomiarów propagacji fal ultradźwiękowych w ośrodku jest więc niezupełne, a gdy widma są bardziej złożone (gęste) — jest też niejednoznaczne. Modele widm relaksacyjnych łatwiej jest tworzyć, jeśli przebieg μ w funkcji ν/p (gdzie p — ciśnienie) czy T dzieli się na wyraźne maksima (rys. 2). W wyborze modelu własności relaksacyjnych są pomocne, oprócz pomiarów tłumienia, również pomiary prędkości propagacji fal ultradźwiękowych c. Dla prostego procesu relaksacyjnego blisko maksimum μ leży punkt przegięcia (środek obszaru dyspersji) prędkości fazowej i ekstremum prędkości grupowej. Obserwacja charakterystycznych punktów przebiegów prędkości pomaga rozdzielić dyskretne widmo relaksacyjne na właściwą liczbę składników.



Rys. 4. Wpływ wyżarzenia na widmo relaksacyjne polikrystalicznego niobu

zastosowania

Przykładami zastosowań spektroskopii ultradźwiękowej ciał stałych są: badania dyfuzji atomów gazu w metalach, wyznaczanie gęstości dyslokacji i średniej długości pętli dyslokacyjnych, badanie koncentracji defektów punktowych i domieszek w sieci krystalicznej (rys. 4). Spektroskopię ultradźwiękową stosuje się również do badania obszaru przejścia fazowego od stanu normalnego do stanu nadprzewodzącego (rys. 5). Tego rodzaju zmiany obserwuje się wyraźnie tylko w bardzo czystych materiałach. Typowymi zastosowaniami w badaniach cieczy i gazów są: badanie izomerii (rys. 3), badanie przejść monomer–dimer, pomiary małych koncentracji domieszek i składu mieszanin, wyznaczenie prędkości reakcji chemicznych, badanie dysocjacji roztworów. Niekiedy celowe jest łączenie metod spektroskopii ultradźwiękowej z innymi metodami spektroskopii, np. z metodą jądrowego rezonansu magnetycznego czy elektronowego rezonansu spinowego (→ Akustyczne zjawiska kwantowe).



Rys. 5. Pochłanianie ultrakrótkich fal przez monokryształ ołowiu w zależności od temperatury

Defektoskopia i mikrodefektoskopia ultradźwiękowa

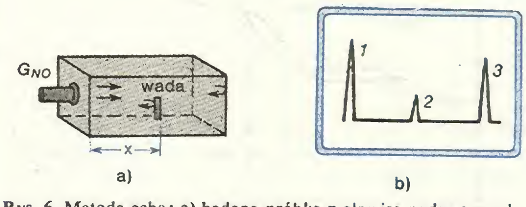
Bogumił Linde

badanie makrostruktury i mikrostruktury

W defektoskopii i mikrodefektoskopii wykorzystuje się wpływ niejednorodności materiału na prędkość i tłumienie rozchodzących się w nim ultradźwięków. Badanie materiału polega na wprowadzeniu do niego fali ultradźwiękowej i na jej odebraniu. Do wytwarzania i rejestrowania ultradźwięków służą przetworniki piezoelektryczne (wielkie częstości) oraz magnetostrykcyjne (małe częstości). Z czasu przelotu fali w materiale i z jej tłumienia lub wzmocnienia można wnioskować o rozmieszczeniu i wielkości niejednorodności. Od częstości użytej fali zależy, jakiej wielkości defekty można w ten sposób wykrywać. Np. rozmiary ziaren można badać stosując fale o wielkich częstościach, gdyż ich tłumienie zależy bardzo silnie od wielkości ziaren. Do badania makrostruktury materiałów, tj. niejednorodności, pęknięć, wytrzymałości, twardości itp. stosuje się fale ultradźwiękowe o częstości 10^5–10^8 Hz. Do badań mikrostruktury, tj. defektów punktowych sieci krystalicznej, rozmiarów ziaren, dyslokacji, załamania lub przesunięcia płaszczyzn sieci itp. stosuje się ultradźwięki i hiperdźwięki o częstości 10^7–10^{11} Hz.

metoda echa

Istnieją trzy podstawowe metody badań stosowane zarówno w defektoskopii jak i mikrodefektoskopii. Najbardziej rozpowszechniona jest metoda echa. Polega ona na nadawaniu krótkich impulsów ultradźwiękowych i ich odbiorze po odbiciu od niejednorodności w badanej próbce. Odległość defektu określa czas liczony od chwili nadania impulsu do jego powrotu po odbiciu (rys. 6), zaś wielkość tego defektu obrazuje amplituda impulsu odbitego. Impulsy obserwuje się

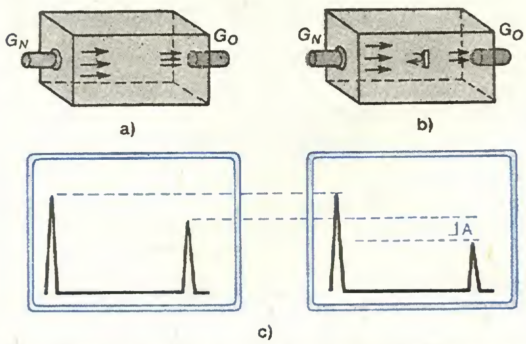


Rys. 6. Metoda echa: a) badana próbka z głowicą nadawczo-odbiorczą G_{NO}, x oznacza odległość wady od powierzchni; b) oscylogram, 1 impuls nadany, 2 impuls odbity od wady, 3 impuls odbity od końca próbki

zwykle na ekranie lampy oscyloskopowej, znając zaś jej podstawę czasu można odczytać odległość czasową impulsu nadanego i odbitego, a stąd znaleźć odległość wady od powierzchni z prostego wzoru: $2x = ct$, gdzie x jest odległością wady od powierzchni, c — prędkością ultradźwięków w badanym materiale, t — czasem przelotu.

metoda cienia

Metoda cienia (przepuszczenia) polega na pomiarze amplitudy fali ultradźwiękowej za wadą po przejściu przez badany materiał (rys. 7). Używa się w niej dwóch głowic — nadawczej i odbiorczej. Niejednorodności w badanym materiale powodują osłabienie energii fal docierających do głowicy odbiorczej. Z różnicy amplitudy impulsu nadanego i odebranego wnioskuje się o wielkości wady. Metodę cienia jak i metodę echa można stosować także do pomiaru grubości.



Rys. 7. Metoda cienia: a) próbka bez wady, b) badana próbka z wadą; G_N głowica nadawcza, G_O głowica odbiorcza, c) oscylogramy z zaznaczoną różnicą amplitud ΔA

metoda rezonansowa

Metoda rezonansowa polega na znalezieniu kolejnych częstości rezonansowych badanego elementu. Zjawisko rezonansu zachodzi wtedy, gdy grubość G badanego przedmiotu jest wielokrotnością połowy długości fali; wówczas $G = c/2\Delta f$, gdzie Δf jest różnicą dwóch kolejnych częstości rezonansowych. Metodę tę stosuje się w pierwszym rzędzie do pomiaru grubości, można ją też stosować do wykrywania rozwarstwień, braków przyczepności w połączeniach spawanych, lutowanych, klejonych, a także istnienia korozji.

metoda drgań własnych

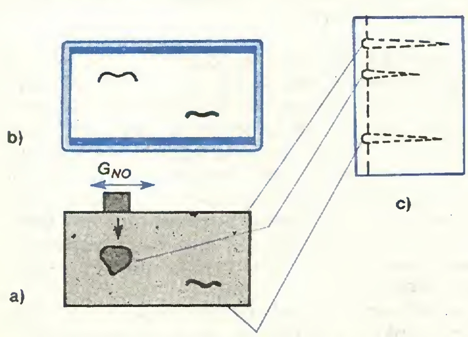
Prócz wymienionych są jeszcze dwie metody defektoskopii stosowane rzadziej. Metoda drgań własnych opiera się na badaniu i analizie drgań własnych układu wzbudzonego uderzeniem; stosuje się ją do badania układów wielowarstwowych (wad samych warstw i połączeń sztywnych międzywarstwowych).

metoda impedancji

W metodzie impedancji wykorzystuje się zależność akustycznego oporu (impedancji) wejściowego układu od obciążenia na jego końcach. Za pomocą tej metody można wykrywać zakłócenia w sztywnych połączeniach cienkiej osłony z masywnym podłożem. Bezpośrednią miarą zakłóceń jest zmiana sił reakcji badanego urządzenia, kóre wzbudzają w czujniku drgania giętne o częstości dźwiękowej.

prezentacja B

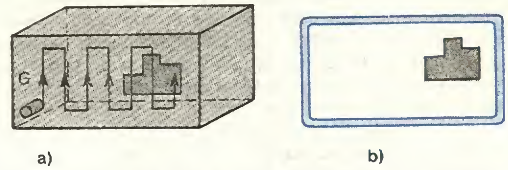
Obecnie rozwijają się metody umożliwiające bezpośrednią obserwację badanych wad. Znajdują przy tym zastosowanie wskaźniki oscyloskopowe (rys. 8), za pomocą których można otrzymać prezentacje A, B (dokładniej omówione w paragrafie Badanie ośrodków biologicznych za pomocą metody echa) oraz C.



Rys. 8. Wizualizacja wad za pomocą prezentacji B: a) badana próbka z głowicą nadawczo-odbiorczą G_{NO}, b) ekran lampy oscyloskopowej, c) obraz jak w metodzie echa

prezentacja C

Przy zastosowaniu prezentacji C (rys. 9) prędkość ruchu plamki podstawy czasu na ekranie jest proporcjonalna do prędkości ruchu głowicy po przedmiocie. Plamka świeci tylko w miejscach wykrytych wad. Przy zastosowaniu odpowiedniej lampy oscyloskopowej obraz może być utrzymany na ekranie dowolnie długo. Pozwala to na dokładną ocenę wad.



Rys. 9. Wizualizacja wad za pomocą prezentacji C: a) badana próbka z głowicą G, b) ekran lampy oscyloskopowej

Ultradźwiękowe metody diagnostyczne

Leszek Filipczyński

Warunki rozchodzenia się fal ultradźwiękowych oraz zjawiska, które ich rozchodzeniu się towarzyszą, zależą od takich makro- i mikroskopowych właściwości ośrodków, jak sprężystość, gęstość, rozciągłość, niejednorodność, anizotropia, budowa molekularna, a w przypadku ruchu ośrodka zależą od jego prędkości. Stąd też wynikają możliwości zastosowania fal ultradźwiękowych jako narzędzia badawczego pozwalającego na uzyskanie o własnościach materii martwej i żywej oraz o jej ruchu cennych informacji, które w inny sposób zdobyć bardzo trudno, a niekiedy w ogóle nie można. W przypadku ośrodków biologicznych ultradźwięki umożliwiają nawet wizualizację wnętrza niedostępnego dla oka bez dokonywania krwawego zabiegu, a także — co jest szczególnie cenne — bez żadnych szkodliwych skutków ubocznych.

możliwości zastosowań diagnostycznych

Dzięki temu ultradźwięki znalazły obecnie szerokie zastosowanie diagnostyczne w medycynie, w takich jej dziedzinach jak położnictwo, kardiologia, okulistyka, neurologia, onkologia, chirurgia naczyniowa itp.

Te wielkie możliwości stosowania ultradźwięków w diagnostyce medycznej wynikają z bardzo korzystnych warunków rozchodzenia się fal ultradźwiękowych w tkankach miękkich. Tłumienie fal ultradźwiękowych w tych tkankach jest przeważnie wprost proporcjonalne do częstości w dolnym jej zakresie aż do 100 MHz, podczas gdy w innych ośrodkach występuje szybszy wzrost tłumienia z częstością. Przy częstości 1 MHz współczynnik tłumienia zawiera się w granicach od 0,1 dB/cm (w ciałku szklistym oka) do 3,3 dB/cm (w tkance mięśniowej w poprzek włókien). Z tego to względu do badań jamy brzusznej, gdzie głębokość penetracji dochodzi do 30 cm, stosuje się fale ultradźwiękowe o częstości ok. 2 MHz, natomiast do badania oczu, gdzie głębokość penetracji wynosi ok. 3 cm, stosowane są częstości wyższe, dochodzące do 20 MHz.

rozchodzenie się ultradźwięków w tkankach miękkich

W tkankach miękkich rozchodzą się praktycznie jedynie podłużne fale ultradźwiękowe, fale poprzeczne są bardzo silnie tłumione; w tkance kostnej mogą się rozchodzić również inne typy fal ultradźwiękowych — poprzeczne, giętne itp. Ponieważ prędkość rozchodzenia się fal ultradźwiękowych w tkankach miękkich wynosi ok. 1500 m/s zatem w podanym wyżej zakresie częstości długości fal ultradźwiękowych wynoszą odpowiednio 0,75–0,075 mm, czyli przekraczają zaledwie o trzy lub dwa rzędy wielkości długość fal świetlnych. Tak krótkie fale można łatwo formować w równoległe lub skupione wiązki, których wymiary poprzeczne są rzędu 1 mm; zapewnia to rozdzielczość wystarczającą do celów diagnostycznych.

W ostatnich latach w mikroskopie ultradźwięko-

wym zastosowano częstości 100 MHz i większe, i uzyskano rozdzielczość odpowiadającą minimalnej odległości 25 µm. Jednakże bardzo duże tłumienie ogranicza możliwość badania w ten sposób tkanek do grubości zaledwie kilku mm.

opór akustyczny tkanek miękkich

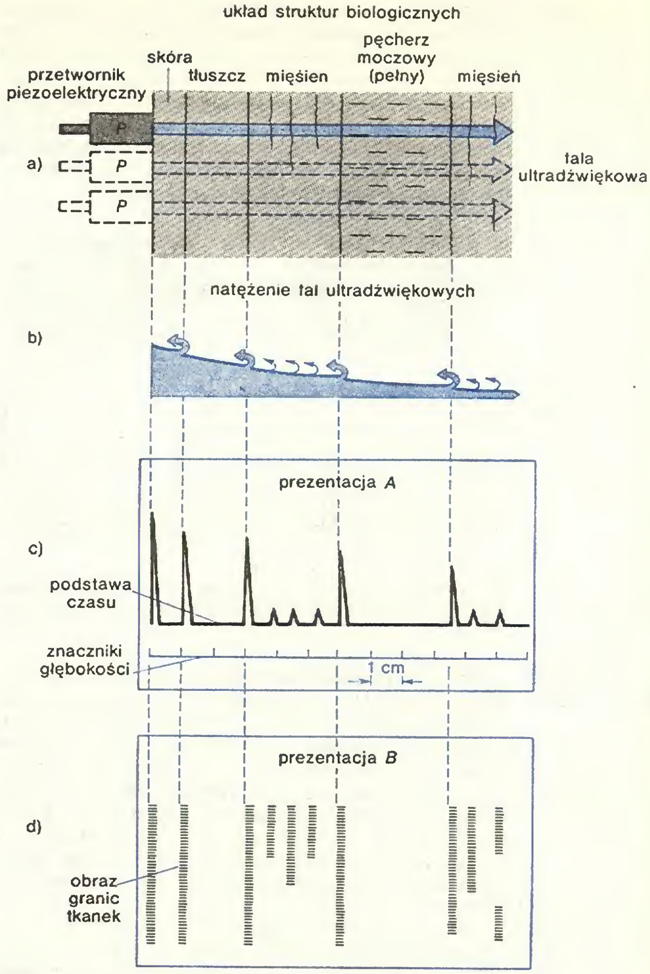
Następnym bardzo korzystnym czynnikiem, który umożliwia zastosowanie ultradźwięków w diagnostyce medycznej, jest fakt, że opór akustyczny właściwy (równy iloczynowi gęstości ośrodka i prędkości rozchodzenia się fali) różnych tkanek miękkich niewiele się różni. Wskutek tego fala ultradźwiękowa padająca prostopadle na granicę dwóch tkanek prawie całkowicie przez nią przenika, zaledwie ok. 1% energii fali padającej odbija się od granicy.

Jeśli się stosuje impulsy fali ultradźwiękowej, to powstaje niewielkie echo, które jest jednak wystarczająco duże, by po wzmocnieniu w układach elektronicznych można je było zarejestrować na ekranie lampy oscyloskopowej. Większość energii fali ultradźwiękowej przenika coraz dalej w głąb ośrodka biologicznego z niewielką stratą energii i daje echa od każdej napotkanej na swej drodze granicy tkanek. Dzięki temu mamy możliwość wykrycia granic wielu różnych, położonych kolejno za sobą tkanek miękkich.

wykrywanie granic tkanek

Przeszkodą nie do przeniknięcia są dla fal ultradźwiękowych wszelkie obszary wypełnione gazami. Opór akustyczny właściwy gazów jest bowiem o kilka rzędów wielkości mniejszy niż tkanek miękkich, i dlatego na granicy tkanka miękka–gaz następuje prawie całkowite odbicie fali ultradźwiękowej. Podobne zjawisko, choć w mniejszym stopniu, zachodzi na granicy tkanek miękkich i kostnych, a znaczna absorpcja fal ultradźwiękowych w tkankach kostnych utrudnia ich badanie.

Informacje zawarte w falach ultradźwiękowych o spenetrowanym ośrodku biologicznym są zakodowane w czasie ich przejścia, amplitudzie, częstości i fazie. Metoda echa, w której się stosuje impulsy fal ultradźwiękowych, pierwsza dała pozytywne wyniki w diagnostyce ultradźwiękowej. Czas przejścia fali i amplituda są podstawowymi parametrami, z których się uzyskuje informacje o ośrodku badanym metodą echa. W ultradźwiękowych metodach dopplerowskich źródłem informacji jest częstość fali, a w ultradźwiękowych układach holograficznych — faza fali.



Rys. 10. Zasada ultradźwiękowej metody echa w zastosowaniu do badania struktur biologicznych: a) badany układ struktur biologicznych, b) natężenie fal ultradźwiękowych, c) prezentacja *A*, d) prezentacja *B*

Badanie ośrodków biologicznych za pomocą metody echa

Informacje uzyskane w postaci echa impulsów fal ultradźwiękowych odbitych od granic tkanek można przedstawić na ekranie oscyloskopowym w różnych prezentacjach. Najprostsza jest prezentacja *A* (rys. 10c), w której impulsy ultradźwiękowe (ściślej — ich obwiednie) odpowiednio przetworzone elektronicznie, przedstawiane są na podstawie czasu, odpowiadającej głębokości ciała.

prezentacja *A*

prezentacja *B*

Bardziej złożoną prezentacją, dającą obraz rozkładu anatomicznego tkanek, jest prezentacja *B*. W tym wypadku sygnały odbite wewnątrz badanego ciała modulują jasność lampy oscyloskopowej; natomiast podstawa czasu wykonuje zazwyczaj złożone ruchy obrotowo-postępowe, analogiczne do ruchów wiązki ultradźwiękowej przeszukującej wnętrze ciała pacjenta. Echa rozjaśniające ekran oscyloskopowy tworzą wówczas rysunek złożonego układu anatomicznego tkanek.

Na rys. 10d przedstawiono zasadę prezentacji *B*, w zastosowaniu do badania struktur biologicznych a na il. 181 (tabl. 46) — przykład jej zastosowania. Obraz oscyloskopowy badanego narządu ciała powstaje stopniowo w miarę przesuwania przetwornika piezoelektrycznego po powierzchni ciała (trwa to zazwyczaj kilkanaście sekund). Dzięki zastosowaniu ekranu oscyloskopowego z długim czasem poświaty albo zastosowaniu układów pamięciowych obraz utrzymuje się wystarczająco długo, by móc postawić diagnozę (w celu uzyskania dokumentacji badania można go również fotografować).

wizualizacja struktur ruchomych

Odmianą prezentacji *B* jest wizualizacja w czasie rzeczywistym: wiązka ultradźwiękowa wykonuje ruchy obrotowe lub postępowe przeszukujące wnętrze badanego ciała z szybkością kilkunastu razy na sekundę. Jest to niezbędne w celu wizualizacji struktur ruchomych serca lub innych narządów. Innym typem prezentacji umożliwiającym rejestrastrację ruchu struktur ruchomych serca jest prezentacja *M*, wykazująca zarówno pewne cechy prezentacji *A*, jak i *B*.

Dopplerowskie metody badania biologicznych struktur ruchomych

Drugą grupę metod ultradźwiękowych w diagnostyce medycznej stanowią metody oparte na zjawisku Dopplera. Fale ultradźwiękowe padając na ruchome struktury biologiczne (jak np. bijące serce płodu lub płynące ciałka krwi), ulegają odbiciu względnie rozproszeniu i zmieniają swą częstość. Zmianę częstości można przetworzyć elektronicznie na zakres słyszalny (pozwala to usłyszeć np. tętno bijącego płodu w łonie matki) lub też można zmierzyć ją (a na tej podstawie wyznaczyć np. prędkość krwi przepływającej w naczyniu krwionośnym).

metoda fali ciągłej

Dwie odmiany metod dopplerowskich znalazły obecnie szerokie zastosowanie. W pierwszej z nich — w metodzie fali ciągłej — przetwornik piezoelektryczny wytwarza falę ciągłą; w drugiej — zwanej metodą

impulsową — wytwarza impulsy fal. Metoda fali ciągłej dostarcza łączną informację o wszystkich strukturach ruchomych leżących na drodze wiązki ultradźwiękowej; powoduje to pewne trudności interpretacyjne, wtedy gdy wiązka ultradźwiękowa trafia jednocześnie np. na tętnicę i leżącą głębiej żyłę. Wady tej nie wykazuje bardziej złożona metoda impulsowa,

metoda impulsowa

która umożliwia np. pomiar prędkości krwi w wybrabranym zakresie głębokości. Dopplerowska metoda impulsowa dostarcza ponadto informacji o czasie przejścia imdulsów ultradźwiękowych w ośrodku, przez co umożliwia jednoczesne dokonywanie pomiarów średnicy naczynia krwionośnego i rozkładu prędkości krwi wewnątrz naczynia. Umożliwia to wyznaczanie prędkości objętościowej krwi (wydatku) w dużych naczyniach krwionośnych. Istnieją również możliwości zastosowania tej metody do badania przepływu krwi wewnątrz serca.

Poważną zaletą ultradźwiękowych metod dopplerowskich jest możliwość dokonywania pomiarów prędkości krwi z powierzchni ciała, bez konieczności dokonywania krwawego zabiegu, choć metody te mogą być również stosowane podczas operacji na odkrytych naczyniach krwionośnych.

Diagnostyka ultradźwiękowa w położnictwie i chorobach kobiecych, L. Filipczyński i I. Roszkowski (red.), Warszawa 1977; L. Filipczyński i in. *Przepływy krwi. Zarys hemodynamiki i ultradźwiękowe metody dopplerowskie*, Warszawa 1980; L. Filipczyński i in. *Ultradźwiękowe metody badań materiałów*, Warszawa 1963; I. Malecki *Fizyczne podstawy akustyki technicznej*, Warszawa 1964, W. F. Nozdriew, N. W. Fedoroszczenko *Moliekularnaja akustika*, Moskwa 1974; Z. Pawłowski *Badania nieniszczące*, cz. 1, Warszawa 1975; W.P. Mason *Physical Acoustics*, vol. 21, part A, New York 1965.

Akustyczne fale powierzchniowe i ich zastosowanie

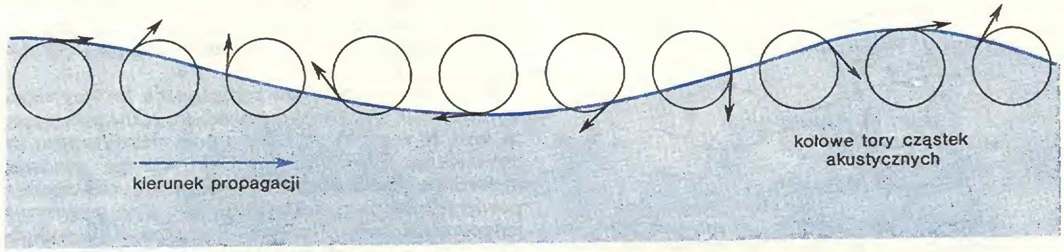
Antoni Śliwiński

Fale powierzchniowe są to fale sprężyste, rozchodzące się na powierzchni ograniczającej ośrodek, który poza tą powierzchnią traktować można jako nieograniczony.

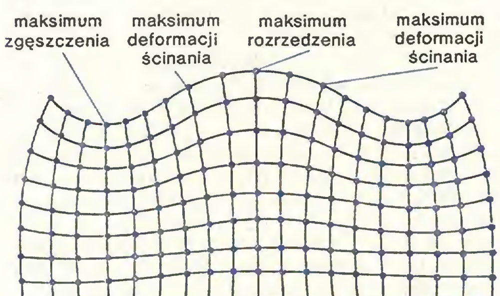
W zależności od charakteru ograniczenia mogą powstawać różne fale powierzchniowe. Wyróżnia się cztery ich rodzaje: fale Rayleigha, Stonleya, Lamba i Love'a. Najbardziej znane z nich są fale Rayleigha

fale Rayleigha

i mówiąc o falach powierzchniowych często je właśnie



Rys. 1. Poglądowy rozkład drgań cząsteczek ośrodka w powierzchniowej fali Rayleigha na zakłóconej powierzchni wody

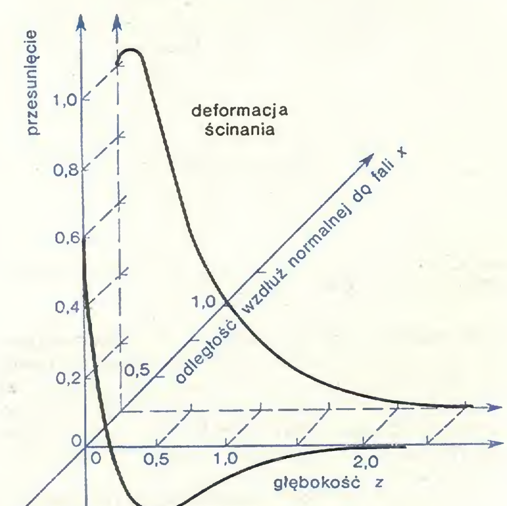


Rys. 2. Modelowe przedstawienie deformacji ośrodka w fali Rayleigha na powierzchni ciała stałego

ma się na myśli. Są to fale rozchodzące się na swobodnej powierzchni cieczy lub ciała stałego graniczącego z próżnią lub gazem — takie jak np. na zakłóconej powierzchni wody (rys. 1) lub na powierzchni ciała stałego (rys. 2). Drgania punktów odpowiadają ruchom po torach eliptycznych (w przybliżeniu kołowych) w płaszczyźnie pionowej, równoległej do kierunku rozchodzenia się fali. W ciałach izotropowych ruch ten można rozłożyć na dwie pary składowych: podłużne i poprzeczne (ścinania) równolegle do powierzchni ciała i normalne do niej (głębokościowe). Składowe głębokościowe przesunięć szybko zanikają i nie sięgają głębiej jak na trochę więcej niż jedna długość fali. Poglądowe przedstawienie współdziałania tych dwóch składowych w powierzchniowej fali Rayleigha zaznaczone jest na rys. 2. Większą amplitudę ma składowa poprzeczna o drganiach prostopadłych

składowa poprzeczna i podłużna

do powierzchni. Składowa podłużna jest równoległa do powierzchni i jej drgania odbywają się równolegle



Rys. 3. Przesunięcie poprzeczne i podłużne cząstki akustycznej w fali Rayleigha jako funkcja głębokości z i odległości wzdłuż kierunku rozchodzenia się x. Jako jednostkę przyjęto długość fali

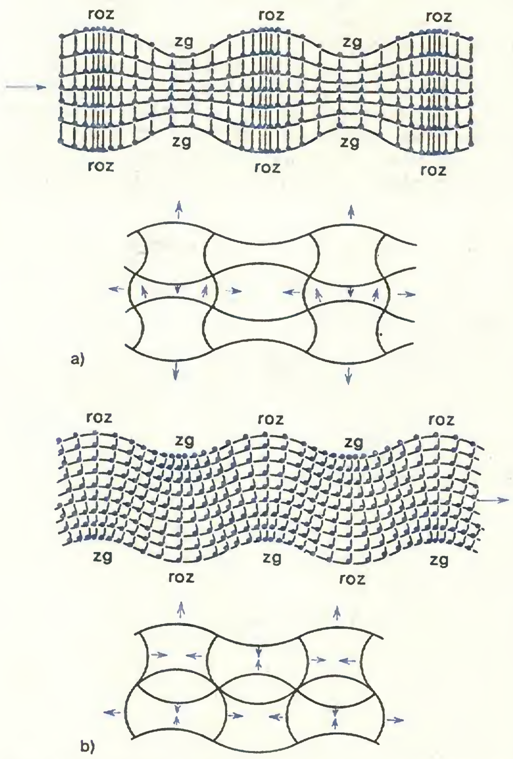
do kierunku rozchodzenia się. Składowe poprzeczna i podłużna są zawsze przesunięte w fazie o 90°. Ten stosunkowo skomplikowany układ przesunięć w formie przestrzennego wykresu przedstawia rys. 3.

fale Stonleya

Falami Stonleya nazywa się fale, które powstają na powierzchni rozdzielającej dwa ośrodki stałe. Mogą one rozchodzić się tylko przy ściśle określonych stosunkach oporów właściwych (→ Przedmiot i zakres akustyki, rozdz. Teoria ośrodków ciągłych i fal sprężystych) obydwu ośrodków. Fale te są bardzo podobne do fal Rayleigha co do rozkładu na składowe podłużną i poprzeczną oraz charakteru ich wnikania w ośrodki po obu stronach powierzchni rozdzielającej te ośrodki.

fale Lamba

Fale Lamba, które często nazywa się płytowymi, powstają w ośrodku stałym ograniczonym dwiema równoległymi powierzchniami (np. w blachach, niegrubych płytach, powłokach itp.), których odległość wzajemna jest porównywalna z długością fali, a nie przekracza kilku długości fali. Przesunięcia cząstek są wynikiem nakładania się na siebie dwóch fal Rayleigha biegnących na obydwu powierzchniach w tym samym kierunku. Fale Lamba rozłożyć można na dwa podstawowe typy: o postaci symetrycznej i antysymetrycznej (rys. 4). W płycie o danej i stałej grubości, której powierzchnie ograniczające są płaszczyznami,

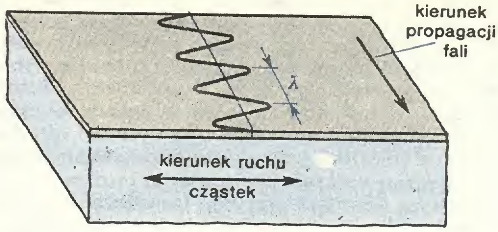


Rys. 4. Fale Lamba: a) postać symetryczna, b) postać antysymetryczna (roz rozrzedzenie, zg zgęszczenie)

może powstać wiele fal o określonych postaciach (podstawowe i wyższe harmoniczne), przy czym o powstaniu danej postaci (sposobu drgań, czyli modu) fali decyduje stosunek grubości płyty do długości fali poprzecznej lub podłużnej. Fale Lamba są falami dyspersyjnymi, to znaczy, że ich prędkość fazowa zależy od częstości.

fale Love'a

Fale Love'a są falami poprzecznymi rozchodzącymi się w układach warstwowych. Cząstki akustyczne drgają w nich równolegle do powierzchni warstwy i prostopadle do kierunku rozchodzenia się. Są to więc fale sprężyste spolaryzowane w płaszczyźnie wyznaczonej przez ograniczenie warstwy. Schematyczny rysunek fali Love'a w cienkiej warstwie spoczywającej na grubszym podłożu (rola podłoża jest tylko pomocnicza — utrzymuje ono warstwę w pozycji poziomej) pokazuje rys. 5. Fale Love'a są silnie dyspersyjne.



Rys. 5. Fala Love'a w cienkiej warstwie umieszczonej na grubym podłożu

Chociaż znane w fizyce od dawna, akustyczne fale powierzchniowe stały się w ostatnim dziesięcioleciu, szczególnie fale o wielkich częstościach, przedmiotem intensywnych badań, ponieważ znalazły zastosowanie w wielu dziedzinach. Stosuje się różne metody wytwarzania takich fal, w zależności od tego, gdzie mają być zastosowane. Fale powierzchniowe o częstościach mikrofalowych zostały wykorzystane w nowej technologii elektronicznej, której rozwój opiera się na układach scalonych (→ Mikroelektronika), w takich urządzeniach jak np. linie opóźniające dla sygnałów elektrycznych, filtry z programowanym przesuwaniem fazy, modulatory, układy do kompresji sygnałów, układy wytwarzające splot dwóch sygnałów (fizyczny odpowiednik splotu w sensie matematycznym), deflektory i modulatory światła (→ Akustyczne zjawiska kwantowe).

Do rozwoju technologii akustycznych fal powierzchniowych przyczyniło się opanowanie konstrukcji przetworników elektro-akustycznych (nadajników i odbiorników fal powierzchniowych) dla dużych częstotliwości.

fale Rayleigha na powierzchniach piezoelektrycznych

W jednym ze sposobów wytwarzania fal Rayleigha wykorzystuje się zjawisko piezoelektryczne polegające na tym, że w wyniku działania pola elektrycznego na powierzchni kryształu piezoelektrycznego powstają naprężenia mechaniczne. Pobudza się elektrycznie powierzchnie piezoelektryczne tak, aby wytworzyć jednocześnie dwie składowe — poprzeczną i podłużną, przesunięte względem siebie w fazie o 90° (rys. 3). Warunki takie można stworzyć na powierzchni płytki piezoelektrycznej odpowiednio wyciętej w stosunku do osi krystalograficznych, przykładając w specjalny sposób pole elektryczne.

Odpowiednio zmienne odkształcenia na powierzchni piezoelektryka wytwarza się za pomocą zmian pola elektrycznego w układzie elektrod (rys. 6) osadzonych na powierzchni kryształu w formie dwóch grzebieniowych, wzajemnie się przenikających układów palczastych. Położenie tych elektrod należy tak dobrać, aby powstające na powierzchni piezoelektryka naprężenia odpowiadały kierunkom tych modułów piezoelektrycznych, które decydują o sprzężeniu elektromechanicznym przy generacji fali powierzchniowej Rayleigha w odpowiednim kierunku. Dla danego kryształu jest to kierunek ściśle określony. Istotne przy wytworzeniu fal powierzchniowych są: składowa E_2 prostopadła do powierzchni i składowa E_3 równoległa do powierzchni. Z rys. 6 widać, że maksimum pola E_2 pojawia się na samych palcach elektrod grzebieniowych, podczas gdy maksimum pola E_3 — pomiędzy nimi. Ponieważ kierunki składowych pola elektrycznego E_2 i E_3 zmieniają się pomiędzy przyległymi parami palców elektrod, powstające składowe deformacji także się zmieniają; przechodzą na przemian od zgęszczeń do rozrzedzeń pomiędzy kolejnymi parami palców elektrod i od poprzecznych drgań skrętnych przeciwnych do ruchu wskazówek zegara do drgań zgodnych z ruchem wskazówek zegara na kolejnych parach palców elektrod.

Pewien dodatkowy przyczynek do procesu generacji fal powierzchniowych Rayleigha w materiałach ferroelektrycznych (a takim jest właśnie $LiNbO_3$) dają efekty elektrostrykcyjne wynikające z istniejących silnych wewnętrznych pól elektrycznych, spowodowanych polaryzacją spontaniczną. Pola te są modulowane przez przyłożone pole elektryczne i wypadkowe naprężenia są wynikiem nałożenia się na siebie efektów piezoelektrycznych i elektrostrykcyjnych. Zjawisko elektrostrykcyjne polega na objętościowym odkształceniu dielektryków w zewnętrznym polu elektrycznym; powstające naprężenia są proporcjonalne do kwadratu natężenia zewnętrznego pola elektrycznego, podczas gdy w zjawisku piezoelektrycznym naprężenia są wprost proporcjonalne do natężenia pola elektrycznego.

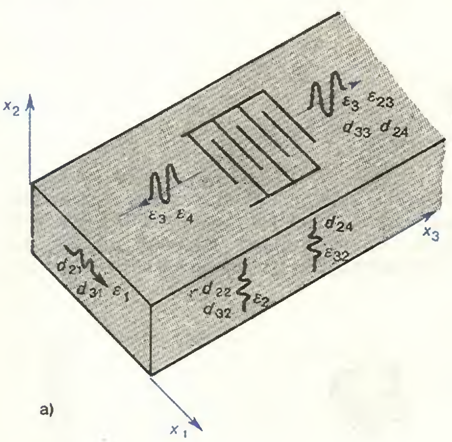
W rozpatrywanym układzie generującym powierzchniowe fale Rayleigha można uzyskać rezonans przy takiej częstości, przy której odległość d (rys. 6b) pomiędzy środkami przyległych palców elektrod staje się równa połówce długości sprężystej fali powierzchniowej, to jest przy:

warunek rezonansu

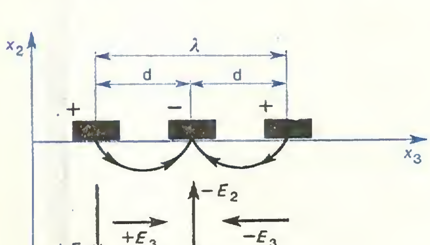
$$f_0 = \frac{v_R}{2d},$$

gdzie v_R — prędkość rozchodzenia się fali Rayleigha.

W warunkach rezonansu składowe odkształceń ε_4 w kierunku oznaczonym w krystalografii jako [2 3] oraz ε_3 w kierunku [3 3] rozchodzą się jako fale powierzchniowe z prędkością v_R od każdego palca elektrody w obydwu kierunkach i docierają do następnych przyległych palców elektrod dokładnie w czasie potrzebnym na zmianę kierunku pól elektrycznych E_2 i E_3. W tych więc warunkach składowe odkształceń propagują się w fazie ze zmianami pól elektrycznych i wytwarzają falę powierzchniową w obydwu kierunkach do palców elektrod. Pozostałe składowe deformacji niepożądane dla generacji fal powierzchniowych propagują się niezależnie od częstości i stąd w warunkach rezonansu są znacznie słabsze, tym niemniej wyraźnie występują.



a)



b)

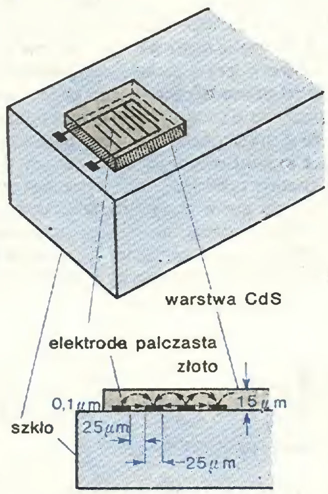
Rys. 6. Metoda wytwarzania fal Rayleigha na powierzchni materiału piezoelektrycznego ($LiNbO_3$) za pomocą elektrod grzebieniowych (palczastych): a) rozkład naprężeń (E_j) i wartości modułów piezoelektrycznych (d_{ij}) w krysztale dla generacji fal Rayleigha w kierunku x_3, b) rozkład pola elektrycznego na powierzchni kryształu $LiNbO_3$

W większości przetworników powierzchniowych dla fal Rayleigha szerokość elektrod palczastych (pojedynczych palców) jest rzędu $^1/_4$ długości fali λ. Gdy częstości są mniejsze niż 300 MHz, długość fali dla zwykle stosowanych materiałów jest większa niż 10 µm i elektrody wykonuje się konwencjonalnymi metodami fotoakwaforty (połączenie techniki fotograficznej z techniką akwaforty). Przy częstościach wyższych (przy 1 GHz, $\lambda \approx 3$ µm) — technika nakładania elektrod jest już specjalna i odbywa się elektronicznie.

Za pomocą opisanej metody można generować fale Rayleigha o częstościach dochodzących do 2 GHz. Fale powierzchniowe można generować również wykorzystując zjawisko magnetostrykcji na powierzchni materiałów magnetycznych, takich jak np. granat itru lub stosując cienkie warstwy magnetyczne na podłożach niemagnetycznych. (Zjawisko magnetostrykcji polega na zmianie kształtu i wymiarów magnetyka podczas magnesowania).

przetworniki cienkowarstwowe

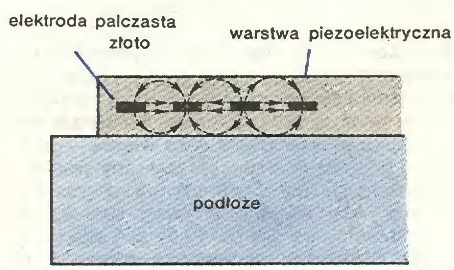
Fale Rayleigha można także wytworzyć w podłożu niepiezoelektrycznym (np. na szkle) pobudzając polem elektrycznym za pomocą elektrody palczastej (rys. 7) cienką warstwę piezoelektryczną, napyloną na powierzchnię podłoża.



Rys. 7. Cienkowarstwowy piezoelektryczny przetwornik fal powierzchniowych Reyleigha (wg J. de Klerka)

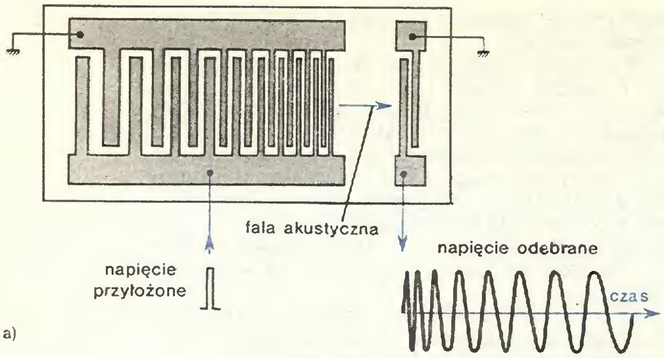
generacja fal Lamba

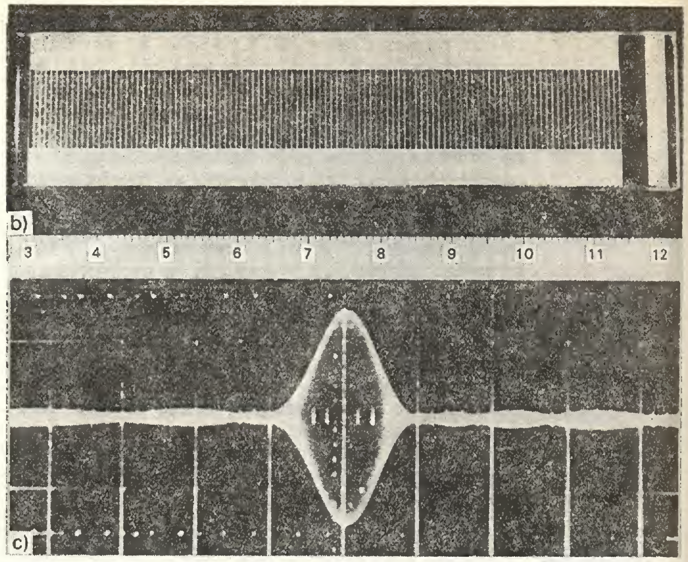
Dzięki odpowiedniej konstrukcji przetworników powierzchniowych można generować fale Love'a i fale Lamba. Jako przykład na rys. 8 przedstawiony jest układ do generacji warstwowej symetrycznej fali Lamba. Palczaste elektrody umieszczone są we wnętrzu (w płaszczyźnie środkowej) naparowanej warstwy piezoelektrycznej. Przyłożenie pola elektrycznego wywołuje symetryczne deformacje warstwy, których rozkład jest zaznaczony na rysunku.



Rys. 8. Przetwornik wytwarzający fale warstwowe Lamba

Jako jeden z wielu możliwych przykładów zastosowań fal powierzchniowych do formowania sygnałów w układach współczesnej elektroniki przedstawiono na rys. 9 filtr dopasowany do sygnału modulowanego





Rys. 9. Filtr dopasowany do sygnału modulowanego częściowo wykorzystujący fale powierzchniowe: a) zasada konstrukcji i działania, b) fotografia filtru dopasowanego do sygnału modulowanego częstościowo z jednym przetwornikiem. Prążki przedstawiają pary elektrod palczastych napylonych na podłożu kwarcowym. Filtr taki używany jest w układach radarowych. Skala jest w centymetrach, c) odpowiedź filtru na dopasowany sygnał. Skala 2 μs na podziałkę

filtry dyspersyjne

częstościowo. Filtr taki działa na zasadzie linii opóźniającej, za pomocą której, dzięki odpowiedniemu zagęszczeniu palców elektrody grzebieniowej, przekształca się np. elektryczny impuls prostokątny (tak jak to pokazano na rys. 9a) na modulowany częstościowo o prawie stałej amplitudzie. Sygnał elektryczny za pomocą pierwszego przetwornika zostaje zamieniony na akustyczną falę powierzchniową, która biegnie z prędkością dźwięku, to jest ok. 10^5 razy wolniej niż sygnał elektryczny biegnący z prędkością światła i zostaje odebrany przez przetwornik drugi, zamieniający go znów na sygnał elektryczny w postaci odpowiednio zmienionej w procesie przetwarzania. Dzięki różnym możliwościom konstrukcyjnym za pomocą takich filtrów (zwanych dyspersyjnymi) można „obrabiać" sygnały elektryczne na wiele bardzo przydatnych w praktyce sposobów, nieosiągalnych innymi metodami. Przykładami zastosowania takiej obróbki sygnałów może być kompresja lub rozszerzenie impulsów w dziedzinie czasu i w dziedzinie częstości, jak również konwolucja czyli uzyskiwanie splotu sygnałów, przekształcanie impulsów z jednej postaci w drugą, kodowanie sygnałów za pomocą impulsów i wiele innych.

E. Dieulesaint, P. Hartmann *Acoustic surface wave filters*, Ultrasonics, 11, 24 (1973); S. Kaliski *Drgania i fale w ciałach stałych*, Warszawa 1966; D. P. Morgan *Surface acoustic wave devices and applications*, Ultrasonics **11**, 121, 211, 255 (1973); A. Śliwiński, E. Ozimek *Akustyka laboratoryjna*, cz. 3, Warszawa 1974; L. A. Viktorov *Rayleigh and Lamb Waves*, New York 1967.

Holografia akustyczna

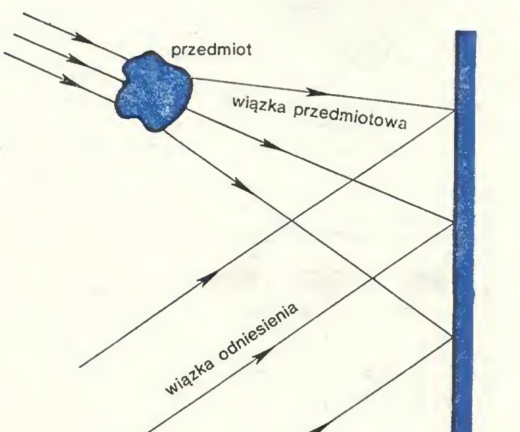
Iwona Wojciechowska

Pod koniec lat 60-ych XX w. powstała nowa metoda przestrzennej wizualizacji wnętrza nieprzezroczystych dla światła obiektów nazwana holografią akustyczną (od greckich słów *hólos* — 'cały', *gràphō* — 'piszę', oraz przymiotnika akustyczny, określającego typ fal służących do tego zapisu).

Podstawowymi zjawiskami fizycznymi, wykorzystywanymi w tej metodzie są, tak jak w holografii optycznej (→ Holografia), dyfrakcja i interferencja. Zasada otrzymywania hologramów jest podobna w obu przypadkach, z tym że zamiast rozkładu natężenia światła w wyniku interferencji, mamy do czynienia w holografii akustycznej z rozkładem natężenia fali dźwiękowej. W holografii akustycznej fale akustyczne wykorzystywane są tylko w pierwszym procesie — tworzenia hologramu, natomiast odtwarzanie odbywa się za pomocą fal świetlnych. Hologram otrzymany w wyniku interferencji dwóch wiązek — przedmiotowej i odniesienia (rys. 1), zapisany jest w ośrodku czułym na zmiany amplitudy lub natężenia fali akustycznej. Aby obraz akustyczny stał się widzialny, odtwarzanie musi się odbywać za pomocą fal świetlnych. Techniczne rozwiązanie procesów rejestracji i odtwarzania w holografii akustycznej polega na odpowiednim wykorzystaniu i dopasowaniu zjawisk oddziaływania ośrodków i światła z falami akustycznymi.

akustyczny zapis, optyczne odtwarzanie

Ze względu na możliwości odtwarzania wnętrza nieprzezroczystych dla światła przedmiotów oraz przedmiotów w nieprzezroczystych ośrodkach, jakie daje holografia akustyczna, zastosowanie jej w różnych dziedzinach, takich jak medycyna, biologia, geologia, geofizyka, oceanologia, hydrolokacja itd., dawałoby szerokie możliwości poznawcze. Przy pomocy



Rys. 1. Rejestracja hologramu

możliwości zastosowania

holografii akustycznej istnieje możliwość badania przedmiotów i struktur o różnej wielkości; zdolność rozdzielcza metody zależy bowiem przede wszystkim od długości fali rejestrującej, a fale akustyczne posiadają długości od metrowych do mikronowych. Długość fali rejestrującej musi być odpowiednio dobrana do rozmiarów obiektu i interesujących w tym obiekcie szczegółów. Im większy obiekt tym większą można stosować długość fali. Najbardziej obiecującą dziedziną zastosowania holografii akustycznej jest diagnostyka medyczna, np. wykrywanie ognisk raka w miękkich tkankach ciała ludzkiego nie osłoniętych przez kości. Poza tym holografia akustyczna znajduje zastosowanie w defektoskopii ultradźwiękowej umożliwiającej nieniszczące badania wewnętrznej budowy różnych ciał. Opierając się na zasadach holografii konstruuje się specjalne urządzenia do podwodnego widzenia, co w przyszłości może znacznie zwiększyć stopień wykorzystania bogactw mórz i oceanów.

Bardzo ważne jest to, że źródła spójnych fal akustycznych znane są i stosowane od dawna, a właśnie od spójności fali przedmiotowej i fali odniesienia zależy wierność odtwarzania w holografii. Fale te nie muszą nawet pochodzić z tych samych źródeł, bo generatory fal akustycznych mogą pracować niezwykle stabilnie. Jak widać, zastosowanie fal akustycznych w holografii ma duże zalety, ale są też poważne trudności polegające na dobraniu odpowiedniego ośrodka

rejestracja fal akustycznych

rejestrującego rozkład ciśnienia akustycznego i układu przetwarzającego obraz akustyczny na optyczny. Od wielu lat bada się oddziaływania fal akustycznych na różne ośrodki — znane są chemiczne, mechaniczne i elektryczne zjawiska zachodzące w materii pod wpływem fal dźwiękowych (szczególnie ultradźwiękowych). Od lat 30-ych XX w. bada się oddziaływanie światła i ultradźwięków (→ Akustyczne zjawiska kwantowe).

Istnieją różne detektory rozkładów pola akustycznego i można je podzielić w zależności od rodzaju zjawisk jakie wywołują w ośrodkach fale akustyczne.

Metody rejestracji pola akustycznego

Metoda	Czułość progowa $W/m^2 \cdot 10^{-4}$
Metody chemiczne:	
bezpośrednie oddziaływanie na błonę fotograficzną	1–5
papier światłoczuły w wywoływaczu	1,0
Metody termiczne:	
zmiana stałej fosforescencji	0,05–0,1
zmiana fotoemisji	0,1
termoczułe barwniki	1
ciekłe kryształy	10^{-10}
Metody mechaniczne i optyczne	
deformacja powierzchni cieczy + przetwarzanie światłem spójnym	10^{-3}–10^{-5}
deformacja powierzchni ciała stałego	10^{-6}
Metody elektronowe:	
przetworniki piezoelektryczne ułożone w mozaikę lub pojedynczy śledzący pole	10^{-11}
elektronowe śledzenie przetwornika piezoelektrycznego (np. kamera Sokołowa)	10^{-7}

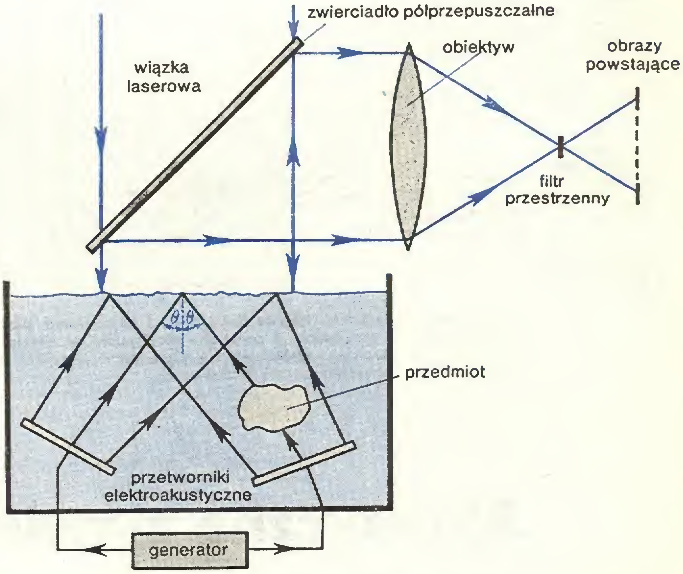
Bardzo ważną cechą detektorów jest czułość progowa, czyli minimalne natężenie ultradźwięków potrzebne do rejestracji obrazu. Poszczególne metody rejestracji mają bardzo różną czułość (tabela). Jak wynika z tabeli największą czułość wykazują metody elektronowe oraz przetworniki akustooptyczne z ciekłym kryształem. Metody elektronowe stosowane są obecnie najczęściej. Trwają natomiast prace nad przetwornikami ciekłokrystalicznymi, które nie mają jeszcze konstrukcji technicznej całkowicie rozwiązanej.

termiczne metody rejestracji

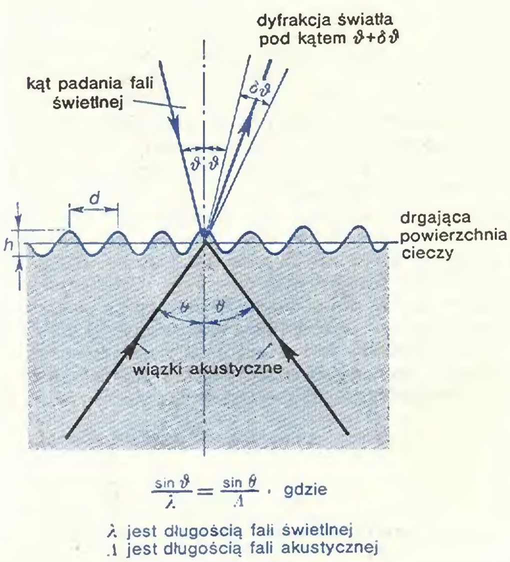
Ogólna zasada rejestracji metodami termicznymi polega na tym, że niektóre substancje organiczne są bardzo czułe na niewielkie zmiany temperatury. Gdy do ośrodka wprowadza się falę ultradźwiękową, to na skutek pochłaniania energii akustycznej występują lokalne zmiany temperatury proporcjonalne do natężenia ultradźwięków. Wraz ze zmianami temperatury zmieniają się lokalnie niektóre własności ośrodków, jak np. zdolność do fotoemisji światła (zmiana barwy przepuszczalnego światła np. w ciekłych kryształach cholesterycznych).

mechaniczne i optyczne metody rejestracji

Już wcześniej stosowane były metody mechaniczne wykorzystujące fakt, że fale akustyczne powodują deformację powierzchni cieczy, wywierając ciśnienie na tę powierzchnię. Schemat układu holograficznego opartego na tym zjawisku przedstawia rys. 2. Rozkład i wysokość zmarszczek na powierzchni cieczy odpowiada rozkładowi natężenia I pola ultradźwiękowego wytworzonego na skutek interferencji wiązki przedmiotowej i wiązki odniesienia. Odległość między zmarszczkami $d = \Lambda/2\sin\theta$, a ich wysokość



Rys. 2. Układ do rejestracji i odtwarzania hologramu ultradźwiękowego, wykorzystujący deformację powierzchni cieczy



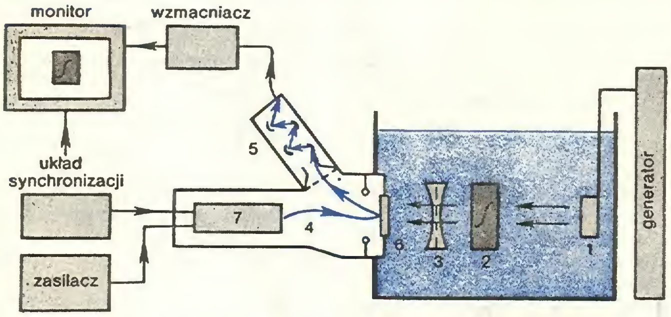
Rys. 3. Schemat wyjaśniający odtwarzanie hologramu akustycznego w metodzie deformacji powierzchni cieczy

$h = I_c/2\pi^2\gamma f^2 \sin^2\theta$, gdzie Λ, λ, c są długością, częstotliwością i prędkością propagacji fali akustycznej, γ — współczynnik napięcia powierzchniowego, θ — kąt między normalną do powierzchni cieczy a kierunkiem

padania wiązki przedmiotowej i odniesienia (rys. 3). Drgająca i zmodulowana powierzchnia cieczy oświetlana jest światłem spójnym, które po ugięciu na powierzchni cieczy i przejściu przez odpowiedni układ optyczny (rys. 2) odtwarza obraz przedmiotu. Sprzężenia układu optycznego z monitorem TV umożliwia osiągnięcie lepszej jakości obrazów oraz badania w tzw. czasie rzeczywistym (tzn., że odtwarzanie obrazu odbywa się równocześnie z rejestracją), co ma ogromne znaczenie m.in. w medycynie.

elektronowe metody rejestracji

Metody elektronowe mają największą czułość spośród wszystkich obecnie stosowanych metod rejestracji i przetwarzania obrazu akustycznego. W tych metodach stosuje się jeden przetwornik śledzący rozkład pola akustycznego lub mozaikę małych przetworników elektroakustycznych (piezoelektrycznych,



Rys. 4. Kamera ultradźwiękowa Sokołowa: *1* przetwornik piezoelektryczny, *2* przedmiot, *3* soczewka akustyczna, *4* wiązka elektronów, *5* powielacz elektronów, *6* płyta piezoelektryczna, *7* układ formujący wiązkę elektronów

elektrostrykcyjnych itp.). Przetworniki zamieniają sygnał akustyczny na elektryczny, sygnał elektryczny przetwarzany jest dalej na optyczny. Jednym z rozwiązań jest kamera Sokołowa (rys. 4). Fala ultradźwiękowa po przejściu przez przedmiot pada na płytę piezoelektryczną, indukując na niej rozkład ładunków, których wartości odpowiadają rozkładowi amplitudy fali. Rozkład ładunków moduluje natężenie śledzącej wiązki elektronowej, zmodulowany sygnał elektronowy po przejściu przez powielacz przekazywany jest na monitor. Czułość większą od kamery Sokołowa posiada układ mozaikowy przetworników, na których napięcia tworzą matryce sygnałów elektrycznych. Ta z kolei przetwarzana jest na matrycę sygnałów optycznych. Urządzenia takie są bardzo złożone pod względem elektronicznym.

kamera Sokołowa

W ciągu ostatnich kilku lat w holografii akustycznej został dokonany ogromny postęp dzięki miniaturyzacji obwodów elektronicznych i jest nadzieja, że holografia akustyczna jako narzędzie badawcze i diagnostyczne wkroczy do nauki i techniki już za kilka lat. Na il. 166 i 167 (tabl. 43) przedstawione są obrazy przedmiotów otrzymane z hologramów akustycznych.

W chwili, gdy czytelnik czyta ten skrótowy opis i ogląda wyniki uzyskiwane za pomocą holografii akustycznej, zapewne w którymś z ośrodków naukowych powstają już znacznie lepsze obrazy od przedstawionych tutaj. Postęp w tej dziedzinie jest bardzo duży i należy się spodziewać szybkiego udoskonalenia metody.

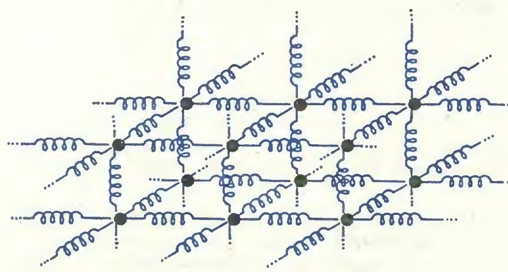
Akusticzeskaja golografija, red. W. G. Prochorowa (tłum. z jęz. ang.) Leningrad 1975; P. ALAIS *Imagerie et holographie ultrasonores*, Rev. de Phys. Appl., suppl. J. de Phys., 11, 559 (1976); A.F. METHERELL i in. *Acoustical Holography*, New York 1969.

Akustyczne zjawiska kwantowe

akustyka kwantowa

Akustyka kwantowa jest bardzo młodym działem akustyki. Chociaż idea kwantowego przedstawienia energii fali akustycznej jest dosyć dawna, sięga lat trzydziestych (wprowadzenie pojęcia fononu przez J. Frenkla w 1932 r.), to dopiero od mniej więcej 15 lat stanowi ona wydzieloną jednoznacznie dziedzinę. Zdecydowały o tym odkrycia bezpośrednich oddziaływań akustoelektronowych (foton–elektron) i akustomagnetycznych (fonon–spin, zjawiska akustycznego rezonansu magnetycznego), których wyjaśnienie wymagało posłużenia się opisem mechaniki kwantowej. Okazało się, że język kwantowy konieczny jest do opisu rozchodzenia się zaburzeń sprężystych w ciekłym helu (akustyczne procesy molekularne), a później w wielu innych procesach, jak np. oddziaływanie fonon–foton czy fonon–fonon. Oddziaływania te są przyczyną powstawania akustycznych procesów nieliniowych (→ Przedmiot i zakres akustyki).

Opis kwantowy jest szczególnie przydatny do wyjaśnienia mechanizmu rozchodzenia się fal sprężystych w ciałach stałych (→ Dynamika sieci krystalicznej). W modelu klasycznym ciała stałego węzły sieci (atomy) są traktowane jako oscylatory sprzężone (można je sobie wyobrazić jako drgające kulki połączone sprężynami; rys. 1, między którymi energia przekazywana jest w sposób ciągły. Model ten jednak okazał się niewystarczający, gdyż w wielu procesach (o których wspomniano wyżej), szczególnie w niższych temperaturach, energia przekazywana oscylatorom ma charakter kwantowy, czyli odpowiada określonym kwantom (porcjom) energii sprężystej.



Rys. 1. Model mechaniczny sieci krystalicznej

W modelu kwantowym ciała stałego poszczególne oscylatory klasyczne zastępuje się kwantowymi, tj. takimi, które mogą być wzbudzone przez dostarczone im kwanty energii. Obrazem sieci krystalicznej z drgającymi atomami jest zbiór takich indywidualnych oscylatorów, w którym ich poszczególne stany wzbudzone są sobie wzajemnie przekazywane jako strumień kwantów energii (fononów). Zbiór tych kwantów, poruszających się swobodnie w sieci krystalicznej, stanowi tzw. gaz fononowy (analogia do gazu elektronowego w przewodnikach).

fonon — kwant energii sprężystej

Leżące u podstaw akustyki kwantowej pojęcie fononu jako kwantu energii fali sprężystej jest analogiczne do pojęcia fotonu jako kwantu energii fali elektromagnetycznej. Jednak foton uważa się za cząstkę elementarną swobodną, natomiast fonon jest tzw. kwazicząstką, czyli przedstawia wzbudzenie elementarne ściśle związane z układem wielu cząstek (np. węzłów sieci krystalicznej, zderzających się cząsteczek cieczy czy gazu). Kwazicząstkę można tylko w przybliżeniu traktować jako swobodną.

Fononowi, podobnie jak fotonowi, przypisuje się energię $E = h\nu$, gdzie h — stała Plancka, ν — częstość drgań, przy czym oba te rodzaje cząstek mają spin całkowity i podlegają statystyce Bosego-Einsteina.

Jest jednak istotna różnica, polegająca na tym, że fotony mają dwie wartości własne spinu odpowiadające dwom stanom polaryzacji (skrętność ± 1), natomiast spin fononów może przybierać wartości -1, 0, $+1$, co odpowiada dwom sposobom drgań poprzecznych o dwóch stanach polaryzacji (skrętność ± 1) oraz drganiom podłużnym (polaryzacja 0).

Fononowi przypisuje się także pęd, który może być określony ściśle jedynie dla węzłów sieci krystalicznej, a traci sens fizyczny dla „pustych" przestrzeni między węzłami sieci. Dlatego traktując fonon jako swobodny, mówimy o kwazipędzie fononu.

fala akustyczna — strumień fononów

Rozchodzącą się falę akustyczną w modelu kwantowym przedstawiamy jako strumień fononów, stąd więc gęstość energii fali $\overline{E}$ wyrazić można: $\overline{E} = Nh\nu = N\hbar\omega$, a następnie natężenie fali $J = \overline{E}c = N\hbar\omega c$, gdzie N — liczba fononów w jednostce objętości, $\hbar = h/2\pi$, ω — częstość kołowa, c — prędkość rozchodzenia się dźwięku.

Akustyczne efekty kwantowe występują wyraźnie w wielu oddziaływaniach, jeżeli energia fononów jest porównywalna z energią innych kwantów, przekazywanych np. przez elektrony, magnony, fotony czy neutrony w procesach zderzeń.

W oddziaływaniach fonon–fonon oddziaływanie jest silniejsze, gdy długość fali akustycznej staje się porównywalna z odległością między węzłami sieci. Warunki takie zachodzą przy bardzo wysokich częstościach, rzędu gigaherców (GHz, powyżej 10^9 Hz), tj. w zakresie hiperdźwiękowym.

Właściwe rozumienie procesów oddziaływań fal akustycznych (zwłaszcza wysokich częstości) z innymi mikroskopowymi obiektami fizycznymi na podstawie akustyki kwantowej umożliwiło wykorzystanie tych oddziaływań. Na przykład oddziaływania fonon–elektron wykorzystuje się w akustoelektronice do bezpośredniego wzmacniania ultradźwięków, oddziaływania fonon–foton w układach pamięciowych, liniach opóźniających (→ Akustyczne fale powierzchniowe i ich zastosowanie), faserach kwantowych.

Oddziaływania fonon–elektron

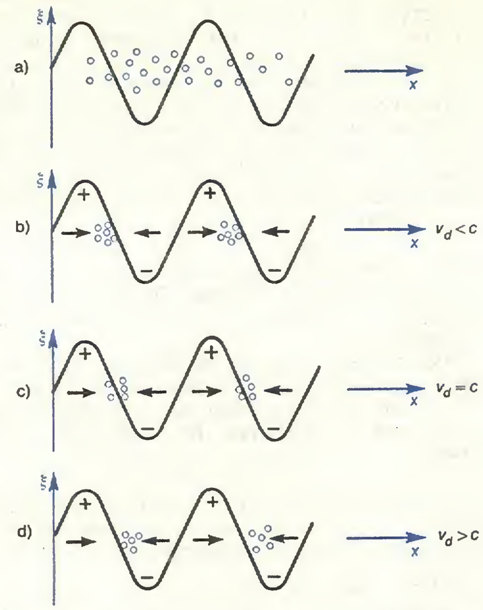
Antoni Śliwiński

Wzajemne oddziaływania fali sprężystej z polem elektrycznym zachodzą głównie w ośrodkach posiadających własności piezoelektryczne. Tego rodzaju własności wykazują kryształy nie mające środka symetrii (należące do niecentrosymetrycznych klas krystalograficznych), lecz posiadające oś biegunową, zwaną osią piezoelektryczną. W takich kryształach może zachodzić zjawisko piezoelektryczne polegające na tym, że deformacje mechaniczne, zwłaszcza wywołane wzdłuż osi biegunowej, powodują powstawanie wolnych ładunków elektrycznych polaryzujących kryształ. Jest to tzw. proste zjawisko piezoelektryczne — w odróżnieniu od odwrotnego zjawiska piezoelektrycznego, polegającego na wystąpieniu deformacji kryształu pod wpływem zewnętrznego pola elektrycznego przyłożonego wzdłuż osi biegunowej. Zjawisko piezoelektryczności jest liniowe i charakteryzują je moduły piezoelektryczne, które są określone jako współczynniki proporcjonalności pomiędzy odkształceniem a powstałym natężeniem pola elektrycznego.

własności piezoelektryczne

Fala ultradźwiękowa rozchodząca się w krysztale piezoelektrycznym w kierunku, w którym moduł piezoelektryczny jest różny od zera, powoduje powstawanie wewnętrznych zmiennych pól elektrycznych. Jeżeli materiał piezoelektryczny jest jednocześnie przewodnikiem lub półprzewodnikiem elektrycznym, to istniejące w nim swobodne nośniki ładunku (np. elektrony lub dziury) podlegają silnemu działaniu tych zmiennych pól elektrycznych (rys. 2).

W dobrym przewodniku nośniki ładunku reagują na działanie pól piezoelektrycznych natychmiast i zupełnie je kompensują. Warunki takie zachodzą wówczas, gdy $\omega_\sigma \gg \omega$, gdzie ω — częstość zmian pola piezoelektrycznego (równa częstości fali ultradźwiękowej), a $\omega_\sigma = \sigma/\varepsilon$, gdzie σ — współczynnik przewod-



Rys. 2. Poglądowe przedstawienie mechanizmu oddziaływania fali sprężystej z nośnikami ładunku w piezoelektryku: a) przy braku sprzężenia, b) przy silnym sprzężeniu i małej prędkości v_d dryfu nośników w polu elektrycznym, c) przy silnym sprzężeniu i prędkości v_d dryfu nośników równej prędkości fali sprężystej c, d) przy silnym sprzężeniu i prędkości v_d dryfu nośników większej od prędkości fali sprężystej c

nictwa elektrycznego, ε — przenikalność elektryczna materiału; można wtedy oczywiście powiedzieć, że materiał jest efektywnie niepiezoelektryczny. Jeżeli jednak w piezoelektrycznym materiale przewodzącym będzie spełniony warunek: $\omega_\sigma < \omega$, to reakcja nośników ładunku na zmienne pole piezoelektryczne będzie znacznie wolniejsza. Taki stan rzeczy bywa często w piezoelektrycznych półprzewodnikach (typu GdS, ZnS) w zakresie wyższych częstości ultradźwiękowych. Wtedy pomiędzy falą ultradźwiękową (strumieniem fononów) a nośnikami ładunków występuje bardzo silne sprzężenie. Jeśli tymi nośnikami są elektrony, to mają one tendencję do gromadzenia się w miejscach, gdzie potencjał elektryczny pola posiada minimum i dzięki temu fala sprężysta może pociągnąć je za sobą (rys. 2b). Jeżeli próbkę umieści się w stałym polu elektrycznym o kierunku zgodnym z kierunkiem rozchodzenia się fali sprężystej, to elektrony zaczną dryfować w tym polu i zmieni się sprzężenie pomiędzy nimi a falą sprężystą. W ten sposób zewnętrzne pole elektryczne pomaga nośnikom podążać za falą, przez co biegnąca fala ulega mniejszemu tłumieniu — traci mniej energii na oddziaływania z elektronami (rys. 2c). Inaczej można powiedzieć, że energia pola zewnętrznego za pośrednictwem nośników ładunku zostaje przekazana fali akustycznej. Przy odpowiednio dużej wartości zewnętrznego pola elektrycznego może nastąpić nie tylko skompensowanie tłumienia, ale nawet wzmocnienie biegnącej fali ultradźwiękowej kosztem energii zewnętrznego pola elektrycznego (rys. 2d).

sprzężenie w półprzewodnikach piezoelektrycznych

oddziaływanie fali ultradźwiękowej z elektronami

Oddziaływanie fali ultradźwiękowej z elektronami (lub z innymi nośnikami ładunku) możemy rozpatrywać jako proces zderzeń fononów i elektronów (lub innych nośników ładunku). Rozpatrzmy najpierw sytuację, w której ośrodek jest półprzewodnikiem i nie ma zewnętrznego pola elektrycznego. Zderzenia są niesprężyste, co oznacza, że przy zderzeniu cała ener-

współczynnik tłumienia

gia fononu zostaje przekazana elektronowi. Z zasady zachowania pędu można obliczyć zmiany prędkości elektronów wynikłe ze zderzeń z fononami; są one równoznaczne z pojawieniem się w półprzewodniku dodatkowego prądu, zwanego akustoelektrycznym. Ubytek energii fononów zużytych na wytworzenie tego prądu objawia się jako tłumienie biegnącej fali sprężystej, co można scharakteryzować za pomocą współczynnika tłumienia α.

Jeżeli z kolei rozpatrzymy sytuację, w której zostało przyłożone pole elektryczne E o kierunku zgodnym z biegiem fali, to stwierdzimy, że na swobodne elektrony w półprzewodniku działa dodatkowa siła, która je wprawia w ruch (dryfowanie) z prędkością proporcjonalną do natężenia pola w kierunku jego działania i wynoszącą:

$$v_d = \mu E,$$

gdzie μ jest ruchliwością elektronów, przy czym $\mu = (e/m)\tau$, gdzie e — ładunek elektronu, m — masa elektronu, τ — średni czas życia elektronów między zderzeniami.

Współczynnik tłumienia fali sprężystej w obecności pola α_E ma inną wartość niż bez pola. Zależność tłumienia fali ultradźwiękowej od pola elektrycznego w procesie oddziaływań fonon–elektron opisuje wzór:

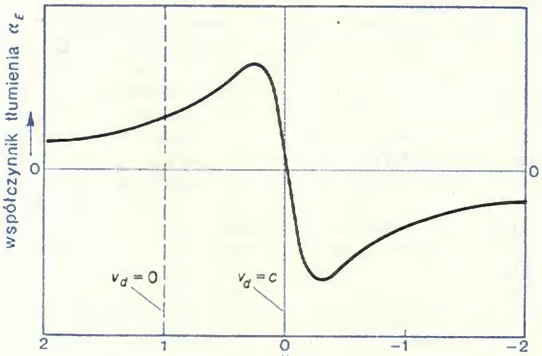
$$\alpha_E = \alpha(1 - v_d/c),$$

gdzie c jest prędkością fali ultradźwiękowej w półprzewodniku. Z tego wzoru wynika, że współczynnik tłumienia w obecności pola elektrycznego α_E może być:

dodatni	$\alpha_E > 0$,	gdy $v_d < c$,	(rys. 1b)
równy zeru	$\alpha_E = 0$,	gdy $v_d = c$,	(rys. 1c)
ujemny	$\alpha_E < 0$,	gdy $v_d > c$.	(rys. 1d)

wzmocnienie fali sprężystej

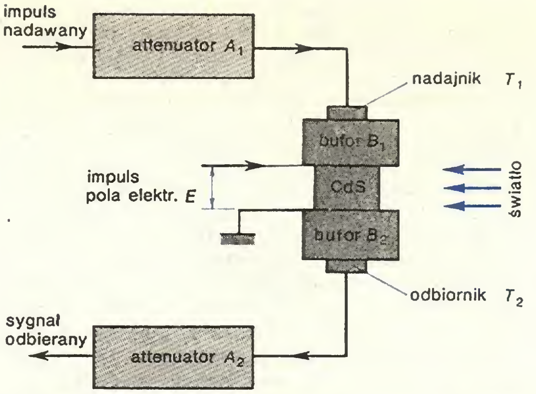
W tym ostatnim przypadku ujemny współczynnik tłumienia oznacza wzmocnienie fali sprężystej kosztem energii zewnętrznego pola elektrycznego. Aby wzmocnienie nastąpiło, pole musi przekroczyć pewną wartość krytyczną $E = E_k$ odpowiadającą $v_d = c$, żeby prędkość dryfu elektronów przekroczyła prędkość fononów c. Na rys. 3 podany jest wykres zależności α_E od $\gamma = 1 - (v_d/c)$.



Rys. 3. Zależność współczynnika tłumienia ultradźwięków α_E od wartości dryfowania γ

Pierwsze doświadczenia, w których uzyskano bezpośrednio wzmocnienie fal ultradźwiękowych na wyżej opisanej zasadzie, przeprowadzili w 1961 r. A. R. Hutson, J. H. McFee i D. L. White w krysztale siarczku kadmu (CdS). Ich układ pomiarowy przedstawia rys. 4.

Przetworzony przez przetwornik T_1 sygnał ultradźwiękowy, którego amplitudę można regulować potencjometrem decybelowym (attenuatorem A_1), przechodzi przez pręt buforowy do kryształu CdS. Sygnał jest odbierany przez przetwornik T_2, przy czym jego wzmocnienie można również regulować, attenuatorem A_2. Na powierzchniach kryształu CdS znajdują się elektrody, pozwalające na przykładanie zewnętrznego pola elektrycznego E. Kryształ CdS



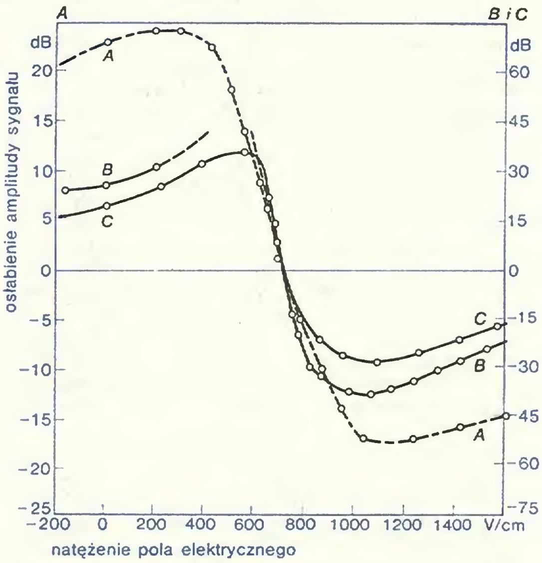
Rys. 4. Schemat układu do obserwacji wzmocnienia fal ultradźwiękowych w CdS zastosowany przez A. R. Hutsona, J. H. McFee i D. L. White'a

jest fotoczuły, co można dodatkowo wykorzystać do regulacji stężenia nośników ładunku w krysztale oświetlając go światłem o różnym natężeniu. Można więc regulować współczynnik przewodnictwa σ.

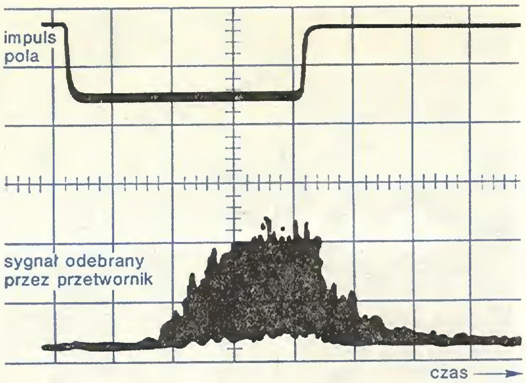
przykład — absorpcja w CdS

Na rys. 5 podana jest zależność (uzyskana przez wyżej wspomnianych autorów) względnego osłabienia poziomu amplitudy sygnału (w dB), który przeszedł przez próbkę CdS, od przyłożonego pola elektrycznego E (pola powodującego dryf elektronów). Poziomem odniesienia krzywych przedstawionych na rysunku są wartości, jakie się uzyskuje dla kryształu nie oświetlonego (gdy $\omega_\sigma = \sigma/\varepsilon = 0$). Widać, że gdy $E < 700$ V/cm, absorpcja jest dodatnia i wzrasta ze wzrostem pola do momentu, w którym prędkość dryfu elektronów v_d staje się porównywalna z prędkością fali ultradźwiękowej ($E \approx 700$ V/cm). Wtedy absorpcja maleje, staje się równa zeru i zmienia znak na przeciwny. Gdy $E > 700$ V/cm, absorpcja jest ujemna, czyli fala ultradźwiękowa ulega wzmocnieniu.

Na rys. 6 jest zapis sygnału, jaki można uzyskać w układzie z rys. 4 po stronie odbiorczej, gdy w nieobecności fali ultradźwiękowej przyłoży się krótkotrwały impuls pola elektrycznego $E > E_k$. W rezultacie oddziaływania elektronów z fononami termicznymi powstanie strumień fononów (koherentnych), które tworzą hiperdźwiękową falę sprężystą.



Rys. 5. Zależność absorpcji fal ultradźwiękowych (osłabienie poziomu amplitudy sygnału w dB względem wartości odpowiadającej próbce nieoświetlonej) w CdS od natężenia pola dryfującego: krzywa A dla $\omega_\sigma/\omega = 1,2$ i częstości 15 MHz, krzywa B dla $\omega_\sigma/\omega = 0,24$ i częstości 45 MHz, krzywa C dla $\omega_\sigma/\omega = 0.21$ i częstości 45 MHz



Rys. 6. Powstanie spontanicznego impulsu akustycznego w CdS po przyłożeniu impulsu pola elektrycznego przekraczającego wartość progową; górny przebieg — impuls pola, dolny przebieg — sygnał odebrany przez przetwornik o częstotliwości własnej 15 MHz (przebiegi zaobserwowano na ekranie oscyloskopu)

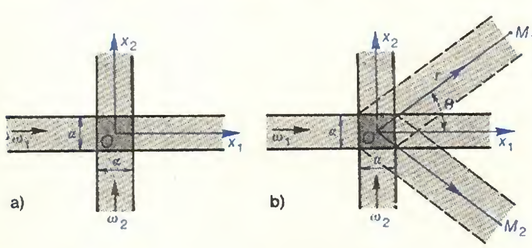
Zjawiska oddziaływania fonon–elektron stały się od kilkunastu lat przedmiotem wielu badań i zastosowań w układach akustoelektronicznych o wzmocnieniu bezpośrednim (→ Akustyczne fale powierzchniowe i ich zastosowanie). Badania są szeroko rozwijane również w Polsce, gdzie zostały opracowane teoretycznie i skonstruowane urządzenia nazwane faserami. Wcześniej były skonstruowane masery, w których zostało wykorzystane zarówno zjawisko akustomagnetyczne, jak i oddziaływanie fonon–elektron.

Oddziaływania fonon–fonon

Antoni Śliwiński

Zjawisko wzajemnego oddziaływania dwóch fal akustycznych (dwóch wiązek fononów) nie występuje przy akustycznych zjawiskach liniowych, przy których obowiązuje zasada superpozycji; przenikające przez siebie fale nie wytwarzają żadnego dodatkowego zaburzenia, które by wychodziło poza obszar wzajemnego przenikania (w tym obszarze może zachodzić interferencja; rys. 7a). Do oddziaływania dochodzi przy zjawiskach nieliniowych, gdy się spotykają fale o odpowiednio dużej amplitudzie. Wtedy w wyniku oddziaływania pojawiają się fale rozproszone (rys. 7b), ich częstości są sumą lub różnicą częstości fal pierwotnych, które się spotykały. Od-

powstawanie fal rozproszonych



Rys. 7. Poglądowe przedstawienie oddziaływania fonon–fonon, czyli rozproszenia dźwięku na dźwięku: a) przenikanie bez wzajemnego oddziaływania, b) przenikanie i wzajemne oddziaływanie — powstaje fala rozproszona

działywanie jest tym silniejsze, im większa jest energia, a więc im wyższa częstość spotykających się fal sprężystych. Przy wyższej energii mogą powstawać także wyższe harmoniczne takich fal kombinacyjnych.

W akustyce kwantowej energię fali akustycznej wyrażamy jako sumę energii fononów, zatem wymienione wyżej procesy możemy rozważać jako wiele aktów elementarnych zderzeń (rozproszeń) fononów jednej fali z fononami drugiej fali. Schematycznie oddziaływania fonon–fonon można przedstawić jak na rys. 8. W zależności od warunków, w których następuje swobodne zderzenie fononów, są dwie możliwości: albo nowy fonon, który powstaje, będzie miał energię równą energii fononów przed zderzeniem, albo równą różnicy ich energii. W procesie oddziaływania musi być spełniona nie tylko zasada zachowania energii, ale również zasada zachowania pędu (pęd fononu $\vec{p} = \hbar\vec{q}$, gdzie wektor falowy $|\vec{q}| = 2\pi/\Lambda$, Λ — długość fali, $\hbar = h/2\pi$, h — stała Plancka. $\vec{p}$ nie jest pędem w ścisłym sensie, ale tzw. pseudopędem lub kwazipędem, bo określony jest jedynie z dokładnością do wektora sieci odwrotnej pomnożonego przez $\hbar$. Fonony swobodne traktujemy jak cząstki prawdziwe, chociaż są to kwazicząstki, a więc związane z określonym układem — zdefiniowane jako wzbudzenia drgań węzłów sieci krystalicznej). Spełnienie zasad zachowania wymaga, aby:

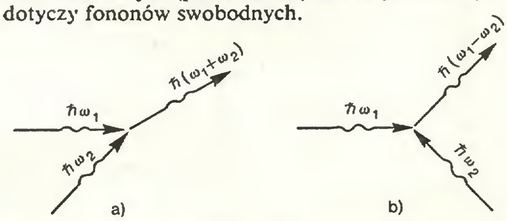
zasady zachowania w zderzeniach fononów

$$\hbar\nu_1+\hbar\nu_2 = \hbar\nu_3 \text{ (zasada zachowania energii)}$$

oraz

$$\hbar\vec{q}_1+\hbar\vec{q}_2 = \hbar\vec{q}_3 \text{ (zasada zachowania pędu)}, \quad (1)$$

gdzie $\vec{q}_1$, $\vec{q}_2$, $\vec{q}_3$ są wektorami falowymi odpowiednich fononów. Powstający w wyniku oddziaływania fonon będzie miał ściśle określony kierunek. W odniesieniu do tych dwu przykładów pokazanych na rys. 8 można powiedzieć, że w pierwszym nastąpiło zwiększenie energii pierwszego fononu o energię drugiego (wzmocnienie fali), natomiast w drugim — zmniejszenie energii pierwszego fononu o energię drugiego (absorpcja, a w następstwie tłumienie fali). Opisany tu proces oddziaływania fonon–fonon nazywa się procesem normalnym (procesem N) lub trójfononowym i dotyczy fononów swobodnych.



Rys. 8. Oddziaływanie fonon–fonon: a) fonon rozproszony zwiększył energię, b) fonon rozproszony zmniejszył energię

W sieci krystalicznej ciała stałego proces oddziaływania fonon–fonon wymaga przy powstaniu nowego fononu uwzględnienia — w myśl zasady zachowania pędu — rekacji kryształu (odrzutu przy wypromieniowaniu fononu), wobec czego trzeba napisać: $\hbar\vec{q}_1+\hbar\vec{q}_2 = \hbar\vec{q}_3+\hbar\vec{G}$, gdzie $\hbar\vec{G}$ oznacza pęd kryształu jako całości; jest to proces czterofononowy zwany procesem U (niem. *Umklapp* 'nakładanie', co w tym wypadku odnosi się do dodania, dołożenia pędu $\hbar\vec{G}$). Spełnienie zasady zachowania pędu wymaga, aby wypadkowy pęd $\hbar\vec{q}+\hbar\vec{G}$ nie wychodził poza pierwszą strefę Brillouina, tj. poza ten obszar w przestrzeni wektora falowego, w którym energia zmienia się w sposób kwasi-ciągły (na granicach tej strefy zmienia się ona skokowo). Strefa Brillouina wiąże się ściśle z odstępami atomowymi w sieci krystalicznej i określona jest przez składowe wektora $\vec{G}$, które np. w wypadku struktury regularnej wynoszą:

proces U

$$G_x = \pm\frac{\pi}{a},\ G_y = \pm\frac{\pi}{a},\ G_z = \pm\frac{\pi}{a}, \text{ gdzie } a \text{ — stała}$$

sieci (→ Wzbudzenia elementarne w ciałach stałych).

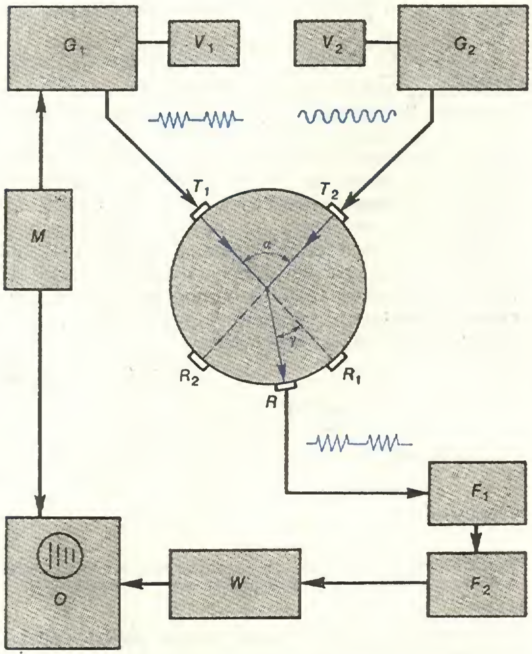
Jeżeli wiązka fal sprężystych pochodzi z jednego źródła, to zwykle stanowi zbiór fononów koherentnych (jest spójna). Przykładem niespójnej wiązki fononów jest strumień ciepła (zbiór fononów termicznych, zwanych też debayowskimi).

fonony koherentne i niekoherentne

Efekty rozpraszania ultradźwięków na ultradźwiękach (fononów koherentnych na fononach koherent-

rozpraszanie fononów koherentnych na fononach koherentnych

nych) w ciele stałym zaobserwowali Su-Fen, L. K. Zarembo i W. A. Krasilnikow w 1962 r., badając powstawanie fali rozproszonej o częstości sumacyjnej. Rysunek 9 przedstawia blokowy schemat doświadczenia. Dwa nadajniki ultradźwiękowe T_1 i T_2 pobudzane do drgań z generatorów G_1 i G_2 promieniują fale ultradźwiękowe o częstościach ν_1 i ν_2, które się przenikają w walcowym bloku ciała stałego (aluminium) pod kątem α. Rozproszoną pod kątem γ falę o częstości kombinacyjnej rejestruje rezonansowy odbiornik R. Odebrany sygnał elektryczny — po przejściu przez filtry F_1 i F_2 i rezonansowy wzmacniacz W — rejestruje się na oscylografie, którego podstawa czasu jest zsynchronizowana za pomocą modulatora M, sterującego również generator G_1. Jeżeli generatory G_1 i G_2 pracują impulsowo, trzeba zagwarantować takie warunki, aby wysłane impulsy spotkały się ze sobą. Wygodniej jest więc pracować z jedną falą ciągłą (jak to zaznaczono na rys. 9),



Rys. 9. Schemat układu do obserwacji rozproszenia dźwięku na dźwięku. G_1 i G_2 generatory drgań, T_1 i T_2 nadajniki ultradźwięków, V_1 i V_2 mierniki napięcia, R_1, R_2 i R rezonansowe odbiorniki F_1 i F_2 filtry, W wzmacniacz, O oscylograf, M modulator

a drugą impulsową. Sygnał rozproszony będzie też impulsowy. Do bloku aluminiowego wprowadzono dwie fale poprzeczne $T(\nu)$ o równych częstościach i w wyniku ich oddziaływania odebrano jako falę rozproszoną falę podłużną $L(2\nu)$ o częstości dwukrotnie wyższej, zgodnie ze schematem:

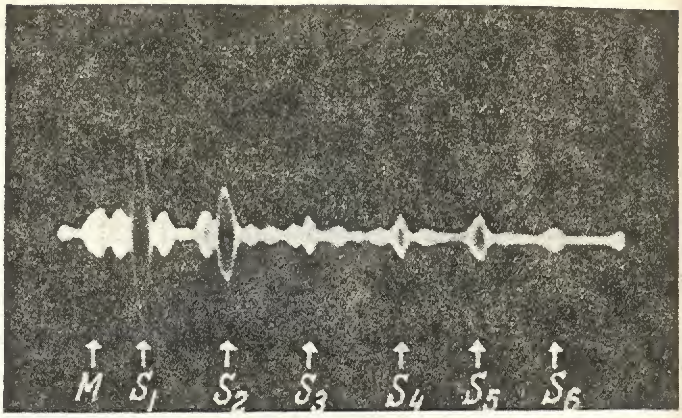
$$T(\nu)+T(\nu) \rightarrow L(2\nu), \qquad (2)$$

odpowiadającym procesowi trójfononowemu określonemu równaniami (1). Na oscylogramie (rys. 10) widać kolejne impulsy sygnałów fali rozproszonej.

Na podstawie zjawisk oddziaływania fonon–fonon, szczególnie w teorii ciała stałego, wyjaśnia się wiele zjawisk fizycznych — jak rozszerzalność cieplna, oporność termiczna sieci, utrzymywanie się równowagi termicznej — przy czym jest to możliwe tylko wtedy, gdy fale sprężyste (fonony termiczne) wytwarzane przez drgania sieci zachowują się nieliniowo.

anharmoniczność sieci

Cecha ta nazywa się anharmonicznością sieci i może być wyrażona przez zależność prędkości dźwięku od naprężeń, co jest równoważne z zależnością częstości drgań fononu $\vec{q}$ o danej polaryzacji (sposobie drgań) od naprężeń (→ Wzbudzenia elementarne w ciałach



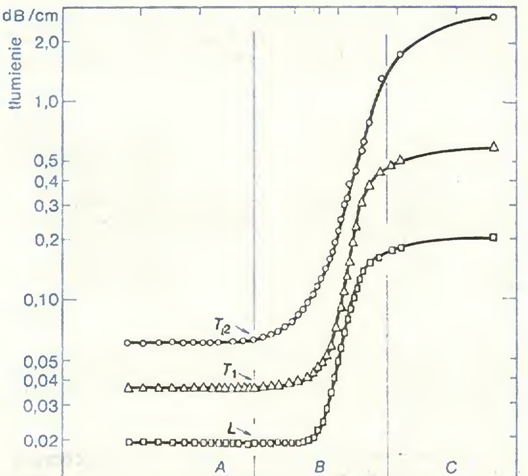
Rys. 10. Oscylogramy drugiej harmonicznej powstałej w wyniku rozproszenia impulsu poprzecznej fali ultradźwiękowej o częstości 3 MHz na ciągłej fali ultradźwiękowej o takiej samej częstości zgodnie z relacją (2). $S_1, S_2, ..., S_6$ kolejne impulsy fali rozproszonej, M sygnał odniesienia

stałych). Mówiąc bardziej obrazowo, im wyższa temperatura, tym bardziej wzrasta amplituda drgań sieci, a przy dużej amplitudzie drgania przestają być harmoniczne i przechodzą w anharmoniczne, czyli takie, w których siły działające przestają być wprost proporcjonalne do wychylenia, a zależą od wyższych jego potęg (są to akustyczne zjawiska nieliniowe).

rozpraszanie fali ultradźwiękowej na fononach termicznych

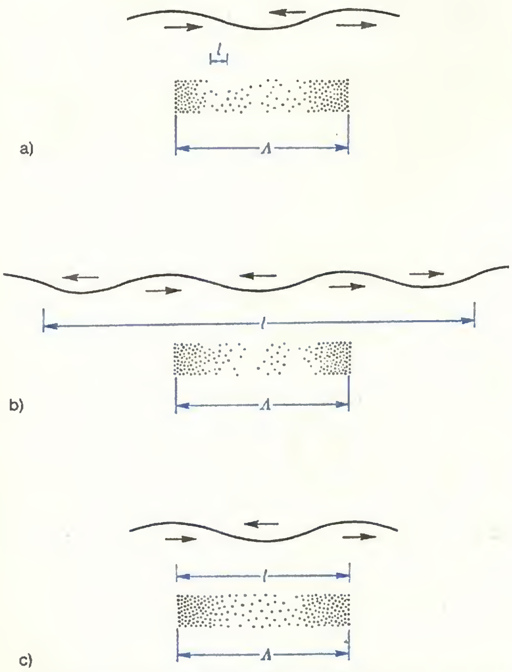
Anharmoniczne oddziaływania pomiędzy fononami w ciałach stałych powodują pochłanianie fal ultradźwiękowych (rozpraszanie fononów fali ultradźwiękowej — koherentnych, na fononach termicznych — niekoherentnych) we wszystkich ciałach stałych, a szczególnie w izolatorach, gdzie nie ma absorpcji pochodzącej od swobodnych elektronów, czyli oddziaływania fonon–elektron.

Rysunek 11 przedstawia typowy wykres tłumienia fali ultradźwiękowej w zależności od temperatury, tłumienia spowodowanego rozproszeniem na fononach termicznych w ciele stałym. Mechanizm rozpraszania fali ultradźwiękowej na fononach termicznych zależy od stosunku ich średniej drogi swobodnej do długości fali ultradźwiękowej. Gdy droga swobodna l fononu jest mała w porównaniu z długością fali, $l \ll \lambda$ (rys. 12a), a $\nu\tau_t \ll 1$, to na fonony wpły-



Rys. 11. Zależność tłumienia ultradźwięków o częstości rzędu GHz w rubinie (Al_2O_3) od temperatury. Obszar A — efekt rozproszenia na fononach termicznych jest do pominięcia; obszar B — tłumienie rośnie jak T^9; obszar C — tłumienie prawie nie zależy od temperatury, T_1, T_2 fale poprzeczne, L fala podłużna

wa tylko niewielki gradient naprężenia spowodowany falą ultradźwiękową. I odwrotnie, gdy $l \ll \lambda$ (rys. 12b), $\nu\tau_t \gg 1$, należy wyobrazić sobie falę ultradźwiękową jako strumień fononów, które biegną razem z fononami termicznymi, przy czym jedyne oddziaływanie polega na zderzeniach. Średnia droga swobodna fononów określona przez zderzenia fonon–fonon jest bardzo silnie malejącą funkcją temperatury przy niskich temperaturach.



Rys. 12. Stosunki pomiędzy drogą swobodną fononu termicznego l i długością fali ultradźwiękowej Λ, która ulega tłumieniu: a) $l \ll \Lambda$, b) $l > \Lambda$, c) $l = \Lambda$

Istnieje taki obszar temperatur, gdzie $l \approx \lambda$ (rys. 12c), a $\nu\tau_t \approx 1$, i wtedy oddziaływania stają się bardzo silne, zderzenia bardzo częste, przy czym procesy zderzeń mają charakter relaksacyjny.

Obszar $\nu\tau_t \ll 1$ nazywamy albo obszarem termoelastycznym, albo obszarem Akhiesera, albo obszarem lepkości fononowej, natomiast obszar, gdzie $\nu\tau_t \gg 1$, nazywa się obszarem Landaua–Rumera. Obszary te rozgranicza obszar pośredni, gdy $\nu\tau_t \approx 1$.

Oddziaływania foton–fonon

Marek Kosmal

Dział fizyki obejmujący zjawiska związane z oddziaływaniami fotonów z fononami, czyli oddziaływaniem fal świetlnych z falami ultra- i hiperdźwiękowymi, nosi nazwę akustooptyki. Praktycznym zastosowaniem efektów akustooptycznych zajmuje się optosonika.

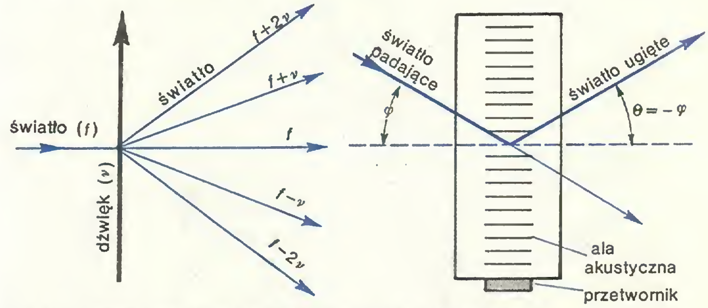
akustooptyka i optosonika

Oddziaływanie fali świetlnej z falą akustyczną rozchodzącą się w ośrodku sprężystym może prowadzić do zjawisk, które nie tylko zależą od rodzaju i cech użytego światła i fal akustycznych (monochromatyczność, zakres częstości, natężenie wiązki, polaryzacja itp.), ale także od rodzaju środowiska (ciała izotropowe, kryształy) i geometrii eksperymentu.

dyfrakcja Ramana–Natha

I tak np. fala akustyczna rozchodząca się w ośrodku powoduje jego sprężyste odkształcenie, a pod wpływem powstałych naprężeń ośrodek ten się zmienia, pojawia się dodatkowa anizotropia (zjawisko fotosprężystości lub efekt elastooptyczny). Na przykład w kryształach normalnie nieczynnych optycznie pojawia się dwójłomność wymuszona. Światło rozchodzące się w płaszczyźnie prostopadłej do wymuszonej osi optycznej kryształu ulega podwójnemu załamaniu i zmienia charakter polaryzacji. Fala akustyczna, rozchodząc się w ośrodku, powoduje pojawienie się okresowego (z okresem tej fali) przestrzennego rozkładu współczynnika załamania światła. Fala świetlna, przechodząc przez taki ośrodek prostopadle lub pod małym kątem do fali akustycznej (rys. 13a), ulega dyfrakcji analogicznie jak na optycznej siatce dyfrakcyjnej; powstaje symetryczny obraz dyfrakcyjny składający się z szeregu maksimów interferencyjnych.

Analogia do optycznej płaskiej siatki dyfrakcyjnej nie jest jednak pełna, gdyż rozkład natężeń w poszczególnych prążkach światła ugiętego zależy od natężenia fali akustycznej; dodatkowo światło ugięte ma zmienioną częstość w wyniku efektu Dopplera. Zjawisko to obserwuje się, gdy długość fal akustycznych jest dużo większa niż długość fal świetlnych; nosi ono nazwę dyfrakcji Ramana–Natha.



Rys. 13. Schemat ugięcia światła na fali ultradźwiękowej: a) ugięcie typu Ramana–Natha, b) ugięcie typu Bragga

akustyczno-optyczna dyfrakcja Bragga

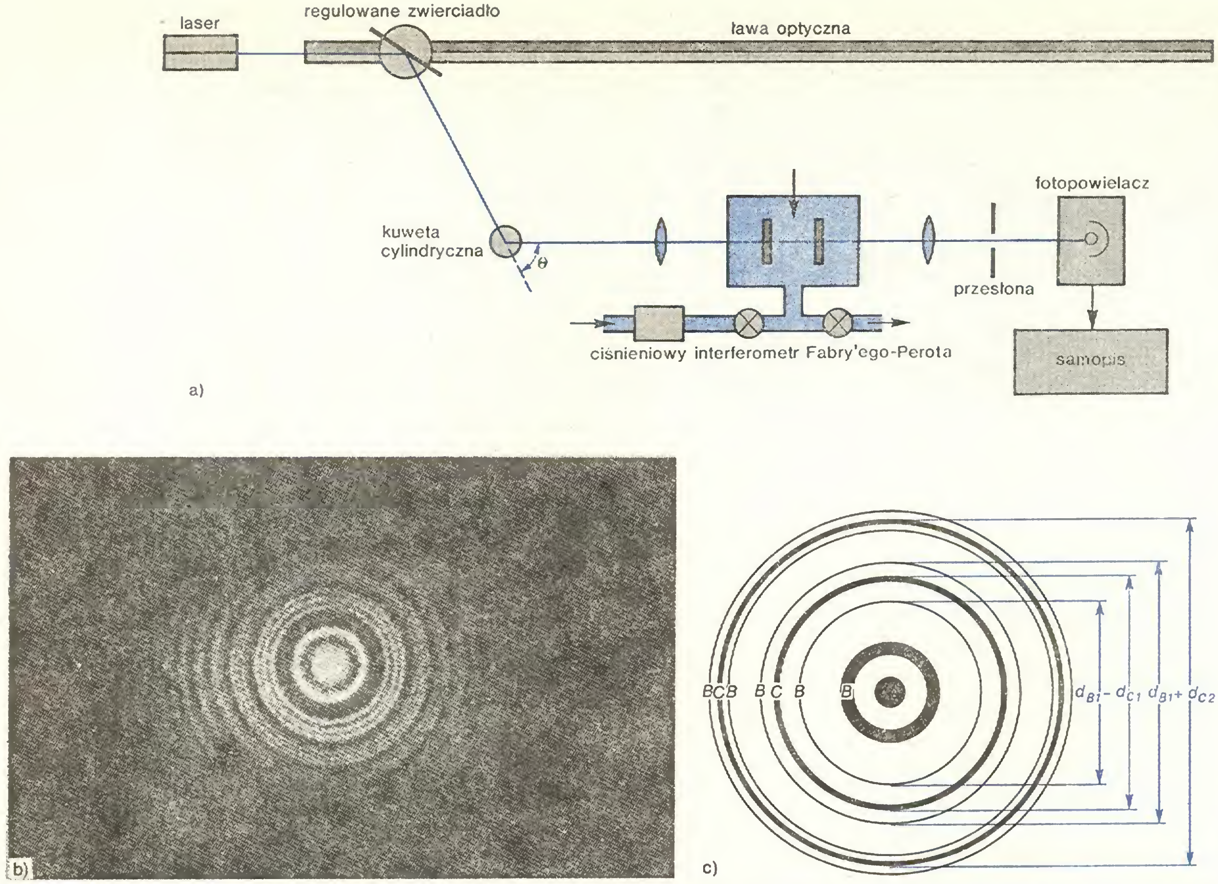
Przy użyciu silnych monochromatycznych wiązek światła (np. promieniowania laserów) na powyższy efekt nakładają się nieliniowe efekty optyczne, prowadzące np. do wytwarzania wyższych harmonicznych światła. W rezultacie ulega zmianie uzyskany obraz dyfrakcyjny — pojawiają się dodatkowe maksima interferencyjne. Przy przejściu fali świetlnej przez pole akustyczne o wysokiej częstości skierowane pod pewnym kątem (kąt Bragga) do fali świetlnej obserwuje się obraz dyfrakcyjny składający się tylko z jednego prążka dyfrakcyjnego. Zjawisko to nosi nazwę akustooptycznej dyfrakcji Bragga (rys.13b).

rozproszenie Mandelsztama–Brillouina

Monochromatyczna wiązka światła, przechodząc przez ośrodek, może także ulec rozproszeniu na niejednorodnościach wywołanych termicznymi fluktuacjami gęstości (cieplne fale akustyczne o częstościach hiperdźwiękowych — fonony termiczne). W zjawisku tym obserwujemy zmianę kierunku rozchodzenia się światła pod kątem Bragga z jednoczesną zmianą częstości (składowa stokesowska i antystokesowska). Zjawisko rozproszenia na fononach akustycznych jest analogiczne do znanego z optyki rozpraszania ramanowskiego na fononach optycznych (→ Wzbudzenia elementarne w ciałach stałych). Natężenie światła rozproszonego jest proporcjonalne do natężenia światła padającego. Rozproszenie tego typu nosi nazwę rozproszenia Mandelsztama–Brillouina (rys. 14).

wymuszone rozproszenie Mandelsztama–Brillouina

Jeśli się stosuje bardzo silne wiązki światła monochromatycznego (np. promieniowania laserów impulsowych), rozproszenie światła na fononach termicznych ma nieco inny charakter. W świetle rozproszonym obserwuje się tylko składową stokesowską, której natężenie nie jest liniową funkcją natężenia fali padającej, a jednocześnie obserwuje się powstawanie fali o częstości hiperdźwiękowej. Fala hiperdźwiękowa powstaje kosztem energii fali świetlnej. Efekt ten nosi nazwę wymuszonego rozproszenia Mandelsztama–Brillouina.



Rys. 14. Rozproszenie Mandelsztama–Brillouina: a) typowy zestaw doświadczalny służący do obserwacji, b) fotografia światła rozproszonego otrzymanego przy użyciu interferrometru Fabry'ego–Pérota, c) schemat analizy częstościowej; B — składowa Brillouinowska, C — składowa centralna

Optosonika

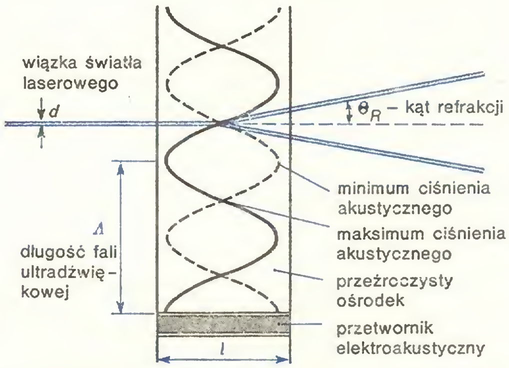
Iwona Wojciechowska

Powstanie tej nowej dziedziny (ok. 1960 r.), wykorzystującej zjawiska akustoopyczne, związane jest z rozwojem techniki wytwarzania różnego typu fal akustycznych, rozszerzeniem zakresu częstotliwości w kierunku MHz i GHz oraz powstaniem źródeł światła spójnego.

Dzięki efektom akustycznym można uzyskać amplitudową, fazową oraz przestrzenną (lub ich kombinację) modulację wiązki laserowej. Metody modulacji i odpowiednio zmodulowane światło spójne znalazły już szerokie zastosowanie w badaniach mikro- i makrostruktury materii, w nowoczesnej elektronice, przy formowaniu impulsów świetlnych, w komunikacji laserowej (→ Lasery — zastosowanie, Optoelektronika półprzewodnikowa), w mikroskopach ultradźwiękowych, holografii akustycznej, rejestracji i odczycie danych pamięci cyfrowych i in. Wykorzystanie zjawisk akustooptycznych do badania materii umożliwia wyznaczenie: stałych fotosprężystych, współczynników sprężystości, lepkości, energetycznych poziomów wzbudzenia termicznego, własności elektrycznych i magnetycznych drobin ośrodka i in. Omówimy teraz przykładowo niektóre sposoby zastosowania zjawisk akustooptycznych.

migawka ultradźwiękowa

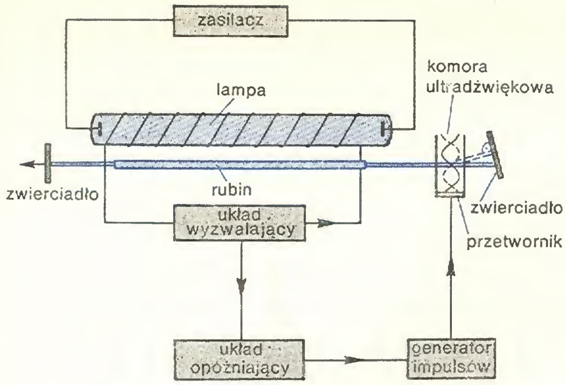
Współczesne zainteresowanie zastosowaniem światła laserowego skierowało uwagę uczonych na sprawę wytwarzania impulsów świetlnych o wielkiej mocy i krótkim czasie trwania (→ Ultrakrótkie impulsy świetlne). Jednym z układów stosowanych do wytwarzania impulsów świetlnych jest migawka ultradźwiękowa, której działanie przedstawiają rys. 15 i 16. Układ ten jest bardziej efektywny i prostszy od rozwiązań mechanicznych, a czas trwania i częstość impulsów laserowych można regulować częstotliwością ultradźwięków.



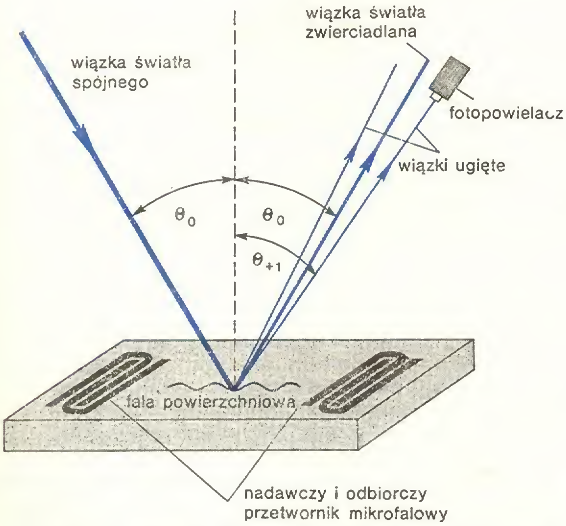
Rys. 15. Refrakcja wiązki świetlnej na fali ultradźwiękowej o długości większej od średnicy wiązki światła $d < \Lambda$. Sinus kąta refrakcji wynosi $\sin\theta_R \approx (2\pi nl/\Lambda)\cos\omega t$, gdzie n jest maksymalną zmianą współczynnika załamania ośrodka ω — częstość kołowa fali ultradźwiękowej

badanie akustycznych fal powierzchniowych

Metoda oparta na zjawisku rozpraszania światła laserowego na akustycznych falach powierzchniowych jest bardzo efektywną metodą pomiaru stałych propagacji fal powierzchniowych w różnych warunkach (różne parametry ośrodka, inny typ przetworników elektroakustycznych itp.). Podstawowy schemat tej metody przedstawiony jest na rys. 17. Pomiar ogranicza się do dokładnego wyznaczenia kątów θ_m



Rys. 16. Schemat układu doświadczalnego — laser rubinowy z migawką ultradźwiękową



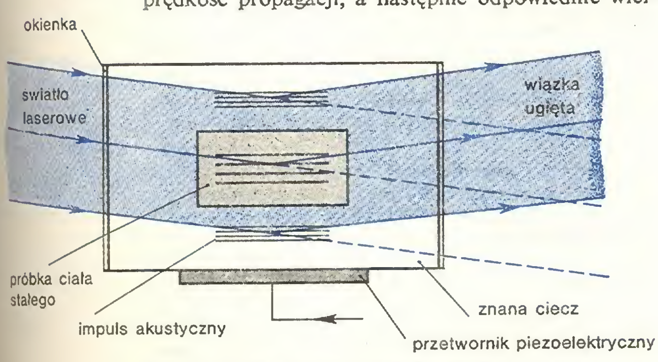
Rys. 17. Badanie fal powierzchniowych; $\sin\theta_m = \sin\theta_0 + m\gamma/\Lambda_s$, $m = 0, \pm1, \pm2, \ldots$ rząd dyfrakcji, λ długość fali świetlnej, Λ_s długość fali powierzchniowej. Jeżeli $\theta_1 = \theta_0$, to $\arcsin \gamma/2\Lambda_s = \theta_B$ (θ_B kąt Bragga)

i rozkładu natężenia rozproszonego światła, z tych wielkości wyznacza się np. prędkość propagacji i tłumienie.

pomiar stałych sprężysto-optycznych ośrodka

Stałe sprężysto-optyczne ośrodków można mierzyć wykorzystując dyfrakcję światła na falach objętościowych. W tym celu stosuje się zarówno dyfrakcję Ramana–Natha, jak i dyfrakcję Bragga.

Rysunek 18 przedstawia schemat układu służącego do pomiaru stałych sprężysto-optycznych ciała stałego. Mierzy się kąt dyfrakcji, na podstawie którego można wyliczyć długość fali ultradźwiękowej, z tego prędkość propagacji, a następnie odpowiednie wielkości charakteryzujące ośrodek (np. moduły sprężystości, przenikalność elektryczną itd.). Tłumienie fali ultradźwiękowej wyznacza się z rozkładu natężenia światła przed i za komorą ultradźwiękową; można w ten sposób otrzymać informacje o wewnętrznej strukturze ośrodka i o procesach zachodzących w ośrodku (→ Badania ośrodków za pomocą ultradźwięków).

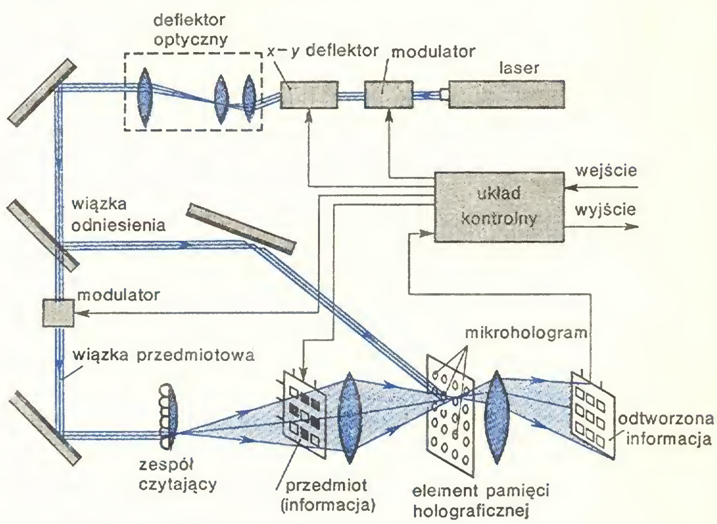


Rys. 18. Schemat układu do pomiaru wielkości sprężysto-optycznych ciała stałego

Zalety optosonicznych metod badania materii: badania te nie niszczą materiału badanego; dają możliwość stosowania niewielkich próbek; śledzenie zachowania się wąskiej wiązki światła umożliwia otrzymanie rozkładów punktowych i badanie nieliniowych własności ośrodków.

ultradźwiękowe modulatory i deflektory światła

Zastosowanie odwrotne (tzn. do znanych ośrodków) wyżej opisanych układów daje możliwość dowolnej modulacji i odchylania wiązki światła (deflektory światła). Modulatory ultradźwiękowe są w nowoczesnej elektronice stosowane do formowania sygnałów elektrycznych pośrednio, poprzez formowanie sygnałów optycznych. Deflektory i modulatory znalazły zastosowanie w optycznych pamięciach maszyn cyfrowych. Pojemność pamięci optycznych (holograficznych) z ultradźwiękowymi deflektorami jest zbliżona do pojemności pamięci innych typów, a prędkość rejestracji i odtwarzania danych jest w nich większa (odpada bezwładność układów mechanicznych). Na rys. 19 przedstawiony jest układ pamięci holograficznej z ultradźwiękowymi modulatorami i deflektorami, służącymi do nadania żądanego przebiegu czasowo-przestrzennego wiązki laserowej. Informacja (obraz przedmiotu) zostaje zarejestrowana



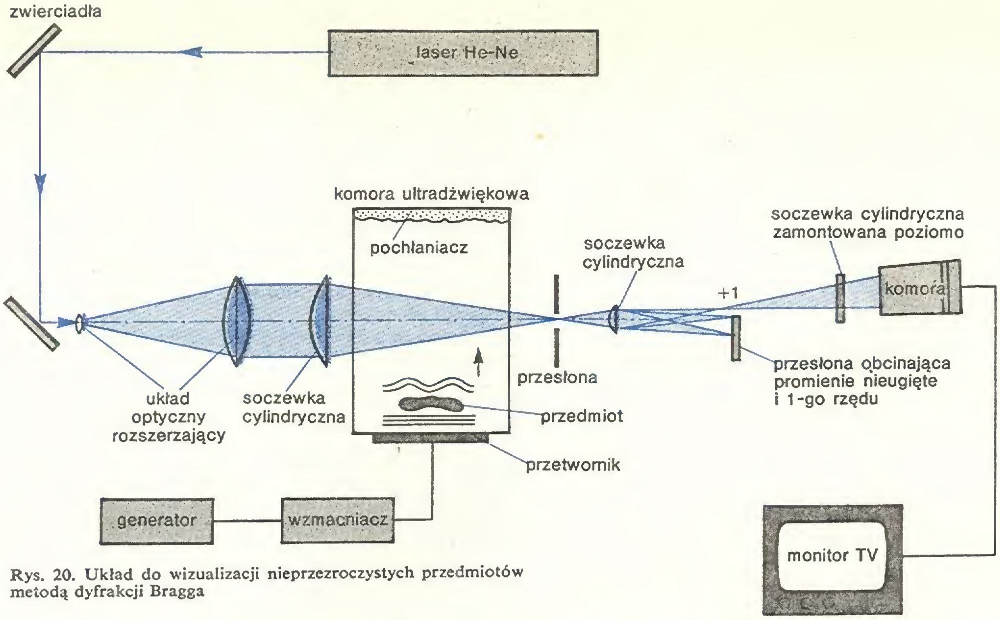
Rys. 19. Układ pamięci holograficznej z modulatorami i deflektorami akusto-optycznymi

na materiale światłoczułym lub termoplastycznym, powstaje mikrohologram. Zapisem lub odczytem żądanej informacji kieruje układ kontrolny za pośrednictwem odpowiednich deflektorów, odchylających w żądany sposób wiązkę przedmiotową i (lub) odniesienia.

wizualizacja nieprzezroczystych dla światła przedmiotów

Bardzo ważna jest możliwość wizualizacji nieprzezroczystych dla światła przedmiotów za pomocą fal akustycznych. Stosuje się do tego metodę dyfrakcji Bragga.

Przedstawiony na rys. 20 układ akustooptyczny służy do otrzymywania obrazów przedmiotów nieprzezroczystych „oświetlanych" ultradźwiękami o wysokiej częstotliwości. Rozkład natężenia światła w ugiętej wiązce, po wyjściu z komory ultradźwiękowej rejestrowany na monitorze TV, odpowiada rozkładowi pola akustycznego rozproszonego na przedmiocie. Na ekranie powstaje optyczny obraz przedmiotu. Układ ten, dostosowany do fal ultradźwiękowych,



Rys. 20. Układ do wizualizacji nieprzezroczystych przedmiotów metodą dyfrakcji Bragga

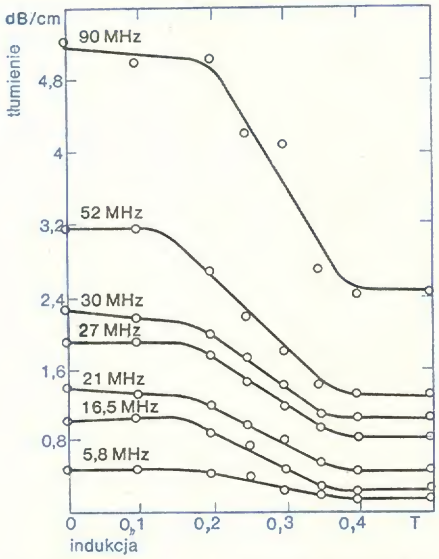
jest jednym z rozwiązań przetwarzania obrazu akustycznego na optyczny w mikroskopach ultradźwiękowych. Zdolność rozdzielcza mikroskopu ultradźwiękowego i jakość otrzymywanych z niego obrazów jest prawie taka jak obrazów uzyskiwanych za pomocą mikroskopów świetlnych (il. 165, tabl. 43). Przytoczone przykłady nie obejmują wszystkich możliwych rodzajów zastosowania optosoniki.

Zjawiska akustomagnetyczne

Czesław Lewa

zjawiska magnetostrykcji

Do zjawisk akustomagnetycznych zalicza się zjawiska fizyczne polegające na zmianie magnetycznych własności substancji pod wpływem zmiennej w czasie deformacji mechanicznej, jak również zjawiska odwrotne, zwane magnetostrykcją, polegające na zmianie kształtu i wymiarów materiałów magnetycznych podczas ich magnesowania. Zjawiska te wykorzystuje się np. w przetwornikach magnetostrykcyjnych.



Rys. 21. Zależność tłumienia ultradźwięków w niklu od indukcji magnetycznej

Z mikroskopowego punktu widzenia deformacja mechaniczna może wpływać na zachowanie się momentów magnetycznych całych domen magnetycznych w ferromagnetykach, ponieważ powoduje ona przemieszczanie ścian domenowych, co z kolei zmienia podatność magnetyczną materiałów. Ruch ten jest przyczyną zwiększonego tłumienia fal akustycznych w materiałach ferromagnetycznych oraz zależności tłumienia i prędkości rozchodzenia się fal akustycznych w tych materiałach od indukcji magnetycznej (rys. 21). Deformacja mechaniczna może również wpływać na zachowanie się momentów magnetycznych cząsteczek lub atomów, jak i momentów magnetycznych jąder atomowych (→ Teoria magnetyzmu).

zjawisko mechanostrykcji

Zmiany własności magnetycznych wywołane deformacją mechaniczną w materiałach magnetycznych powodują — na skutek zjawiska magnetostrykcji — dodatkową deformację sprężystą tych materiałów. Zjawisko to nosi nazwę mechanostrykcji. Fale akustyczne wywołują więc w środowisku magnetycznym, nawet w nieobecności zewnętrznego pola magnetycznego, zmianę kierunków i wartości spontanicznego namagnesowania domen. Towarzyszą temu zmiany wymiarów materiału. W materiałach magnetycznych mechanostrykcja może powodować odchylenia od prawa Hooke'a. Ze zjawiskiem mechanostrykcji wiąże się również wiele innych zjawisk fizycznych zachodzących w magnetykach, np. zmiana modułu sprężystości pod wpływem pola magnetycznego, wpływ naprężeń na magnetostrykcję.

zjawiska giromagnetyczne

Do zjawisk akustomagnetycznych należy również grupa zjawisk giromagnetycznych, w których się obserwuje zależność między momentem pędu i momentem magnetycznym atomów, np. zjawisko Barnetta — polegające na magnesowaniu się ferromagnetyków poddanych ruchowi obrotowemu, zjawisko Einsteina–de Hassa–Richarda, w którym się obserwuje obrót namagnesowanego ciała wokół osi pokrywającej się z kierunkiem pola magnetycznego.

Akustyczny rezonans magnetyczny

Czesław Lewa

Niektóre substancje umieszczone w stałym polu magnetycznym pochłaniają selektywnie (rezonansowo) energię pola akustycznego.

własności paramagnetyczne mikroobiektów

Wiele mikroelementów materii ma własności paramagnetyczne (para- i diamagnetyki). Takimi elementami są:

— atomy lub molekuły posiadające pewną liczbę nie sparowanych elektronów, np. molekuły O_2, NO, NO_2,

— defekty sieci w kryształach,

— swobodne rodniki chemiczne,

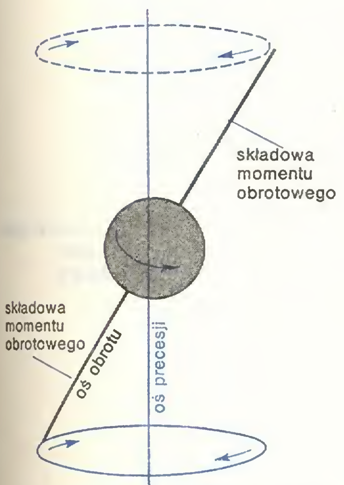
— atomy i jony o nie sparowanych wewnętrznych powłokach elektronowych, np. Fe, Ni, Co, Cr, Mn, Cu, Ti, V, pierwiastki ziem rzadkich,

a więc elementy, których własności paramagnetyczne wynikają z odpowiedniego stanu powłok elektronowych, oraz

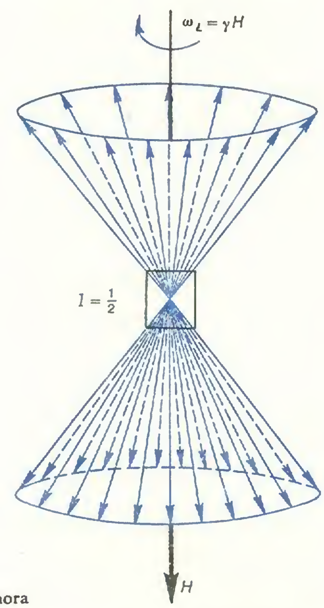
— jądra atomowe o nieparzystej liczbie protonów i neutronów (kwantowa liczba spinowa I przybiera wówczas wartości całkowite od 1 do 6) lub nieparzystej liczbie nukleonów (I przybiera wartości połówkowe od $^1/_2$ do $^9/_2$),

a więc elementy, których własności paramagnetyczne wynikają z odpowiedniej struktury i stanu jąder atomowych.

Umieszczenie takich elementów w polu magnetycznym (np. w polu ziemskim albo w polu wytworzonym przez nabiegunniki magnesu lub przez sąsiadujące elementy materii o własnościach paramagnetycznych) wymusza na nich przyjęcie określonych orientacji względem tego pola. Należy uwzględnić również moment pędu, którym są obdarzone paramagnetyczne mikroelementy, w wyniku czego ich zachowanie w polu magnetycznym podobne jest do zachowania bąka w polu grawitacyjnym (rys. 22). Moment pary sił, oddziaływania momentu magnetycznego rozważanych elementów z polem magnetycznym, wprawia je w ruch precesyjny (zwany precesją Larmora) o prędkości kątowej proporcjonalnej do natężenia pola magnetycznego H (rys. 23):



Rys. 22. Ruch precesyjny wirującego ciała



Rys. 23. Precesja Larmora

precesja Larmora

$$\omega_L = \gamma H,$$

gdzie γ jest współczynnikiem giromagnetycznym charakteryzującym rozważany rodzaj elementów.

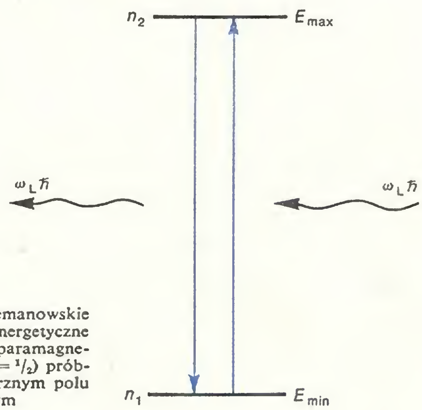
Z praw mechaniki kwantowej wynika, że elementy paramagnetyczne mogą przyjmować tylko takie orientacje względem kierunku pola magnetycznego, przy jakich energia ich oddziaływania z tym polem jest całkowitą wielokrotnością $\hbar\omega_L$, tzn.

$$E = -m\omega_L\hbar,$$

gdzie m to magnetyczna liczba kwantowa, $\hbar$ — stała Plancka dzielona przez 2π. Liczba m może przybierać wartości od $-I$ do $+I$. Oznacza to, że każdy element paramagnetyczny może przyjąć jedną z $(2I+1)$ orientacji, tzn. może się znaleźć w jednym z $(2I+1)$ stanów energetycznych (zwanych poziomami zeemanowskimi).

poziomy zeemanowskie

W próbce zawierającej wiele elementów paramagnetycznych umieszczonej w polu magnetycznym występują wszystkie możliwe orientacje, tzn. zajęte są wszystkie poziomy zeemanowskie. W stanie równowagi termicznej najwięcej momentów magnetycznych przyjmuje orientację o najniższej energii E_{min}, tzn. równoległą do kierunku pola magnetycznego, najmniej zaś — o najwyższej energii E_{max}, tzn. antyrównoległą do pola magnetycznego (rys. 24).



Rys. 24. Zeemanowskie poziomy energetyczne elementów paramagnetycznych ($I = ^1/_2$) próbki w zewnętrznym polu magnetycznym

Zmiana orientacji może zachodzić tylko skokowo, z zachowaniem reguł zwanych w mechanice kwantowej regułami wyboru. Według nich możliwe są tylko takie zmiany orientacji, podczas których m zmienia się o $\Delta m = \pm 1$, tzn. zmiany prowadzące do przyjęcia orientacji bezpośrednio sąsiadującej z poprzednią. Inaczej mówiąc — energia danego elementu może ulec zmianie tylko o

$$\Delta E = E(m) - E(m \pm 1) = \pm\hbar\omega_L = \hbar\gamma H.$$

Przejście elementu paramagnetycznego z jednego stanu energetycznego do sąsiedniego może nastąpić w wyniku absorpcji lub emisji kwantu energii $\omega_L\hbar$.

rezonans paramagnetyczny

Absorpcja energii przy przejściach między zeemanowskimi poziomami mikroelementów paramagnetycznych stanowi istotę rezonansu paramagnetycznego. Prawdopodobieństwo emisji energii zależy głównie od oddziaływań między rozważanym elementem paramagnetycznym i pozostałymi elementami materii w próbce.

Zależnie od tego, czy własności paramagnetyczne rozważanych elementów są pochodzenia elektronowego czy jądrowego, wprowadza się podział na dwie metody: elektronowego rezonansu paramagnetycznego ERP (skrót ang. EPR — *electronic paramagnetic rezonance*) i jądrowego rezonansu magnetycznego JRM (skrót ang. NMR — *nuclear magnetic rezonance*). Podział ten, choć z punktu widzenia przebiegu zjawiska jest czysto formalny, znajduje uzasadnienie techniczne (różne są zakresy częstości $\nu_L = \omega_L/2\pi$ precesji Larmora mimo zastosowania takich samych natężeń pól magnetycznych) i naukowo-poznawcze (obie metody mają różne zastosowania i dostarczają różnych, często uzupełniających się tylko informacji naukowych).

ERP i JRM

Porcje energii odpowiadające rezonansowym przejściom między zeemanowskimi poziomami energetycznymi dostarczać można do próbki zawierającej omawiane elementy paramagnetyczne w postaci kwantów $\hbar\omega$ promieniowania elektromagnetycznego lub fononów $\hbar\omega$ promieniowania mechanicznego (ultradźwiękowego). Metody, w których się wykorzystuje rezonansową absorpcję fononów, noszą odpo-

AERP i AJRM

wiednio nazwy: akustycznego elektronowego rezonansu paramagnetycznego (AERP) i akustycznego jądrowego rezonansu magnetycznego (AJRM).

Przekazanie energii fal ultradźwiękowych układowi elementów paramagnetycznych zachodzi wówczas, gdy w substancji występuje wewnętrzne oddziaływanie między tymi elementami a elementami materii wprawionymi w ruch drgający przez falę ultradźwiękową, tzn. gdy występuje sprzężenie spin–sieć. Uważając energię drgań sieci za zbiór fononów, możemy proces przekazywania energii traktować jako zbiór elementarnych oddziaływań fonon–spin. Akustyczne drgania wprawiają w ruch ładunki elektryczne oraz dipole magnetyczne zawarte w substancji, w wyniku czego zostaje w niej wytworzone zmienne pole elektromagnetyczne o częstości równej częstości drgań akustycznych. Gdy częstość drgań ν równa się częstości precesji Larmora ν_L, wytworzone pole elektromagnetyczne wywołuje przejścia pomiędzy poziomami zeemanowskimi elementów paramagnetycznych. Prawdopodobieństwa absorpcji i emisji energii w przeliczeniu na jeden element paramagnetyczny są jednakowe. W równowadze termicznej zbioru elementów paramagnetycznych zawartych w próbce liczba obsadzeń niższych poziomów zeemanowskich jest wyższa. Wystąpi więc rezonansowe pochłanianie fali akustycznej o częstości $\nu = \nu_L$, prowadzące do wzrostu liczby obsadzeń wyższych poziomów zeemanowskich, a tym samym do obniżenia absorpcji fali akustycznej.

oddziaływanie fonon–spin

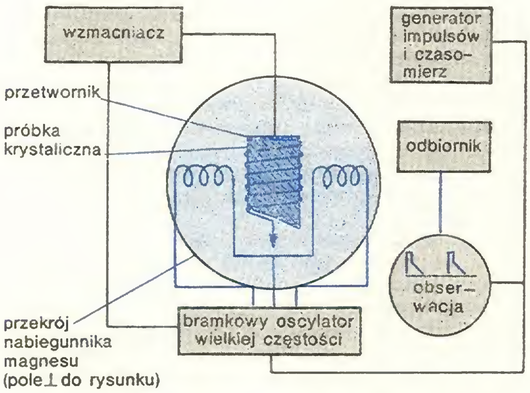
relaksacja spin–sieć

Oddziaływanie fonon–spin jest również odpowiedzialne za proces odwrotny — przekazywanie energii od zbioru elementów paramagnetycznych do sieci, tzn. za proces relaksacji spin–sieć charakteryzowany czasem relaksacji spin–sieć T_1. Szybkość tego procesu zależy od struktury i dynamiki molekularnej w materii. Pomiary czasów relaksacji dostarczają więc cennych informacji o strukturze i mikroskopowych procesach molekularnych zachodzących w badanej substancji w różnych stanach skupienia.

Możliwość wzbudzenia układu spinów przez fonony o częstości rezonansowej przewidzieli niezależnie od siebie A. J. Kastler (w 1952 r.) i S. A. Altshuler (w 1955 r.).

pierwsze doświadczenia

W pierwszych doświadczeniach, wykonanych przez W. G. Proctora i jego współpracowników w 1956 r., obserwowano absorpcję fononów przez zmianę różnicy obsadzeń poziomów zeemanowskich w próbce nadźwiękawianej rezonansową falą ultradźwiękową, tzn. metodą akustycznego nasycenia. Absorpcję fononów określano ze zmian absorpcji sygnału elektromagnetycznego w obecności pola ultradźwiękowego i bez niego. Fonony rezonansowe fali ultradźwiękowej, wzbudzając elementy paramagnetyczne do wyższych poziomów energetycznych, powodują częściowe nasycenie układu i obniżenie absorpcji sygnału elektromagnetycznego. Tak w opisywanym eksperymencie, jak i w późniejszych badaniach, najczęściej do tego celu stosowano tradycyjne urządzenia do JRM i ERP z wbudowanym nadajnikiem ultradźwiękowym.



Rys. 25. Schemat układu impulsowego do obserwacji AJRM

Pierwsze obserwacje absorpcji fali akustycznej przez jony paramagnetyczne wykonywane były w 1959 r. przez E. H. Jacobsena i współpracowników.

układ impulsowy do obserwacji AJRM

Na rys. 25 przedstawiono schemat układu impulsowego do obserwacji AJRM. Silne pole magnetyczne H ma kierunek prostopadły do płaszczyzny rysunku. Impuls elektromagnetyczny o częstości ν_L jest podawany przez cewki nadajnika. Nadźwiękowienie próbki uzyskuje się z przetwornika ultradźwiękowego. Za pomocą cewki odbiorczej, nawiniętej na próbce, rejestruje się wymuszony w próbce impuls i jego zanikający transjent. Amplituda tego impulsu (zwanego sygnałem precesji swobodnej) jest proporcjonalna do różnicy obsadzeń poziomów zeemanowskich. Maleje ona po nadźwiękowieniu próbki falą o częstości rezonansowej. Na rys. 26 pokazano sekwencję impulsów stosowaną w doświadczeniu Proctora. Po przyłożeniu impulsu elektromagnetycznego a i zarejestrowaniu sygnału precesji swobodnej o amplitudzie



Rys. 26. Impulsy elektromagnetyczne a i sygnały precesji swobodnej dla próbki nie nadźwiękowionej A_0 i nadźwiękowionej A

A_0 przykładano impuls ultradźwiękowy. Absorpcja energii akustycznej przez elementy paramagnetyczne przeszkadza ich powrotowi do stanu równowagi. Po wyłączeniu sygnału ultradźwiękowego mierzy się amplitudę sygnału precesji swobodnej A, uzyskanego przez wprowadzenie kolejnego impulsu elektromagnetycznego. Stosunek amplitud sygnału precesji swobodnej następującego po sygnale ultradźwiękowym A do amplitudy sygnału, gdy nie ma ultradźwięków A_0, jest miarą stosunku obsadzeń poziomów zeemanowskich w tych dwu wypadkach.

zastosowanie AJRM i AERP

Metody AJRM i AERP są szeroko stosowane w badaniach materiałów krystalicznych, natomiast nie udało się dotychczas osiągnąć powodzenia w zastosowaniu ich do cieczy i ciał amorficznch.

Metoda AERP została również zastosowana w faserach kwantowych.

Fasery

Mieczysław Szustakowski

Faserami nazywa się układy generujące dźwięki wysokich częstości — hiperdźwięki. Nazwa faser została wprowadzona przez polskiego uczonego S. Kaliskiego w 1969 r. i utworzona z pierwszych liter wyrazów angielskich określających zasadę działania kwantowych generatorów hiperdźwięków — *Fonon Ampcalifitionby Stimulated Emission of Radiation.*

Rozróżnia się dwa rodzaje faserów: fasery akustoelektryczne, w których wykorzystuje się oddziaływanie fonon–elektron, i fasery kwantowe, często zwane maserami fononowymi albo akustycznymi, w których wykorzystuje się oddziaływanie fonon–spin.

Fasery akustoelektryczne

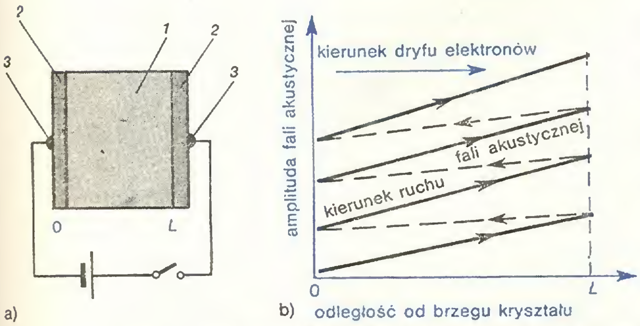
generacja fal akustycznych

W faserach akustoelektrycznych generacja drgań następuje w wyniku wzmocnienia spontanicznych drgań sieci (fononów termicznych) przez dryfujący strumień elektronów w kryształach piezopółprzewodnikowych.

Podstawowym elementem fasera akustoelektrycznego jest płytka piezopółprzewodnikowego kryształu

budowa fasera akusto-elektrycznego

(np. CdS, CdSe, ZnO) o powierzchni kilku mm^2 i grubości ułamka mm. Oś c (oś heksagonalna) kryształu skierowana jest prostopadle do powierzchni płytki. Dokładnie wypolerowane równoległe boki płytki pokryte są cienką warstwą metalu (rys. 27a). Warstwy metaliczne tworzą kontakty omowe z kryształem i umożliwiają doprowadzanie stałego pola elektrycznego E zwanego napięciem dryfu. Napięcie to powoduje ruch elektronów, a efekt piezoelektryczny sprzęga dryfujący strumień elektronów ze spontanicznie drgającą siecią kryształu. Gdy napięcie dryfu przekroczy wartość progową $E > E_p$, przy której prędkość elektronów osiąga prędkość równą prędkości fal akustycznych w krysztale, amplituda drgań atomów sieci, których kierunek ruchu jest zgodny z kierunkiem dryfu elektronów, będzie narastać (przebieg tego procesu



Rys. 27. Faser akustoelektryczny: a) schemat budowy, *1* kryształ piezopółprzewodnikowy, *2* warstwa metalu, *3* kontakty omowe, b) narastanie amplitudy fali akustycznej w zależności od kierunku ruchu fali względem kierunku dryfu elektronów

został już dokładnie omówiony w rozdziale pt. Oddziaływania fonon–elektron). W ten sposób z nieuporządkowanych, termicznych drgań sieci zostaje wygenerowana podłużna fala akustyczna (a w wypadku impulsowego napięcia dryfu — paczka falowa), która po dojściu do ścianki kryształu z dodatnią elektrodą odbija się i biegnie w kierunku przeciwnym do dryfu elektronów, jak na rys. 27b. Amplituda fali akustycznej biegnącej w kierunku przeciwnym do strumienia elektronów jest tłumiona, ale znacznie słabiej niż jest wzmacniana, gdy biegnie w kierunku elektrody dodatniej, czyli oddziaływanie fonon–elektron nie jest symetryczne względem kierunku ich wzajemnego ruchu. Niesymetryczność oddziaływania powoduje, że amplituda wielokrotnie odbitej fali akustycznej rośnie i dopiero, gdy zaczynają występować efekty nieliniowe uruchamiające dodatkowe mechanizmy tłumienia, osiąga poziom nasycenia, tzn. ustala się stan równowagi między wzmocnieniem i tłumieniem fali akustycznej. Narastanie amplitudy (wzmocnienie dryfowe) obejmuje tylko te drgania, których częstość kołowa ω spełnia warunek rezonansu określony przez odległość L między ściankami płytki:

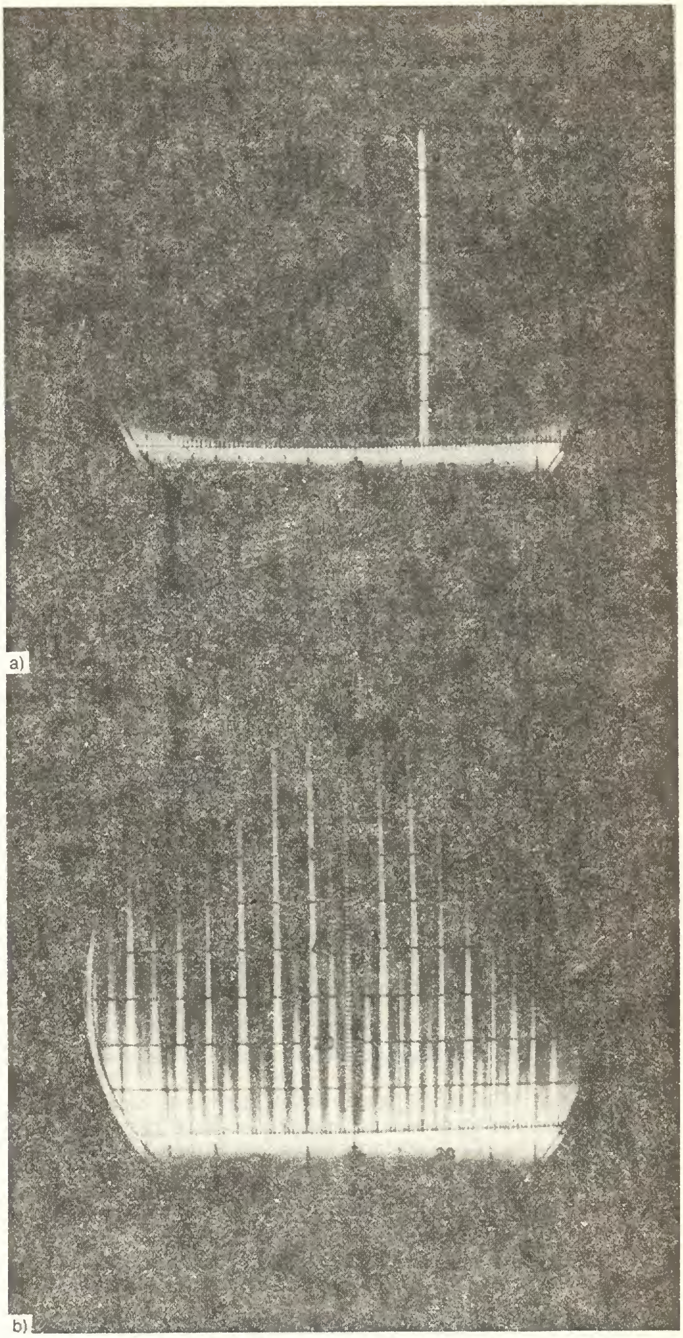
warunek rezonansu

$$\omega L\left(\frac{1}{v_p} - \frac{1}{v_0}\right) + \varphi = 2\pi n,$$

gdzie: $n = 1, 2, 3...$, v_p, v_0 oznacza prędkości fali akustycznej zgodnej i przeciwnej do kierunku dryfu, φ — różnicę faz powstałą przy odbiciu.

mody drgań

Faser generuje ciąg drgań akustycznych zwanych modami o częstości równej wielokrotności częstości podstawowej (częstość podstawowa przy $n = 1$). Wzbudzonym drganiom akustycznym w piezopółprzewodzącym ośrodku towarzyszy pole elektryczne o tej samej częstości wywołane efektem piezoelektrycznym. Dlatego ze ścianki kryształu prostopadłej do kierunku drgań akustycznych możemy wyprowadzić drgania w postaci fal akustycznych przy obciążeniu fasera falowodem akustycznym lub w postaci drgań elektrycznych przy obciążeniu obwodem elektrycznym.



Rys. 28. Oscylogram widma drgań fasera akustoelektrycznego: a) drgania jednomodowe, b) drgania wielomodowe

przykład — faser na CdS

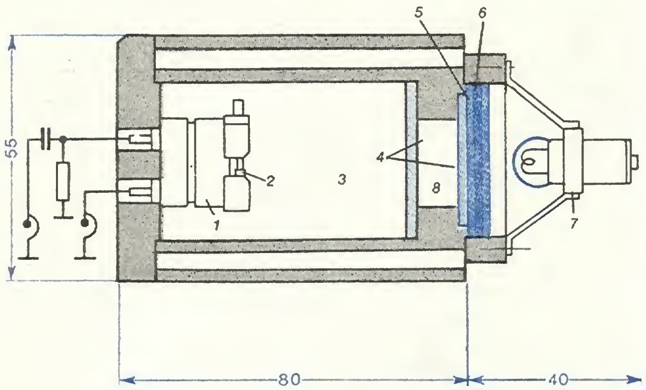
Rysunek 28 przedstawia zdjęcie z oscyloskopu analizatora widma drgań fasera skonstruowanego na bazie kryształu CdS. Dobierając odpowiednie napięcie pola dryfu i rozmiary próbki można uzyskać jednomodową pracę na częstości podstawowej (rys. 28a) lub pracę wielomodową (rys. 28b). Piki napięcia na rys. 29b odpowiadają różnym częstościom drgań — modom.

Kryształy CdS, CdSe są fotoczułe, tzn. że gęstość dryfującego w nich strumienia elektronów (przewodność) można regulować za pomocą oświetlenia. Przy oświetleniu światłem o długości fali bliskiej krawędzi absorpcji (np. dla CdS λ = 520–530 nm) światło wnika tylko na głębokość kilkudziesięciu mikrometrów, i wytwarza przypowierzchniową warstwę przewo-

dzącą. Przyłożenie do tak oświetlonej płytki napięcia dryfu powoduje powstanie fal powierzchniowych. Są to poprzeczne fale akustyczne o amplitudzie zanikającej z głębokością od powierzchni oświetlonej, których wychylenia cząstek (polaryzacja drgań) są równoległe do tej powierzchni (→ Akustyczne fale powierzchniowe i ich zastosowanie).

pierwsze modele faserów

Pierwsze modele faserów zostały opracowane w latach 1969–72. Schemat modelu fasera opracowanego w Polsce w 1970 r. przedstawia rys. 29. Podstawowe parametry tego fasera akustoelektrycznego są następujące: kryształ CdS o wymiarach 0,66×1,67×1,70 mm, częstość pracy 21703048 Hz, napięcie wyjściowe 0,2 mV, napięcie dryfu 80 V, prąd dryfu



Rys. 29. Schemat fasera akustoelektrycznego z podświetleniem: *1* oprawka na kryształ z dociskiem mechanicznym, *2* kryształ CdS, *3* komora termostatu, *4* płytki szklane, *5* matówka, *6* filtr barwny, *7* oświetlacz z żarówką, *8* okno do oświetlenia kryształu

0,3 mA, stałość częstości $\Delta f/f = 4 \cdot 10^{-6}$ na godzinę, czas ustalania się częstości przy oświetlonym krysztale ok. 1 h. Fasery akustoelektryczne nie znalazły jednak szerszego zastosowania w technice z powodu małej stabilności termicznej częstości drgań oraz konieczności chłodzenia ich i oświetlenia.

Fasery kwantowe

generacja hiperdźwięków

W faserach kwantowych generacja hiperdźwięków następuje w wyniku emisji fononów przy przejściach na niższe poziomy energetyczne mikroelementów paramagnetycznych. Faser kwantowy działa na zasadzie podobnej jak maser (→ Spektroskopia mikrofalowego rezonansu rotacyjnego) z tą tylko różnicą, że funkcję fotonów w maserze spełniają w faserze fonony, dzięki czemu zostaje wygenerowaną fala akustyczna zamiast fali elektromagnetycznej, która powstaje w maserze.

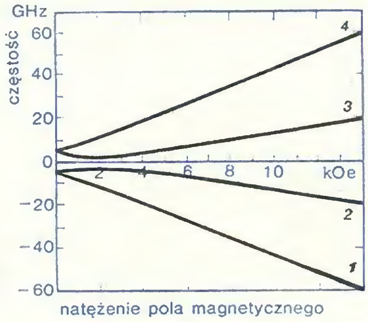
budowa fasera kwantowego

Podstawowym elementem fasera kwantowego jest paramagnetyczny kryształ, w którego dielektrycznej osnowie umieszczone są jony pierwiastków przejściowych (około 0,05%) z niesparowanym spinem elektronowym np. $Al_2O_3Cr^{+3}$. W zewnętrznym polu magnetycznym momenty magnetyczne jonów przyjmują tylko określone kierunki orientacji względem pola magnetycznego, co odpowiada różnym dyskretnym poziomom energii. Zmiana kierunku orientacji spinu odpowiadająca przejściu z niższego poziomu na wyższy wymaga dostarczenia energii równej różnicy poziomów, $(E_a - E_b)$, natomiast w sytuacji odwrotnej uzyskuje się kwant energii o częstości kołowej $\omega = (E_a - E_b)\hbar$. Uzyskane w ten sposób kwanty energii mogą być wyemitowane w postaci fotonów — wówczas mówimy o efekcie maserowym, lub fononów — efekt faserowy (oddziaływanie akustomagnetyczne).

Faser kwantowy został opracowany przez uczonych amerykańskich na początku lat 60. Możliwość generacji fononów za pomocą oddziaływania spin-fonon w paramagnetycznych kryształach przewidział w swych pracach w 1961 r. C. Kittel, później w 1963 r. E. B. Tucker uzyskał eksperymentalne potwierdzenie tego efektu.

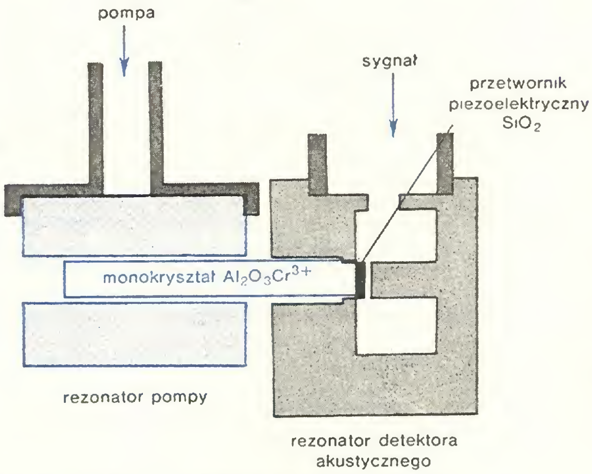
faser Tuckera

Do budowy fasera kwantowego Tucker użył monokryształu $Al_2O_3Cr^{3+}$. Dolne poziomy energetyczne jonów Cr^{3+} w rubinie są dubletami Cramersa rozdzielonymi w zerowym polu magnetycznym przerwą energetyczną ok. 38,6 m^{-1}. W polu magnetycznym *H* dublety rozszczepiają się tworząc cztery poziomy energetyczne. Przy orientacji pola pod kątem 54°44′ do osi *c* rubinu (trygonalna oś) rozszczepienia poziomów przyjmują postać jak na rys. 30. Różnice energii



Rys. 30. Poziomy energetyczne rubinu ($Al_2O_3Cr^{3+}$) przy orientacji pola pod kątem 54°44′ do osi trójkrotnej

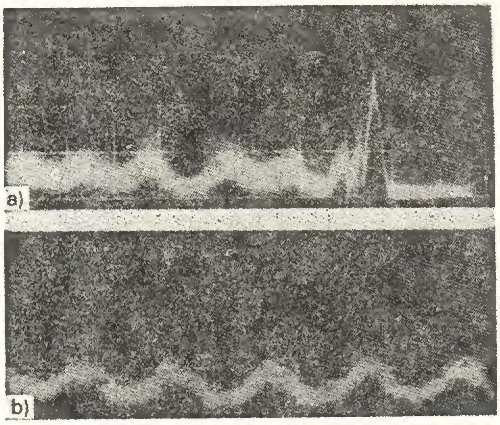
między poziomami *1–2* i *3–4* są jednakowe. W stanie równowagi termodynamicznej największa liczba jonów zajmuje najniższe poziomy energetyczne. Zmianę gęstości obsadzeń poziomów, czyli tzw. inwersję obsadzeń niezbędną dla uzyskania akcji faserowej, otrzymał Tucker za pomocą pola elektromagnetycznego (pola pompującego) o częstości 24 GHz odpowiadającej różnicy poziomów energetycznych *1–2* i *3–4*. Schemat ideowy fasera kwantowego konstrukcji Tuckera przedstawia rys. 31. Do czoła rubinowego walca przyklejony był przetwornik piezoelektryczny mający formę płytki kwarcowej milimetrowej grubości. Przetwornik służył do detekcji pola akustycznego i był umieszczony w rezonatorze mikrofalowym.



Rys. 31. Schemat głowicy fasera kwantowego Tuckera

Cały układ badawczy został zanurzony w ciekłym helu o temperaturze 4,2 K. Najpierw badano oddziaływanie pola akustycznego z układem spinów, nadźwiękawiając kryształ paramagnetyczny podłużną falą akustyczną. Zaobserwowano wzmocnienie fali akustycznej rzędu 10% na częstości 9,3 GHz. Na tejże częstości otrzymano generację podłużnych fal akustycznych o odebranej mocy elektromagnetycznej 10^{-11} W, co odpowiada mocy akustycznej rzędu 10^{-7} W lub amplitudzie deformacji rzędu 10^{-8}. Wzbu-

dzające pole elektromagnetyczne (pole pompujące) miało moc 40 mW, a temperatura kryształu była poniżej 4,2 K.
Na podstawie częstości otrzymanych drgań stwierdzono, że efekt generacji fononów otrzymuje się w wyniku przejść między poziomami *3–2*. Przejścia te są indukowane przez termiczne drgania sieci o częstości odpowiadającej różnicy energii między poziomami *3–2*, co interpretuje się jako wzmocnienie drgań sieci o tej częstości. Podobnie jak w faserze akustoelektrycznym wzmocnieniu ulegają te rodzaje fal akustycznych, dla których długość pręta paramagnetycznego jest całkowitą wielokrotnością połowy długości fali akustycznej. Generator kwantowy generuje wiele



Rys. 32. Oscylogram pokazujący drgania na wyjściu fasera kwantowego: a) drgania w stanie przejściowym po włączeniu pola pompy, b) zdudnienia dwóch sąsiednich rodzajów (modów) drgań, częstość dudnień 240 kHz

mody fasera kwantowego

rodzajów (modów) drgań, a różnica częstości między sąsiednimi modami wynosi $\Delta f = v/2L$. Przy prędkości $v = 1{,}2 \cdot 10^6$ m/sek i długości $L = 0{,}0254$ m różnica częstości w faserze Tuckera wynosiła 236 kHz. Na oscylogramie rys. 33b zarejestrowano częstość zdudnień dwóch sąsiednich modów, która wynosiła 240 kHz. Za pomocą odbiornika heterodynowego stwierdzono, że pasmo generacji fasera kwantowego wynosiło 11,4 MHz. Na oscylogramie rys. 32 pokazano typową dla fasera fluktuację wyjściowego sygnału związaną z początkiem generacji. Fluktuacja sygnału poprzedza ustalenie się stanu stacjonarnej generacji pokazanego na rys. 32b. Nowsze opracowanie modelu fasera przedstawił w 1974 r. E. M. Ganapolski. Faser Ganapolskiego zbudowany jest również na rubinie ($Al_2O_3Cr^{3+}$) o podobnych parametrach i orientacji co faser Tuckera. Innowacją jest rozwiązanie techniczne układu wzbudzania ośrodka paramagnetycznego (rezonatora pola pompującego) i detekcji (lub generacji) drgań akustycznych. Jako detektora drgań użyto cienkowarstwowego przetwornika piezoelektrycznego z ZnO napylonego bezpośrednio na kryształ rubinu. Ganapolski wykorzystywał faser do badania wzmocnienia podłużnych fal akustycznych pobudzanych przetwornikiem ZnO na częstości 9,3 GHz. Częstość pola pompy wynosiła 23,4 GHz. Otrzymane wzmocnienie fal akustycznych wynosiło 0,9 dB/cm przy temperaturze 1,7 K i koncentracji chromu Cr^{3+} 0,028%. Natomiast przy koncentracji chromu 0,05% wzmocnienie wynosiło tylko 0,4 dB/cm. W swych badaniach Ganapolski stwierdził zależność oddziaływania fonon–spin od koncentracji jonów parametrycznych, kąta orientacji kryształu względem pola magnetycznego, mocy pompy i temperatury.

faser Ganapolskiego

Prace badawcze nad faserami kwantowymi trwają nadal. Obecnie dąży się do uzyskania akcji faserowej w kryształach piezoelektrycznych, np. w $LiNbO_3{:}Cr^{3+}(Fe^{3+})$. W tym wypadku drgania akustyczne można będzie odbierać bezpośrednio z kryształu paramagnetycznego pomijając stosowanie dodatkowych przetworników, które na częstościach gigahercowych powodują duże straty akustycznej energii fasera. Fasery kwantowe a szczególnie zjawisko oddziaływania fonon–spin są wykorzystywane w badaniach akustoelektronowego rezonansu paramagnetycznego (AERP).

AERP

Acoustic Surface Wave and Acousto-optic Devices T.Kallard (ed.), New York 1971; R. T. Beyer, S. V. Letcher *Physical Ultrasonics*, New York 1969; I. Malecki *Podstawy teoretyczne akustyki kwantowej*, Warszawa 1972; W. P. Mason *Physical Acoustics*, vol. 1–12, New York 1964–76; A. Śliwiński, E. Ozimek *Akustyka laboratoryjna*, cz. 3, Warszawa 1974; R. Truell i in. *Ultrasonics Methods in Solid State Physics*, New York 1969; J. W. Tucker, V. W. Rampton *Microwave Ultrasonics in Solid State Physics*, Amsterdam 1972 (tłum. ros. Moskwa 1975).

Modelowanie obiektów akustycznych

Stefan Czarnecki

W badaniach akustycznych szeroko stosuje się metodę doświadczalna zwaną modelowaniem obiektów akustycznych. Polega ona na zastąpieniu badanych obiektów przez inne, których parametry można łatwo zmieniać w celu poszukiwania rozwiązań optymalnych. Metoda ta rozpowszechniona jest w pracach badawczych, jak również w pracach projektowych.

Badany obiekt można odwzorowywać za pomocą modeli cyfrowych lub analogowych.

modele cyfrowe

Modele cyfrowe wymagają pełnego opisu matematycznego badanego obiektu, w celu odpowiedniego zaprogramowania komputera. Szeroko stosowana technika komputerowa zapewnia uniwersalność modelowania oraz możliwość automatycznego poszukiwania rozwiązań optymalnych przez zastosowanie odpowiednich kryteriów optymalizacji. Wadą modelowania cyfrowego jest mały kontakt człowieka z komputerem, co utrudnia proces śledzenia toku procesu modelowania.

modele analogowe

Kontakt człowieka z komputerem jest znacznie większy w wypadku modelowania analogowego, którego dodatkową zaletą jest możliwość analizy badanego obiektu bez znajomości jego pełnego opisu matematycznego.

Modelowanie obiektów akustycznych metodami analogowymi realizowane jest najczęściej za pomocą maszyn analogowych lub hybrydowych, względnie za pomocą modeli geometrycznych.

Model analogowy jest odwzorowaniem badanego obiektu w postaci układu zastępczego, opisywanego takimi samymi zależnościami matematycznymi co rzeczywiste układy modelowe i o parametrach wzajemnie proporcjonalnych. Modele te mają duże znaczenie dydaktyczne, gdyż dają proste, syntetyczne spojrzenie na rozpatrywany obiekt, a dzięki odpowiedniemu doborowi układu zastępczego lub jego modyfikacji umożliwiają polepszenie cech obiektu badanego.

model geometryczny

Często stosowanym modelem analogowym jest model geometryczny. Model geometryczny przedstawia badany obiekt w zmienionej skali (najczęściej zmniejszonej), przy czym, żeby był spełniony warunek wza-

jemnej proporcjonalności parametrów, musi być spełniona zależność:

$$n = l_0/l_m = \nu_m/\nu_0,$$

gdzie n jest współczynnikiem skali, l_0 i l_m — odpowiednio liniowym wymiarem obiektu i modelu, ν_m i ν_0 — częstością pomiarową obiektu i modelu. Badanie modelowe obiektu w funkcji częstości pociąga za sobą konieczność odwzorowania charakterystyk częstościowych właściwości akustycznych materiałów (np. współczynnika pochłaniania) lub właściwości ośrodka (współczynnika tłumienia fali akustycznej przez powietrze), co jest dosyć trudne zwłaszcza przy dużych współczynnikach skali. Warunkiem efektywnego korzystania z modeli geometrycznych, jest ich bardzo precyzyjne wykonanie, co pociąga za sobą dość duże koszty.

Wymienione powyżej wady i zalety poszczególnych modeli sprawiają, że każdy z nich może być stosowany do odmiennych celów.

zastosowanie modeli

Modelowanie cyfrowe stosowane jest np. do obliczeń tłumików hałasu metodą elementu skończonego. Badany tłumik dzielony jest na obszary o znanych lub nie znanych warunkach brzegowych, w wyniku czego otrzymujemy rozwiązanie układu kilkudziesięciu równań algebraicznych. Modelowanie analogowe przydatne bywa przy modelowaniu przetworników elektroakustycznych (głośników, mikrofonów, czujników) lub filtrów akustycznych używanych w tłumikach hałasu. Do analizy tych urządzeń wykorzystywane są również elektryczne modele analogowe.

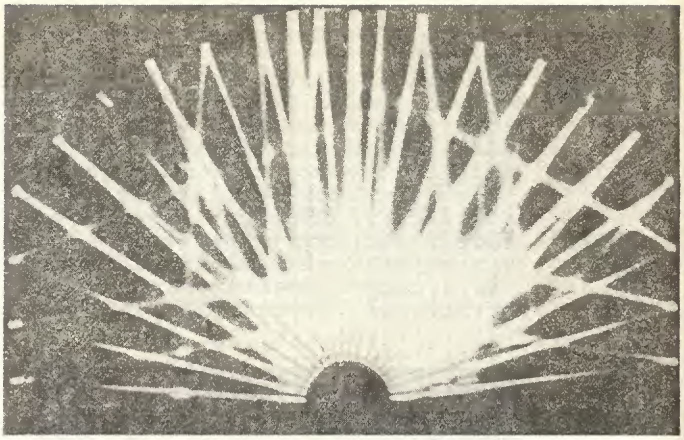
Modelowanie geometryczne stosowane jest przede wszystkim do badania sal, badania wnętrz urbanistycznych oraz badania przegród izolacyjnych (→ Hałas).

Do badania sal wykonuje się model w zmniejszonej skali i stosuje się sygnał o odpowiednio wyższej częstości; bada się w ten sposób rozkład fal odbitych w zależności od kształtu sali jak i rozłożenia materiałów dźwiękochłonnych.

Szeroko rozpowszechnione są badania techniką impulsową, która umożliwia określenie nie tylko rozkładu fal odbitych, lecz również odstępów czasu dochodzenia kolejnych fal odbitych, co ma bardzo duże znaczenie dla akustyki wnętrz.

modele optyczne

Wygodną metodą jakościową badania sal (ilościową nie, gdyż nie ma proporcjonalności parametrów) mogą być realizowane za pomocą modelu optycznego, którego kształt odpowiada kształtowi sali. Własności odbijające ścian odwzorowane są za pomocą powierzchni lustrzanych, natomiast własności pochłaniające — za pomocą powierzchni matowych. Wiązki fal akustycznych zastąpione są wiązkami świetlnymi wychodzącymi z punktowego źródła światła. Fotografia uzyskanego rozkładu fal odbitych (zob. obok) umożliwia znalezienie obszarów zagęszczenia i rozrzedzenia energii fal odbitych, co można korygować poprzez zmianę kształtu modelu lub zmianę rozmieszczenia powierzchni matowych.



Fotografia rozkładu fal odbitych uzyskanego w modelu optycznym

modele urbanistyczne

Modele urbanistyczne przedstawiają w zmniejszonej skali wycinek wnętrza urbanistycznego złożonego z arterii komunikacyjnych i zabudowy oraz odpowiednio odwzorowują źródła hałasu z uwzględnieniem zarówno ich rozmiarów, jak i zakresu częstości. Celem badań jest projektowanie takiej lokalizacji budynków mieszkalnych i budynków ekranujących, by hałas arterii komunikacyjnych dochodzący do wnętrza budynków mieszkalnych był jak najmniejszy. Badania te utrudnia konieczność stosowania dużego współczynnika skali, co przesuwa znacznie zakres częstości w modelu i w dużym stopniu uniemożliwia odwzorowanie współczynnika tłumienia powietrza, który zmienia się wraz z częstością.

badania modelowe przegród i osłon

Badania modelowe własności izolacyjnych przegród i osłon przeciwdźwiękowych nie wymagają tak dużej zmiany skali częstości, gdyż nie jest konieczne stosowanie dużych współczynników skali. Utrudnieniem w tego rodzaju badaniach powodującym to, że wyniki mają charakter przybliżony stanowi nieharmoniczność rozkładu częstości drgań własnych płyt i jego zależności od sposobu zamocowania, wskutek czego zmiana wymiarów badanego obiektu nie zmienia proporcjonalnie częstości rezonansowych. Dlatego też — szczególnie w wypadku osłon — badania modelowe są przydatne w stosunku do materiałów i obiektów silnie tłumionych, dla których wpływ częstości rezonansowych odgrywa mniejszą rolę.

Fale uderzeniowe

Wiktor Jungowski

Fala uderzeniowa jest frontem gwałtownego wzrostu ciśnienia, temperatury, gęstości i prędkości ośrodka ciągłego, którym może być gaz, ciecz lub ciało stałe. Gwałtowność tego wzrostu powoduje nieodwracalność przemiany przy sprężaniu ośrodka przez falę uderzeniową. Oznacza to, że ośrodek rozprężony z powrotem do ciśnienia początkowego ma temperaturę wyższą od początkowej. Wzrost wymienionych wyżej wielkości nie pociąga za sobą nieodzownie ich obniżenia, tak jak w fali akustycznej lub fali na powierzchni cieczy, kiedy to ośrodek zostaje przemieszczony względem stałego średniego położenia.

fale ciśnieniowe

W tzw. falach ciśnieniowych nawet znaczne sprężanie lub rozprężanie ośrodka odbywa się stopniowo i przemiany są odwracalne. W falach akustycznych przyrosty ciśnienia są bardzo małe a prędkość stała (zależna od ściśliwości ośrodka).

W falach ciśnieniowych zgęszczeniowych i w falach uderzeniowych wyróżniamy czoło fali, tzn. stronę, od której rozpoczyna się wzrost ciśnienia, i tył fali — miejsce, w którym wzrost ten się kończy. Analogicznie czołu rozrzedzeniowej fali ciśnieniowej odpowiada początek spadku ciśnienia, a tyłowi — koniec.

prędkość fali uderzeniowej

Prędkość fali uderzeniowej względem ośrodka jest zawsze większa od prędkości dźwięku i rośnie z jej natężeniem, określonym stosunkiem przyrostu ciśnienia do ciśnienia przed falą. W gazach prędkość ta może przekraczać wielokrotnie prędkość dźwięku, w cieczach i ciałach stałych, nawet jeśli fale mają bardzo duże natężenie, jest z nią porównywalna.

Prędkość dźwięku w gazach o umiarkowanej temperaturze jest na ogół mniejsza, a w cieczach większa od 1 km/s, natomiast w ciałach stałych osiąga kilka km/s. Na przykład prędkość dźwięku przy temperaturze 20°C wynosi w powietrzu 0,34 km/s, w wodzie 1,46 km/s i w stali 4,99 km/s. Stosunek prędkości fali uderzeniowej do prędkości dźwięku przed nią jest nazywany liczbą Macha M_s. W wypadku jądrowej eksplozji w powietrzu M_s ma wartość ponad 3000, co odpowiada prędkości fali ok. 1000 km/s. Ze wzrostem natężenia fali uderzeniowej rośnie temperatura za jej czołem i w gazach może wynosić od kilkuset do milionów K. Dolnej granicy odpowiada fala poruszająca się przed kulą karabinową, a górnej — fala wywołana wspomnianą wyżej eksplozją jądrową.

liczba Macha

Fale uderzeniowe powstają na skutek gwałtownego wyzwolenia energii, np. przy eksplozji materiału wybuchowego lub eksplozji jądrowej, pęknięciu zbiornika z parą lub gazem o wysokim ciśnieniu, wyładowaniu elektrycznym, wybuchu wulkanu, wypływie gazu lub pary z prędkością naddźwiękową. Ciała poruszające się z prędkością nadźwiękową względem ośrodka, takie jak meteoryty, pojazdy kosmiczne, pociski lub samoloty, są również źródłem fal uderzeniowych. Fala uderzeniowa jest wtedy wywołana ruchem ciała w ośrodku stawiającym opór. W wyniku tego ciało albo zmniejsza prędkość, tak np. jak kabina pojazdu kosmicznego w atmosferze (energia kinetyczna kabiny zamienia się w fali uderzeniowej na ciepło), albo porusza się ze stałą prędkością, tak jak samolot napędzany silnikami. Wówczas siła oporu jest pokonywana kosztem energii wyzwolonej z paliwa. Zawsze jednak powstanie lub istnienie fali uderzeniowej musi mieć źródło energii. W pobliżu źródła fale mogą być bardzo silne, ale rozprzestrzeniając się zmniejszają swoje natężenie (i prędkość) i stopniowo zanikają, przeradzając się w słabe zaburzenia akustyczne. Proces zanikania dokonuje się na najdłuższej drodze w gazach, na krótszej w cieczach i najkrótszej w ośrodkach stałych.

źródła fal uderzeniowych

znaczenie fal uderzeniowych

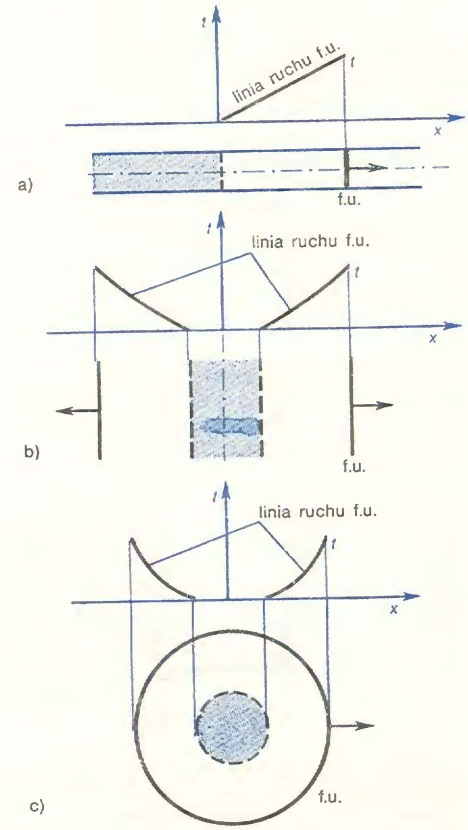
Zjawisko fizyczne, jakim jest fala uderzeniowa, ma dla człowieka różnorodne znaczenie. Jako niekontrolowana siła przyrody fala uderzeniowa może stanowić zagrożenie. Można tu wymienić fale uderzeniowe spowodowane wyładowaniem energii elektrycznej w atmosferze (uderzenie pioruna), ruchem meteorytu w pobliżu powierzchni Ziemi z prędkością naddźwiękową i jego upadkiem oraz wypływem gazów i lawy podczas wybuchu wulkanów. Takimi niekontrolowanymi zjawiskami przyrody są również trzęsienia Ziemi wywołujące fale typu fali uderzeniowej biegnące w płaszczu i skorupie ziemskiej lub fale powodujące ruch olbrzymich mas wody w oceanach (fale tsunami), a także huragany lub tajfuny powodujące powstawanie na powierzchniach mórz i oceanów stromych fal podobnych do fali przypływu. Kontrolowane przez człowieka fale uderzeniowe mogą być środkiem niszczenia, bądź pomocą w kształtowaniu terenu, zarówno na powierzchni Ziemi jak i w jej wnętrzu, czy też na dnie zbiorników wodnych.

zastosowanie fal uderzeniowych

Możliwość wytwarzania wysokich ciśnień i temperatur za pomocą fal uderzeniowych znalazła wiele zastosowań. Na pierwszym miejscu należy wymienić instalacje badawcze: rury uderzeniowe, niektóre tunele aerodynamiczne i wyrzutniki, służące do badań opływu ciał z dużymi prędkościami, badań własności gazów itp. Bezpośrednie praktyczne zastosowanie znalazły fale uderzeniowe w instalacjach zapłonowych, w których wytwarzają wysoką temperaturę gazu, jako źródła dźwięku o bardzo wysokim poziomie (uzyskiwane za pomocą oscylujących fal uderzeniowych), w badaniach geologicznych (rejestracja ruchu fal w gruncie wywołanych wybuchami w różnych punktach badanego terenu), przy produkcji diamentów z grafitu przez poddawanie go bardzo wysokiemu ciśnieniu, w tłoczeniu blach (poddawanych działaniu fal uderzeniowych wytworzonych w wodzie przy wybuchu), w wykonywaniu otworów (impulsowym strumieniem cieczy o bardzo dużej prędkości) oraz w wytwarzaniu silnych wibracji elementów mechanicznych.

Ze względu na liczbę współrzędnych opisujących ruch fali względem ośrodka można rozróżnić fale jedno-, dwu- i trójwymiarowe. Fale jednowymiarowe — to fale płaskie, cylindryczne i kuliste (rys. 1). Wykresy na rys. 1 przedstawiają zmianę położenia x fali w czasie t, poczynając od chwili początkowej $t = 0$ do $t = t'$, inaczej mówiąc — linię ruchu fali. Rysunki pod wykresami uwidoczniają tylko położenia początkowe i końcowe fali uderzeniowej. Można wykazać, pomijając tłumiące działanie lepkości gazu, że prędkość fali płaskiej jest stała, a cylindrycznej i kulistej maleje w miarę oddalania się od punktu wyjściowego. Wszystkie te fale napotykając obiekt, który powoduje ich odbicie i dyfrakcję, mogą się przerodzić w fale dwu- lub trójwymiarowe.

linia ruchu fali



Rys. 1. Linie ruchu fal uderzeniowych (wykresy x–t) obrazujące położenie w czasie: a) fali płaskiej, b) cylindrycznej, c) kulistej. Kolorem niebieskim zaznaczono obszary wysokiego ciśnienia

fale o małym natężeniu

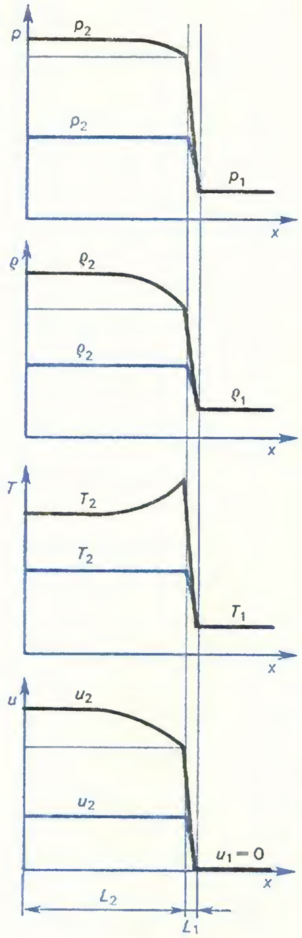
W odniesieniu do fal uderzeniowych występujących w gazie nasuwa się podział na fale o małym natężeniu, w których temperatura pozostaje na umiarkowanym poziomie i odgrywają rolę tylko stopnie swobody związane z przesunięciami i obrotami molekuł, oraz na fale o dużym natężeniu, w których temperatura jest na tyle wysoka, że zostają wzbudzone stopnie swobody związane z drganiami atomów w molekule, dysocjacją, wzbudzeniem atomu i jonizacją. W pierwszym wypadku do osiągnięcia przez gaz stanu równowagi termodynamicznej wystarcza kilka zderzeń. Grubość fali (L_1 na rys. 2) jest wówczas rzędu drogi swobodnej molekuły (w warunkach atmosferycznych $L_1 \approx 0{,}02$ μm). W drugim wypadku natomiast osiągnięcie stanu równowagi po wzbudzeniu wymaga liczby zderzeń większej o kilka rzędów wielkości i wówczas za warstwą L_1 znajduje się znacznie grubsza strefa L_2, w której parametry gazu ulegają

fale o dużym natężeniu

dalszej zmianie, aż do osiągnięcia przez gaz stanu równowagi termodynamicznej.

Podobną strukturę ma fala detonacyjna, w której za warstwą L_1 występuje strefa spalania, a więc strefa reakcji chemicznej. Jeżeli ogrzany gaz za falą uderzeniową silnie promieniuje, to parametry gazu przed falą zmieniają się w miarę jej zbliżania, w wyniku ogrzewania się gazu. W plazmie mogą pojawić się także fale uderzeniowe, w których poza ciśnieniem

porównanie fal o małym i dużym natężeniu



Rys. 2. Zmiana parametrów gazu (ciśnienia, gęstości, temperatury i prędkości) w fali uderzeniowej o małym natężeniu (krzywa niebieska) i fali uderzeniowej o dużym natężeniu (krzywa czarna), biegnących w dodatnim kierunku osi x; L_1 grubość fali o małym natężeniu, L_1+L_2 grubość fali o dużym natężeniu

gazu odgrywają rolę siły wywołane obecnością pola magnetycznego, zależne od przenikalności magnetycznej plazmy, natężenia przepływającego w niej prądu elektrycznego i natężenia pola magnetycznego.

fale uderzeniowe w przestrzeni kosmicznej

W przestrzeni kosmicznej wiatr słoneczny, będący strumieniem cząstek wyrzuconych przez Słońce i poruszających się z prędkością naddźwiękową (ok. 400 km/s), przy zetknięciu z ziemskim polem magnetycznym powoduje powstanie fali uderzeniowej. W istniejących tam warunkach silnego rozrzedzenia cząstek, fala uderzeniowa może mieć grubość tysięcy kilometrów. Fala taka powstaje nie w wyniku zderzeń cząstek, lecz jedynie w wyniku oddziaływań elektromagnetycznych. Oddziaływania te powodują, że temperatura za frontem fali jest co najmniej dziesięciokrotnie wyższa niż w wietrze słonecznym przed nim. Powstanie fali uderzeniowej zmienia strukturę magnetosfery w sposób mogący ulegać fluktuacjom w czasie, ponieważ własności wiatru słonecznego zależą od aktywności Słońca. Jest to przyczyną zjawisk obserwowanych bezpośrednio na Ziemi, takich jak burze magnetyczne czy zorze polarne. Źródłem fal uderzeniowych w przestrzeni kosmicznej są również wybuchy słoneczne.

Rozpatrując fale uderzeniowe w układzie odniesienia związanym z obiektem, na który oddziałują, rozróżniamy fale biegnące, stacjonarne i oscylujące. Oczywiście układ odniesienia nie zmienia własności fal ani związków między parametrami ośrodka przed i za falą. Zależą one od względnej prędkości fali i ośrodka. Natomiast dla obiektu, na który działa fala, jest istotne, czy przemieszcza się ona względem niego czy też nie.

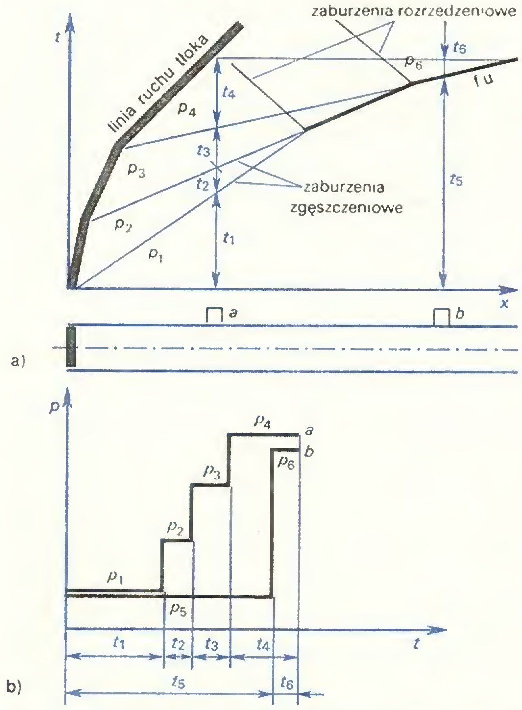
Fale biegnące

Jeżeli gaz o wysokim ciśnieniu (obszary niebieskie na rys. 1) jest oddzielony ścianką od otoczenia, to pęknięcie ścianki powoduje wypływ gazu i utworzenie się fali uderzeniowej. Ilustracja 170 (tabl. 45) przedstawia obraz przepływu, odpowiadający rys. 1c, otrzymany metodą smug (nazywaną również metodą Teplera, w której wykorzystuje się zależność kąta odchylenia promienia świetlnego od pierwszej pochodnej gęstości gazu), po upływie 150 µs od pęknięcia szklanej kuli o średnicy 5 cm. Kula była wypełniona powietrzem i pękła przy ciśnieniu 3,6 MPa. Na zdjęciu widzimy kulistą falę uderzeniową S, front C wypływającego z kuli powietrza i popękaną ściankę G kuli. Przepływ powietrza wydostającego się pomiędzy fragmentami ścianki z wnętrza kuli jest silnie burzliwy, w odróżnieniu od przepływu powietrza otaczającego kulę i wprawianego w ruch przez falę uderzeniową. Fragmenty ścianki kuli, z powodu swojej bezwładności, są niewiele przesunięte względem początkowego położenia.

kulista fala uderzeniowa

Podobne zjawiska występują przy przerwaniu przepony w rurze uderzeniowej — urządzeniu służącym m.in. do wytwarzania i badania fal uderzeniowych. Wzdłuż rury przemieszcza się wtedy płaska fala uderzeniowa (rys. 1a). Źródłem fali uderzeniowej jest również gwałtowne otwarcie zaworu oddzielającego gaz o różnych ciśnieniach. Wypływ gazów prochowych, po opuszczeniu lufy przez pocisk, wywołuje silną falę uderzeniową S_2 (il. 174, tabl. 45). Pocisk jest słabo widoczny w ciemnych gazach prochowych, jego wierzchołek znajduje się blisko fali S_2. Fala uderzeniowa S_1 powstała wcześniej i jest wynikiem dyfrakcji

płaska fala uderzeniowa



Rys. 3. Powstawanie fali uderzeniowej wskutek ruchu tłoka (a) i przebiegi w czasie ciśnień (b) mierzonych przetwornikami a i b

prostopadłej fali uderzeniowej wytworzonej w lufie przed pociskiem. Obie fale i ich odbicia od Ziemi oraz otaczających obiektów są głównymi źródłami słyszanego przy wystrzale huku.

mechanizm generowania fali uderzeniowej

Mechanizm generowania fali uderzeniowej przez poruszający się w przewodzie pocisk lub tłok jest następujący. Tłok, znajdujący się w chwili początkowej w spoczynku, zostaje wprawiony w ruch działającą nań różnicą ciśnień i gwałtownie przyspiesza. Popychając stykający się z nim gaz wytwarza on słabe zaburzenia zgęszczeniowe biegnące wzdłuż przewodu. Zastępując ciągły przyrost prędkości tłoka przyrostami skończonymi możemy przedstawić jego linię ruchu i przemieszczanie się zaburzeń w sposób pokazany na rys. 3a. Prędkość zaburzeń względem przewodu jest sumą prędkości gazu i prędkości dźwięku w gazie. Ponieważ w miarę rozpędzania się tłoka rosną obie prędkości (prędkość dźwięku na skutek podwyższania się temperatury sprężanego gazu), to zaburzenia wytworzone później doganiają wcześniejsze, i czoło fali ciśnieniowej staje się coraz bardziej strome tworząc falę uderzeniową. Znajdujące się w ściance przewodu przetworniki ciśnienia *a* i *b* umożliwiają zarejestrowanie na ekranie oscyloskopu przebiegów ciśnienia w czasie (rys. 3b). Jak widać, dla danej prędkości tłoka przyrost ciśnienia w fali uderzeniowej jest mniejszy niż w izentropowej fali ciśnieniowej ($p_6 < p_4$), co powoduje powstanie zaburzeń rozrzedzeniowych (rys. 3a). Przyczyną tego jest wzrost entropii, wynikający z nieodwracalnej zamiany części energii mechanicznej na cieplną w fali uderzeniowej.

Płaska fala uderzeniowa, biegnąca wzdłuż przewodu o stałym przekroju, zmniejsza stopniowo swoje natężenie na skutek tarcia gazu o ściankę. Gdy fala biegnie wzdłuż przewodu o perforowanej ściance (np. w tłumiku hałasu wylotowego gazu o wysokim ciśnieniu lub w tłumiku odrzutu lufy działa), to ulega ona osłabieniu także z powodu wypływu gazu do otoczenia. Na zdjęciu (il. 168, tabl. 44) przepływu perforowaną rurą z szybami optycznymi w części centralnej są widoczne: płaska fala uderzeniowa S_0 w przewodzie, front *C* gazu wypływającego przez otwory z prędkością krytyczną (czyli równą prędkości dźwięku), wywołane tym zaburzenia akustyczne *W* sięgające do zakrzywionej fali uderzeniowej poruszającej się na zewnątrz przewodu.

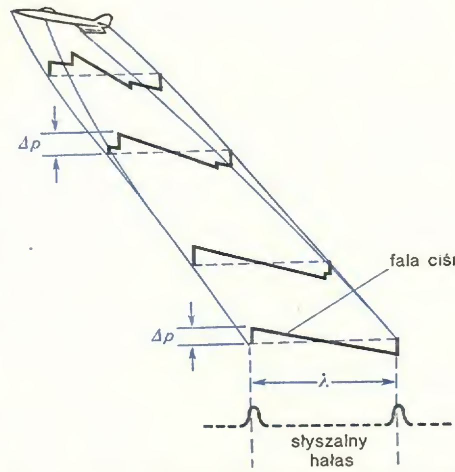
dyfrakcja fali uderzeniowej

Płaska fala uderzeniowa trafiając na zmianę przekroju przewodu lub inną przeszkodę ulega odbiciu i dyfrakcji, powstaje złożony układ fal. Na zdjęciu (il. 169, tabl. 44) wykonanym metodą cieni (wykorzystuje się zależność kąta odchylenia promienia świetlnego od drugiej pochodnej gęstości gazu) jest widoczny fragment pola przepływu w kanale o przekroju prostokątnym, którego ściany boczne wykonano ze szkła optycznego. Można wyróżnić: płaską falę uderzeniową S_0 biegnącą wzdłuż kanału ponad pionowo ustawionym klinem, odbitą od pochyłej ścianki klina falę S_2 i powstałą w wyniku dyfrakcji fali płaskiej zakrzywioną falę S_1. Z linii (na zdjęciu punkt) przecięcia tych fal wychodzi powierzchnia (na zdjęciu krzywa) *C* nieciągłości temperatury i gęstości gazu, która łączy się z wirem *V* przy ostrzu klina. Ponadto słabo widoczne jest czoło fali rozrzedzeniowej *R*, biegnącej pod prąd pomiędzy pochyłą ścianką klina a falą uderzeniową S_2.

Przejście fali uderzeniowej przez kanał może stanowić fazę początkową ustalania się przepływu. Następuje to wówczas, gdy są utrzymywane stałe warunki zasilania. Na il. 175 (tabl. 45) są widoczne kolejne zmiany struktury przepływu za płaską falą uderzeniową, prowadzące do utworzenia stacjonarnych, skośnych fal uderzeniowych w początkowym odcinku kanału. Ich obecność wskazuje, że przepływ jest naddźwiękowy. Prędkość fali płaskiej na wejściu do kanału jest rzędu 1 km/s. Na ostatnim zdjęciu widać dyfrakcję fali po wyjściu z kanału.

Samolot lecący z prędkością naddźwiękową wytwarza układ fal uderzeniowych i rozrzedzeniowych, które w wyniku wzajemnego oddziaływania docierają do powierzchni Ziemi jako tzw. fala N, o dwóch skokowych przyrostach ciśnienia odbieranych przez ucho ludzkie jako dwa odgłosy (rys. 4) z przerwą λ wynoszącą dla samolotu myśliwskiego 0,1 s a dla dużego samolotu transportowego 0,35 s. Wielkość przyrostu ciśnienia i poziom wytwarzanego dźwięku zależą od wielu czynników, np. od ciężaru i wymiarów samolotu, wysokości i prędkości lotu, rozkładu temperatury w atmosferze, turbulencji powietrza,

fala N



Rys. 4. Fala ciśnieniowa N wytwarzana przez samolot lecący z prędkością naddźwiękową

przyspieszenia samolotu i ukształtowania powierzchni Ziemi. Dodatni przyrost ciśnienia w fali N może w szczególnych wypadkach być równoważny niszczącemu działaniu huraganu (o prędkości ok. 35 m/s) na ludzi, pojazdy, budynki i środowisko naturalne. Za tolerowane obciążenia uważa się obciążenia odpowiadające silnemu wiatrowi (prędkość 12 m/s). Jeżeli występują one jednak często, np. w rejonach lotnisk, to mogą przyspieszać starzenie się budynków, a ponadto przez nagłe pojawianie się powodować wypadki pojazdów oraz, ze względu na hałas, wywoływać u ludzi nerwice i choroby serca. Aby zapewnić bezpieczne warunki życia na Ziemi wprowadza się różne ograniczenia dotyczące lotów samolotów naddźwiękowych.

parametry gazu za falą uderzeniową

Istotne jest określenie parametrów gazu za falą uderzeniową: prędkości, temperatury, ciśnienia i gęstości. Natężenie fali uderzeniowej zależy od prędkości jej ruchu względem gazu. Jeśli ruch fali uderzeniowej będziemy rozpatrywać względem nieruchomego układu odniesienia (rys. 5a), to jej prędkość można wyrazić przez sumę prędkości gazu przed falą i prędkości fali względem gazu ($u_1 + W$). Wyróżnimy obszar *1* przed falą i obszar *2* za falą. W obszarze *1* prędkość gazu oznaczymy przez u_1 i prędkość dźwięku przez a_1, a w obszarze *2* — przez u_2 i a_2, przy czym $u_2 > u_1$ i $a_2 > a_1$. Jak już wspominaliśmy prędkość fali uderzeniowej względem gazu przed jej czołem jest zawsze naddźwiękowa, czyli $W > a_1$, a więc

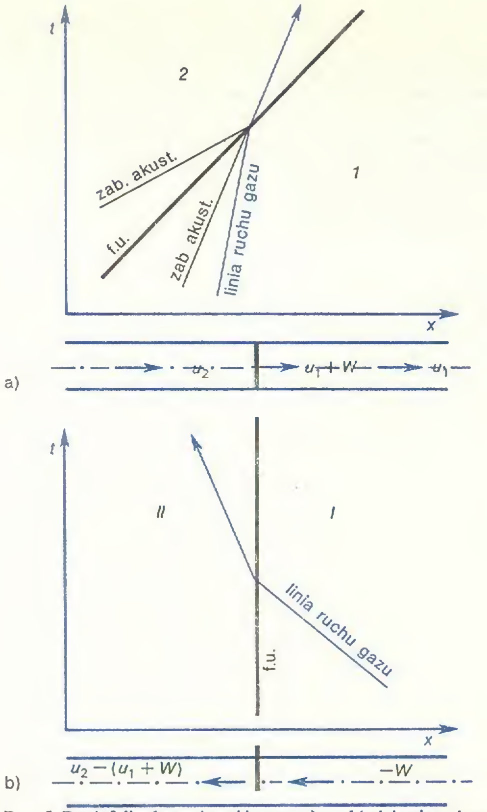
$$u_1 + W > u_1 + a_1.$$

Można również wykazać, że

$$u_1 + W < u_2 + a_2.$$

Z nierówności tych wynika, że fala uderzeniowa dogania zaburzenia akustyczne biegnące przed nią i jest doganiana przez biegnące za nią, co oznacza wchłanianie przez falę zaburzeń biegnących w jej otoczeniu.

Zwiążmy teraz układ odniesienia z ruchomą falą; w układzie tym ruch gazu jest ustalony (jak w wypadku fali stacjonarnej) a jego prędkość jest równa $-W$ w obszarze *I* i $u_2 - (u_1 + W)$ w obszarze *II* (rys.



Rys. 5. Ruch fali uderzeniowej i gazu: a) w układzie nieruchomym, b) w układzie związanym z falą

5b). Stosując do gazu przepływającego przez falę uderzeniową zasady zachowania masy, pędu i energii otrzymamy następujące równania:

$$\varrho_1 W = \varrho_2[W \pm (u_1 - u_2)], \tag{1}$$

$$p_1 + \varrho_1 W^2 = p_2 + \varrho_2[W \pm (u_1 - u_2)]^2, \tag{2}$$

$$\frac{W^2}{2} + h_1 = \frac{[W \pm (u_1 - u_2)]^2}{2} + h_2, \tag{3}$$

w których ϱ jest gęstością gazu, p — ciśnieniem, a h — entalpią ($h = e + (p/\varrho)$, gdzie e — energia wewnętrzna). Znak plus odnosi się do fali biegnącej w dodatnim kierunku osi x, jak na rys. 5a, natomiast minus — do fali biegnącej w kierunku przeciwnym.

parametry gazu — fale słabe

W wypadku słabych fal uderzeniowych (o małym natężeniu), tzn. w zakresie ciśnień i temperatur, przy których wartości ciepła właściwego (przy stałym ciśnieniu i przy stałej objętości) i ich stosunek $\varkappa$ można przyjąć za stałe dla danego gazu, z powyższych równań uzyskuje się zależności wyznaczające: przyrost prędkości gazu

$$\frac{u_2 - u_1}{a_1} = \pm \frac{2}{\varkappa + 1}\left(M_s - \frac{1}{M_s}\right); \tag{4}$$

stosunki temperatur bezwzględnych ($T = 273 + t$°C)

$$\frac{T_2}{T_1} = \left(\frac{a_2}{a_1}\right)^2 = $$

$$= 1 + \frac{2(\varkappa - 1)}{(\varkappa + 1)^2}\left[\varkappa M_s^2 - \frac{1}{M_s^2} - (\varkappa - 1)\right]; \tag{5}$$

stosunki ciśnień

$$\frac{p_2}{p_1} = 1 + \frac{2\varkappa}{\varkappa + 1}(M_s^2 - 1); \tag{6}$$

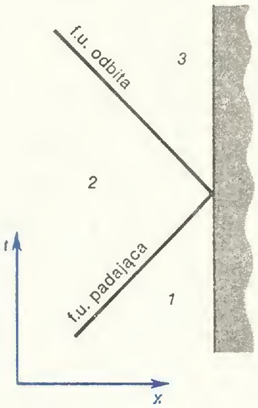
stosunki gęstości

$$\frac{\varrho_2}{\varrho_1} = \frac{1}{1 - \dfrac{2}{\varkappa + 1}\left(1 - \dfrac{1}{M_s^2}\right)}, \tag{7}$$

gdzie $M_s = W/a_1$ jest liczbą Macha fali uderzeniowej. Wyrażenia (4)-(7) są słuszne dla fali płaskiej, cylindrycznej i kulistej.

Obliczmy na przykład parametry gazu za falą i po jej odbiciu od przeszkody, tzn. za falą odbitą, w wypadku fali o $M_s = 2$ (wywołanej pęknięciem zbiornika z gazem będącym pod ciśnieniem kilku MPa lub eksplozją bomby czy pocisku), biegnącej w nieruchomym ($u_1 = 0$) powietrzu ($\varkappa = 1{,}4$) o temperaturze $t_1 = 15$°C ($a_1 = 340$ m/s) i ciśnieniu 0,1 MPa. Z zależności (4)-(6) mamy $u_2 = 425$ m/s, $t_2 = 213$°C i $p_2 = 0{,}45$ MPa. Fala trafiająca na przeszkodę ulega odbiciu, a prędkość przepływu za nią maleje do zera (rys. 6). Wykorzystując warunek $u_3 = 0$, z (4) możemy obliczyć M_s fali odbitej, a z (5) i (6) znaleźć odpowiednio $t_3 = 446$°C i $p_3 = 1{,}5$ MPa. Otrzymane wartości wskazują, że nawet fala uderzeniowa o małym natężeniu stwarza poważne obciążenia obiektów stojących na jej drodze; w rozpatrywanym przykładzie obciążenie 153 t/m² ściany jest obciążeniem niszczącym każdy budynek.

obciążenie obiektów przez falę uderzeniową



Rys. 6. Odbicie płaskiej fali uderzeniowej od ściany

W wypadku fal uderzeniowych o dużym natężeniu, gdy wartości ciepła właściwego zmieniają się wraz z ciśnieniem i temperaturą, parametry gazu za falą możemy wyznaczyć bezpośrednio z równań (1)-(3), wykorzystując tablice lub wykresy podające zależności entalpii od gęstości i ciśnienia dla danego gazu.

Fale stacjonarne (nieruchome)

W strumieniu gazu wypływającym z dyszy i rozpędzonym do prędkości naddźwiękowej mogą powstawać stacjonarne fale uderzeniowe o różnych kształtach — zależnie od stosunku ciśnienia zasilania do ciśnienia otoczenia. Na il. 173 (tabl. 45) jest widoczny strumień gazu wypływający z prędkością krytyczną z dyszy zbieżnej o przekroju kołowym i osiągający dalej prędkość naddźwiękową (stosunek ciśnień 7,5). W pobliżu dyszy powstaje baryłkowata fala uderzeniowa zakończona płaską falą uderzeniową, zwaną dyskiem Macha. Przy wypływie do kanału z dyszy zbieżnej o przekroju prostokątnym powstaje układ skośnych fal uderzeniowych (il. 171, tabl. 45), rozpoczynający się od miejsca, w którym granica strumienia styka się ze ścianką kanału (lewa strona fotografii). Prążki oznaczają linie stałej gęstości gazu; ich zagęszczenie świadczy o dużej zmianie gęstości. Obraz prążków w otoczeniu wylotu dyszy D wskazuje na malejącą gęstość rozprężającego się gazu, a ich zagęszczenie na liniach skośnych — na wzrost gęstości gazu przy przepływie przez falę uderzeniową.

dysk Macha

skośne fale uderzeniowe

Ruch bryły ze stałą prędkością naddźwiękową

w nieruchomym gazie powoduje utworzenie fali uderzeniowej, stacjonarnej względem bryły. Identyczna sytuacja ma miejsce, gdy nieruchoma bryła jest opływana przez gaz. Takie fale są widoczne na il. 177, 178, 172, 176 (tabl. 45 i 46).

odsunięta fala uderzeniowa

Przed kulą (il. 177), poruszającą się z prędkością nieco większą od prędkości dźwięku, w odległości trochę większej od jej średnicy powstaje zakrzywiona i odsunięta fala uderzeniowa S_b, a na kuli następuje oderwanie warstwy przyściennej gazu i tworzy się silnie burzliwy obszar zastoju W. Za odsuniętą falą uderzeniową przepływ względem kuli jest poddźwiękowy, ale przy opływie kuli prędkość wzrasta ponownie do naddźwiękowej. Fakt ten i wychylenie granicy obszaru zastoju na zewnątrz powoduje powstanie stożkowej fali uderzeniowej S_t.

stożkowa fala uderzeniowa

Zakrzywiona i odsunięta fala uderzeniowa S_b powstaje również przed zaokrąglonym stożkiem (il. 178) poruszającym się w powietrzu z prędkością naddźwiękową. Fala ta znajduje się blisko stożka. Ponadto są widoczne: fala rozrzedzeniowa R przy podstawie stożka, burzliwy obszar zastoju T, stożkowa fala uderzeniowa S_t wywołana zakrzywieniem się granicy śladu aerodynamicznego oraz fale akustyczne W spowodowane burzliwością w tym śladzie.

Ilustracja 172 przedstawia opływ kabiny kosmicznej z hiperdźwiękową prędkością (badanie modelowe prowadzone w tunelu aerodynamicznym). Za odsuniętą falą uderzeniową jest widoczny jarzący się gaz o wysokiej temperaturze (5000°C).

Fale uderzeniowe powstające przy płaskim opływie romboidalnego profilu (kąt ostry wynosi 14°) z naddźwiękową prędkością widoczne są na il. 176. Z czołowej krawędzi wychodzą skośne fale uderzeniowe, a w rejonie krawędzi bocznych widać fale rozrzedzeniowe.

parametry prostopadłej fali uderzeniowej

Parametry prostopadłej fali uderzeniowej można wyznaczyć w prosty sposób w nieruchomym układzie odniesienia (rys. 5b). Równania zachowania w tym układzie są następujące:

$$\varrho_{I} u_{I} = \varrho_{II} u_{II}, \tag{1a}$$

$$p_{I} + \varrho_{I} u_{I}^2 = p_{II} + \varrho_{II} u_{II}^2, \tag{2a}$$

$$(u_{I}^2/2) + h_{I} = (u_{II}^2/2) + h_{II}; \tag{3a}$$

u_{I} i u_{II} — prędkości gazu za i przed falą. Z powyższych równań zamiast wzoru (4) wyznaczamy związek między liczbami Macha przed i za falą

$$M_{II}^2 = \left(M_{I}^2 + \frac{2}{\varkappa - 1}\right) \Big/ \left(\frac{2\varkappa}{\varkappa - 1} M_{I}^2 - 1\right). \tag{8}$$

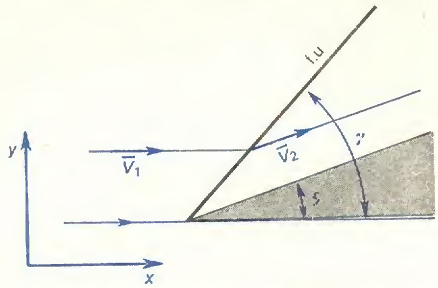
Do określenia stosunków temperatur, ciśnień i gęstości wykorzystujemy zależność (5), (6) i (7) podstawiając do nich $M_s = M_{I}$.

Ze wzoru (8) wynika, że $M_{II} = M_{I}$, gdy $M_{I} = 1$, co oznacza, że przy przepływie z prędkością dźwięku nie powstaje stacjonarna fala uderzeniowa, podobnie jak przy $M = 1$ ruchoma fala uderzeniowa staje się zaburzeniem akustycznym. Gdy $M_{I} \to \infty$, to $M_{II} \to \sqrt{(\varkappa - 1)/2\varkappa}$, co przy $\varkappa = 1,4$ daje $M_{II} \to 0,378$. A więc zakresowi $1 < M < \infty$ odpowiada zakres $1 > M_{II} > 0,378$. Ze wzorów (6) i (7) wynika, że $p_{II}/p_{I} \to \infty$, lecz $\varrho_{II}/\varrho_{I} \to (\varkappa+1)/(\varkappa-1)$ a przy $\varkappa = 1,4$ $\varrho_{II}/\varrho_{I} \to 6$. Z powyższych rozważań wynika, że przepływ przed prostopadłą falą uderzeniową jest zawsze naddźwiękowy ($M_{I} > 1$), a za nią poddźwiękowy ($M_{II} < 1$) oraz, że adiabatyczne sprężanie gazu w fali uderzeniowej może zwiększyć gęstość tylko w ograniczonym zakresie w odróżnieniu od sprężania izentropowego, przy którym wzrost gęstości nie jest ograniczony.

Wykorzystując równania zachowania w odniesieniu do skośnej fali uderzeniowej wywołanej zmianą kierunku płaskiego przepływu o kąt θ (rys. 7) można wykazać, że zależności między parametrami gazu przed i za falą są identyczne jak dla fali prostopadłej, jeżeli się rozpatruje składowe prędkości w kierunku prostopadłym do fali skośnej. Możliwość występowania skośnej fali uderzeniowej jest ograniczona zakresem wartości M_1 i θ. Z równań zachowania wynika równanie rodziny krzywych, z których każda odpowiada określonej prędkości gazu przed skośną falą uderzeniową, a opisuje wektor prędkości gazu za falą w zależności od kąta θ.

biegunowe skośnej fali uderzeniowej

Krzywe te, nazywane biegunowymi skośnej fali uderzeniowej,



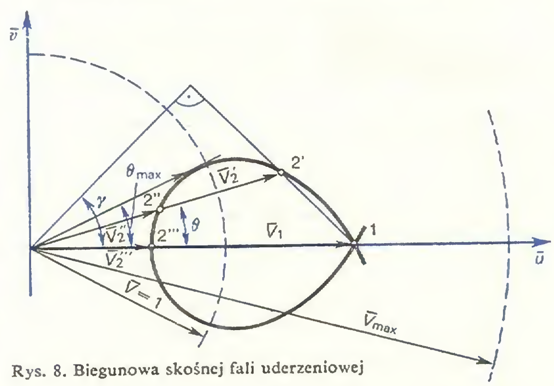
Rys. 7. Skośna fala uderzeniowa przy opływie klina

przedstawia się na ogół w układzie współrzędnych u i v będących składowymi w kierunku odpowiednio x i y prędkości gazu $\overline{V}$, lub w układzie współrzędnych bezwymiarowych $\overline{u} = u/a^*$ i $\overline{v} = v/a^*$. (a^* krytyczna prędkość dźwięku, tzn. prędkość odpowiadająca liczbie Macha $M = 1$). Równanie biegunowej w tym układzie ma postać:

równanie biegunowej

$$\overline{v}_2^2 = \frac{(\overline{u}_1 - \overline{u}_2)^2 (\overline{u}_1 \overline{u}_2 - 1)}{\frac{2}{\varkappa + 1} \overline{u}_1^2 - \overline{u}_1 \overline{u}_2 + 1}. \tag{9}$$

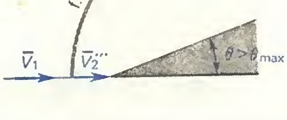
Parametrem jest tu $\overline{u}_1$ równe $\overline{V}_1$, ponieważ kierunek przepływu niezakłóconego pokrywa się z osią x, a więc $\overline{v}_1 = 0$. Na rys. 8 przedstawiono biegunową fali uderzeniowej opisaną wzorem (9). Przy danej wartości



Rys. 8. Biegunowa skośnej fali uderzeniowej

$\overline{V}_1$, w zakresie $0 < \theta < \theta_{max}$, istnieją dwa rozwiązania określone np. punktami *2'* i *2''*. Pierwsze z nich odpowiada mniejszej zmianie prędkości niż drugie, a więc słabszej fali uderzeniowej. Jak wynika z rysunku, za słabszą falą przepływ jest nadal naddźwiękowy ($\overline{V}_2' > 1$), a za silniejszą — poddźwiękowy ($\overline{V}_2'' < 1$). Jeżeli $\theta = \theta_{max}$, istnieje tylko jedno rozwiązanie, a dla $\theta > \theta_{max}$ nie ma rozwiązania dla skośnej fali uderzeniowej. Stwierdzono doświadczalnie, że przy

Rys. 9. Odsunięta fala uderzeniowa przy opływie klina



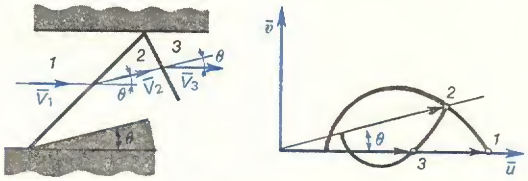
opływ klina

opływie klina swobodnym strumieniem występuje tylko słaba fala, natomiast w kanale zamkniętym, jeśli się wytworzy odpowiedni poziom ciśnienia za

falą uderzeniową, można wymusić pojawienie się silniejszej fali. Gdy $\theta > \theta_{max}$ występuje fala odsunięta (rys. 9), jak przed bryłą tępą (il. 172, 177, 178) a prędkość $\overline{V}_2'''$ gazu za nią, w miejscu gdzie jest ona prostopadła, określa punkt *2'''* (rys. 8).

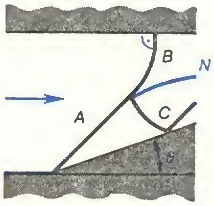
Skośna fala uderzeniowa ulega odbiciu od ściany narzucającej kierunek przepływu (rys. 10). Prędkość $\overline{V}_3$ za falą odbitą zależy od $\overline{V}_2$ i odchylenia strug z powrotem o kąt θ. Gdy biegunowa wychodząca z punktu *2* (rys. 10b) nie przecina osi odciętych, to nie jest możliwe odbicie fali skośnej i wówczas powstaje tzw. odbicie Macha (porównaj dysk Macha na

odbicie Macha



Rys. 10. Odbicie skośnej fali uderzeniowej (a) i odpowiednie biegunowe (b)

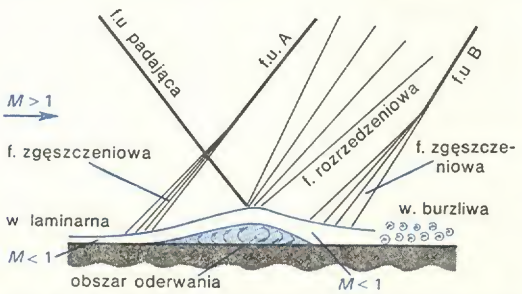
il. 173, tabl. 45) składające się z trzech fal uderzeniowych *A*, *B*, *C* i powierzchni *N* nieciągłości prędkości, gęstości i temperatury gazu (rys. 11). Prędkość $\overline{V}_2$ jest w tym wypadku zbyt mała, aby przy danym kącie θ mogła utworzyć się odbita skośnie fala uderzeniowa. Podobnie przy wypływie z dyszy nie jest możliwe odbicie stożkowej fali uderzeniowej w osi symetrii i po obu jej stronach powstają struktury przepływu (il. 173) odpowiadające schematowi na rys. 11.



Rys. 11. Odbicie Macha; *A*, *B*, *C* fale uderzeniowe, *N* powierzchnia nieciągłości prędkości, gęstości i temperatury gazu

wpływ warstwy przyściennej

Rozpatrując fale uderzeniowe występujące w pobliżu opływanych powierzchni nie zawsze można pominąć obecność warstwy przyściennej wywołanej lepkością gazu, tzn. cienkiej warstwy, w której prędkość przepływu rośnie gwałtownie (od zera na powierzchni) w miarę oddalania się od niej. Wskutek obecności warstwy przyściennej, odbicie skośnej fali uderzeniowej od ciała przebiega w nieco inny sposób niż to pokazano na rys. 10. Gdy warstwa ta jest bardzo cienka, to można jej wpływu nie uwzględniać. Jak już wiadomo, stacjonarna fala uderzeniowa pojawia się tylko w przepływie naddźwiękowym ($M > 1$), a więc musi się kończyć na granicy warstwy (rys. 12), w której gaz porusza się z prędkością poddźwiękową ($M <$ < 1). Wskutek tego wzrost ciśnienia występujący za falą uderzeniową przenika przez poddźwiękową warstwę laminarną (czyli uwarstwioną, nieburzliwą) pod prąd powodując wzrost jej grubości, a nawet oderwanie od ściany i utworzenie obszaru wypełnionego wirami. Pogrubianie warstwy poddźwiękowej działa na przepływ naddźwiękowy jak zakrzywienie ściany wytwarzając falę zgęszczeniową, która w pewnej odległości przeradza się w falę uderzeniową *A*. Fala padająca docierając do oderwanej warstwy poddźwiękowej powoduje z kolei jej odchylenie w kierunku ściany, co jest przyczyną powstawania fali rozrzedzeniowej. Zetknięcie się warstwy ze ścianą generuje falę zgęszczeniową, która przeradza się w falę uderzeniową *B*. W wyniku tych wszystkich procesów poddźwiękowa warstwa staje się grubsza i burzliwa. Ostatecznie w stosunkowo niewielkiej odległości od ściany fale uderzeniowe *A* i *B* oraz znajdująca się między nimi fala rozrzedzeniowa spotykają się i tworzą pojedynczą, odbitą falę uderzeniową.



Rys. 12. Współoddziaływanie skośnej fali uderzeniowej z warstwą przyścienną

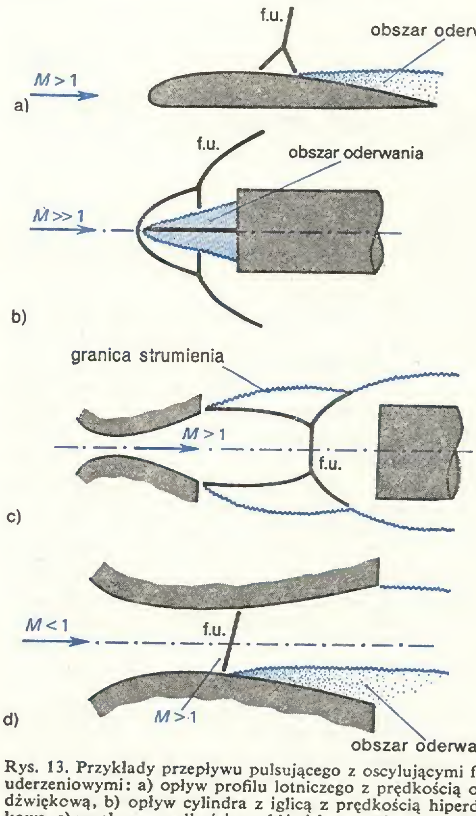
Fale oscylujące

pulsacje fal stacjonarnych

Stacjonarne fale uderzeniowe występujące w przepływie burzliwym nie są nigdy idealnie nieruchome. Wskutek pulsacji prędkości związanej z burzliwością drgają one ze znaczną częstością, ale z bardzo małą amplitudą. W pewnych jednak wypadkach w stacjonarnym naddźwiękowym przepływie mogą pojawić się silne oscylacje fal uderzeniowych, przy których występują duże zmiany położenia oraz natężenia fal, a czasem nawet ich chwilowy zanik.

Ilustracja 179 (tabl. 46) ukazuje zmiany (w czasie jednego okresu) struktury przepływu w kanale o przekroju prostokątnym zasilanym w sposób ciągły przez dyszę zbieżną. Gdy ciśnienie w obszarze zastoju tuż przy dyszy jest małe, to w strumieniu występuje pojedyncza zakrzywiona fala uderzeniowa (il. 179a). Następnie strumień odrywa się od ściany kanału (il. 179b), ciśnienie rośnie, a fala zanika. Stopniowo tworzy się struktura komórkowa (il. 179c, d), która w mia-

struktura komórkowa



Rys. 13. Przykłady przepływu pulsującego z oscylującymi falami uderzeniowymi: a) opływ profilu lotniczego z prędkością okołodźwiękową, b) opływ cylindra z iglicą z prędkością hiperdźwiękową, c) wypływ z prędkością naddźwiękową z dyszy na cylinder, d) przepływ dyszą zbieżno-rozbieżną

rę spadku ciśnienia przeradza się z powrotem w falę uderzeniową (il. 179e, f, g) i cykl się powtarza. W tym wypadku pulsacja przepływu przejawia się zasadniczymi zmianami jego struktury oraz silną oscylacją ciśnienia w całym kanale.

mechanizm pulsacji przepływu

Mechanizm opisanej pulsacji przepływu jest prawdopodobnie następujący. Przy pewnej wartości stosunku ciśnienia zasilania do ciśnienia w obszarze zastoju przy dyszy, rejon przyklejenia strumienia do ścianki kanału powinien znajdować się w określonym miejscu zapewniającym stałość ilości gazu w obszarze zastoju. Inaczej mówiąc, ilość gazu wypływająca wskutek mieszania z obszaru zastoju musi być równa ilości dopływającej z powrotem z rejonu przyklejenia. Odpowiadająca wspomnianemu stosunkowi ciśnień struktura strumienia jest jednak niestabilna i wskutek wzajemnego oddziaływania może tworzyć się tzw. pętla sprzężenia zwrotnego między ciśnieniem w obszarze zastoju, strukturą strumienia i położeniem rejonu przyklejania. Przypadkowe zaburzenie obniżające ciśnienie w obszarze zastoju zwiększa rozprężanie strumienia przy dyszy i odpływ gazu z tego obszaru, co z kolei pogłębia dalej obniżenie ciśnienia. W rezultacie rejon przyklejenia i związana z nim fala uderzeniowa przesuwa się pod prąd w kierunku dyszy. Ciśnienie za falą uderzeniową rośnie na tyle, że poddźwiękowa warstwa odrywa się od ściany i pojawia się silny powrotny przepływ do obszaru zastoju, powodujący wzrost ciśnienia. W rezultacie strumień przy dyszy zwęża się, a rejon przyklejenia oddala od niej. Stopniowo wskutek wzrostu ciśnienia prędkość przepływu w kanale rośnie, a odpływ z obszaru zastoju zwiększa się, to z kolei powoduje obniżenie się ciśnienia i cykl się powtarza. Przy danym stosunku przekroju dyszy do przekroju kanału pulsacja przepływu ma pewne pasmo częstości ale największe amplitudy oscylacji ciśnienia i fal uderzeniowych występują przy określonej długości kanału, który odgrywa rolę rezonatora.

występowanie fal oscylujących

Podobnego typu oscylacje fal uderzeniowych są obserwowane np. przy okołodźwiękowym (M nieco większe od jedności) opływie profilu lotniczego (rys. 13a), przy hiperdźwiękowym ($M \gg 1$) opływie ciała tępego z iglicą (rys. 13b), przy napływie strumienia naddźwiękowego na przeszkodę (rys. 13c) bądź przy wypływie z dyszy zbieżno-rozbieżnej, tzw. dyszy Lavala. Wszystkie te przypadki odznaczają się: niestabilnymi w danych warunkach strukturami fal uderzeniowych, oddzielających obszar naddźwiękowy od poddźwiękowego, obecnością warstwy mieszania na granicy obszaru oderwania (rys. 13a, b, d) lub na granicy strumienia (rys. 13c) oraz wzmożonym wytwarzaniem hałasu o dyskretnych częstościach. Regularne i silne pulsacje przepływu występują wtedy, gdy istnieje sprzężenie zwrotne między poszczególnymi obszarami przepływu, z wystarczającym wzmocnieniem oddziaływań. Zakresy częstości oscylacji fal są zależne od wymiarów geometrycznych ciał i parametrów przepływającego gazu.

I. I. Glass *Fale uderzeniowe i człowiek*, Warszawa 1980; W. M. Jungowski *Some self induced supersonic flow oscillations*, J. Prog. Aerospace Sci. 18, 151 (1978); A.H. Shapiro *The Dynamics and Thermodynamics of Compressible Fluid Flow*, New York 1953–54; J. B. Zieldowicz, J. P. Rajzer *Fizyka udarnych wołn i wysokotiempieraturnych gidrodinamiczeskich jawlenij*, Moskwa 1966.

Hałas

Stefan Czarnecki

Hałasem jest każdy dźwięk, który działa ujemnie na człowieka. Oddziaływanie to może mieć charakter przeszkadzający, uciążliwy lub szkodliwy, co zależy przede wszystkim od natężenia hałasu oraz czasu jego oddziaływania na człowieka.

Hałasy przeszkadzające powodują zmęczenie, utudniają wypoczynek, rozpraszają uwagę, zmniejszają zrozumiałość mowy, a także u niektórych ludzi mogą działać ujemnie na psychikę. Hałasy uciążliwe o większym natężeniu lub działające w sposób długotrwały — oprócz wymienionych wyżej skutków są przyczyną okresowego, odwracalnego przytępienia słuchu, a także stają się źródłem schorzeń o podłożu nerwicowym. Hałasy szkodliwe prowadzą do całkowitej głuchoty i wywołują poważne schorzenia o podłożu nerwicowym. Ponadto prawie całkowicie uniemożliwiają ustne porozumienie się.



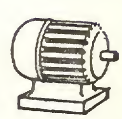
75-85

Ze względu na charakter przebiegu ciśnienia akustycznego w funkcji czasu (rys. 1) hałasy można podzielić na: periodyczne i nieperiodyczne — w zależności od tego, czy w przebiegu ciśnienia akustycznego występują lub nie występują regularne powtarzalności, na: stacjonarne i niestacjonarne — w zależności od tego, czy poziom hałasu jest stały lub zmienny oraz na hałas ciągły i przerywany — w zależności od tego, czy hałas jest długotrwały, czy też występuje tylko w pewnych przedziałach czasu. Jeśli przedziały te są bardzo krótkie (poniżej 1 s), wówczas hałas nosi nazwę hałasu impulsowego.



80-90

Hałasy nie są przebiegami sinusoidalnymi (tonami), lecz wykazują strukturę złożoną. Każdy hałas stacjonarny można poddać analizie Fouriera i przedstawić w postaci rozkładu poziomu ciśnienia akustycznego w funkcji częstości, czyli w postaci widma częstości. W przypadku przebiegów periodycznych jest to rozkład amplitud o określonych częstościach tworzący widmo prążkowe; w przypadku przebiegów nieperiodycznych jest to widmo ciągłe (rys. 2), uzyskane przez obliczenie wartości poziomu ciśnienia akustycznego w odpowiednich przedziałach (pasmach) częstości Δf. Przy $\Delta f \to 0$ rozkład poziomu ciśnienia akustycznego nazywa się gęstością widma. Przykładem hałasów periodycznych jest gwizd lokomotywy, pisk hamulców; przykładem hałasów nieperiodycznych — szum płynącej wody, syk pary, hałas silnika odrzutowego.

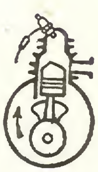


80-90

Widmo częstości hałasów niestacjonarnych lub impulsowych zmienia się w funkcji czasu, co prowadzi do konieczności analizy hałasu zarówno w zależności od częstości jak i czasu.

Ocena i pomiar hałasu

Ocena hałasu za pomocą wielkości fizycznych nie pokrywa się z odczuciem subiektywnym przez człowieka. Stąd też wynika konieczność stosowania dwóch rodzajów kryteriów oceny hałasu: obiektywnej — polegającej na pomiarze parametrów fizycznych hałasu, oraz subiektywnej — prowadzącej do oceny hałasu za pomocą jednej wielkości, która przy równoczesnym uwzględnieniu kilku czynników odzwierciedlałaby w przybliżeniu wrażenia psychoakustyczne. Kryteria subiektywne określa się przez uśrednienie wyników badań wielu ludzi.



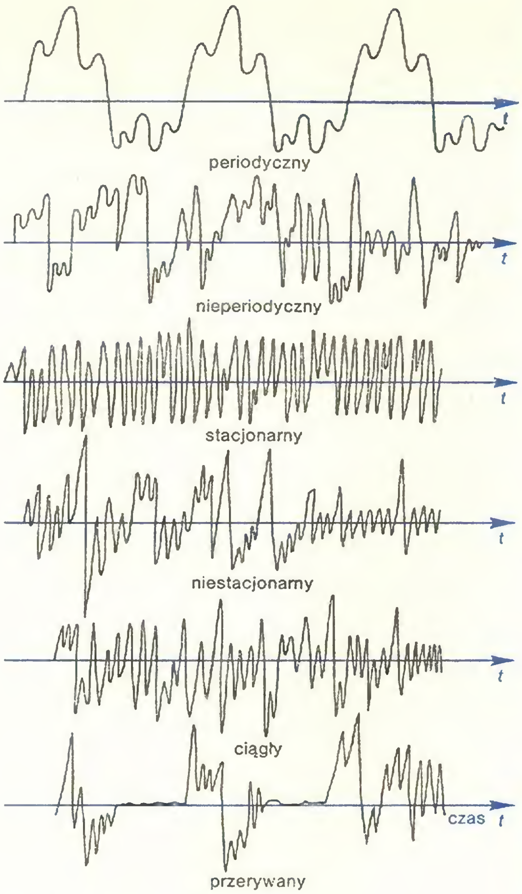
90-100

Podstawową wielkością określającą subiektywne odczucie dokuczliwości hałasu jest hałaśliwość wyrażana w noysach. Zróżnicowanie odczucia stopnia hałasu przy różnych częstościach obrazują krzywe jednakowej hałaśliwości (rys. 3). Można z nich odczytać wartości natężenia dźwięku lub ciśnienia akustycznego, przy których hałasy o różnych częstościach są jednakowo dokuczliwe.

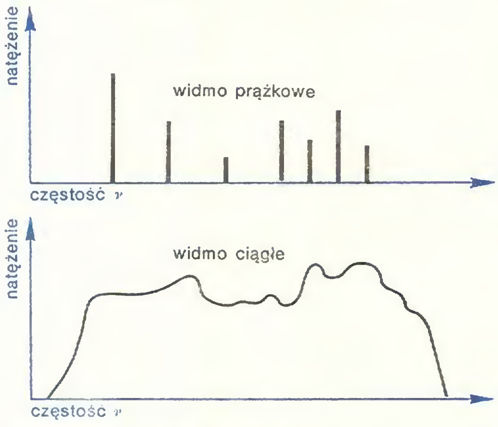


95-105

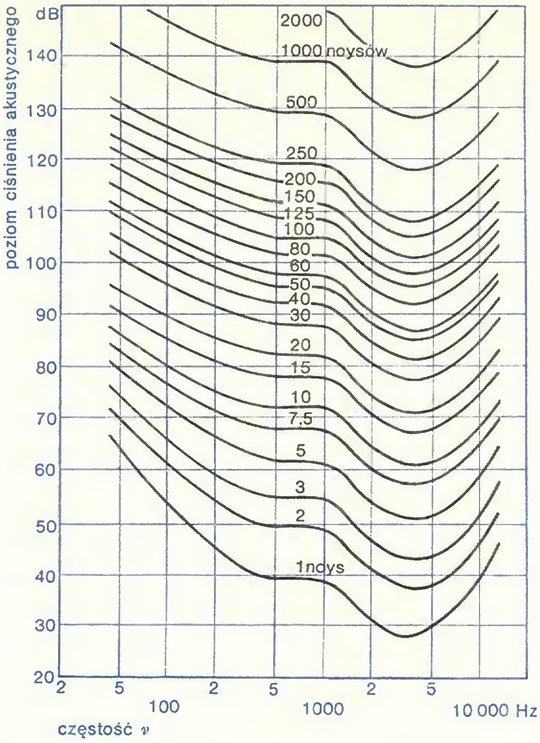
Częściej stosowaną wielkością określającą subiektywne odczucie dźwięków jest głośność. Zmiany w odczuciu głośności są proporcjonalne w przybli-



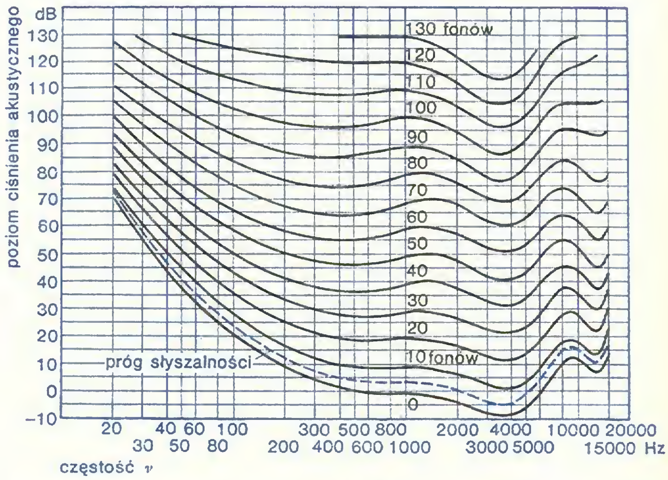
Rys. 1. Rodzaje hałasu



Rys. 2. Rodzaje widm częstości hałasu



Rys. 3. Krzywe jednakowej hałaśliwości



Rys. 4. Krzywe jednakowego poziomu głośności

żeniu do ilorazu natężeń dźwięku, a nie do ich różnic, dlatego głośność wyraża się w skali logarytmicznej (tak jak i wielkości fizyczne — natężenie dźwięku i ciśnienie akustyczne; → Przedmiot i zakres akustyki).

Poziom głośności dźwięku L_s określony jest następująco:

$$L_s = 10 \lg \frac{I_s}{I_{s0}},$$

gdzie I_s i I_{s0} — subiektywnie odczuwane natężenie dźwięku i subiektywnie odczuwana wartość progowa (granica słyszalności). Poziom głośności czystych tonów (dźwięków sinusoidalnych) lub dźwięków wąskopasmowych określa się w fonach za pomocą krzywych jednakowego poziomu głośności, tzw. krzywych izofonicznych (rys. 4). Interpretuje się je w sposób analogiczny jak krzywe jednakowej hałaśliwości.

Do oceny hałasów niestacjonarnych lub przerywanych stosuje się pojęcie poziomu ekwiwalentnego wyrażonego zależnością:

$$L_{eq} = \frac{q}{0,3} \lg \frac{1}{T} \sum_{i=1}^{n} t_i \cdot 10^{0,3 L_i/q},$$

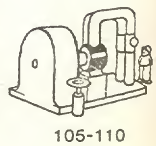
gdzie t_i jest czasem oddziaływania hałasu o poziomie L_i, T — całkowitym czasem oddziaływania hałasu, q — współczynnikiem zależności od charakteru hałasu (q = 3–4). Do pomiaru poziomu ekwiwalentnego używa się dozymetrów hałasu umożliwiających określenie czasu, po którym przekroczony jest poziom dopuszczalny.



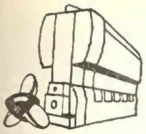
95-105



100-110



105-110



110-115

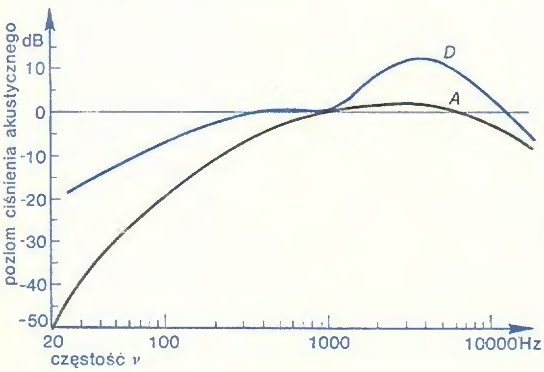
Poziom głośności hałasu nazywany jest zazwyczaj poziomem hałasu. Metody pomiaru hałasu, które dawałyby wyniki odpowiadające jego odczuciu subiektywnemu można podzielić na metody bezpośrednie, polegające na stosowaniu przyrządu z wbudowanym układem korekcyjnym uwzględniającym własności słuchu, oraz metody pośrednie polegające na pomiarze wielkości fizycznych z zastosowaniem odpowiedniej interpretacji wyników.

W obu metodach podstawowymi elementami przyrządu pomiarowego są: mikrofon, który daje na wyjściu wielkości elektryczne proporcjonalne do ciśnienia akustycznego, wzmacniacz, dzielnik zakresów wycechowany w decybelach oraz przyrząd wskazówkowy.



110-120

Do określenia krzywych korekcyjnych dla metod bezpośrednich korzysta się z krzywych jednakowego poziomu głośności. Jak wynika z rys. 4 krzywe te nie przebiegają równolegle, co powoduje, że krzywe korekcyjne filtru uwzględniające zmiany czułości słuchu w funkcji częstości powinny mieć różny kształt przy różnych wartościach poziomu głośności. Dla uproszczenia stosuje się jedną uniwersalną krzywą korekcyjną, tzw. krzywą *A* (rys. 5). Zastosowanie w przyrządzie pomiarowym filtru korekcyjnego o powyższej charakterystyce powoduje mniejsze wzmocnienie w zakresie częstości niskich, a tym samym daje wynik pomiaru zbliżony do subiektywnej oceny poziomu głośności. Dla podkreślenia, że pomiar został wykonany przy użyciu filtru korekcyjnego *A* uzyskane wyniki podawane są w dB (*A*). Oprócz krzywej *A* stosowane są również inne krzywe korekcyjne, z których najnowsza jest krzywa *D* (rys. 5) uwzględniająca krzywe jednakowej hałaśliwości.



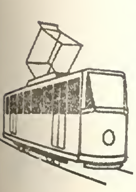
Rys. 5. Krzywe korekcyjne dla poziomów głośności hałasu



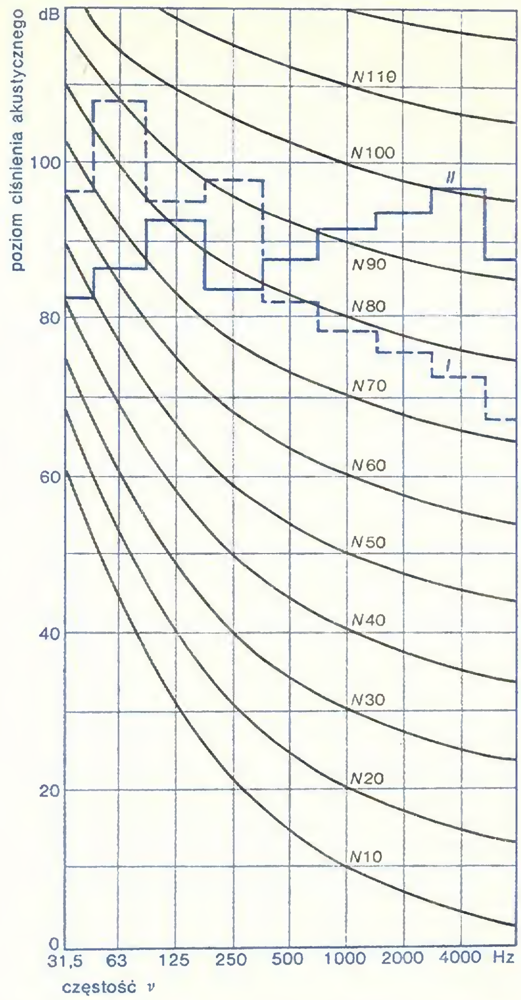
65-75

Najprostszą metodą pośrednią oceny hałasu jest metoda liczb *N* (rys. 6). Metoda ta wymaga zmierzenia widma częstości hałasu, które przy zastosowaniu tej samej skali decybelowej nanosi się na krzywe *N*. Widmo częstości mierzy się w pasmach oktawowych, tzn. takich, których szerokość pasma przepuszczania filtru $f_1 - f_2$ wyrażona jest stosunkiem $f_1/f_2 = 2$, lub w pasmach $^1/_3$-oktawowych, którym odpowiada stosunek częstości skrajnych filtru $f_1/f_2 = \sqrt[3]{2}$. Pomiar polega na przełączaniu kolejnych filtrów i odczytywaniu wartości poziomu ciśnienia akustycznego dla poszczególnych pasm częstości, reprezentowanych przez częstość zawartą w środku pasma.

Liczbę *N* mierzonego hałasu wyznacza się z przecięcia jego widma częstości z najwyżej położną krzywą *N*. Przykład dwóch widm częstości naniesiony na rys. 6 wskazuje, że dla hałasu reprezentowanego przez widmo I (krzywa przerywana) otrzymamy liczbę *N* wynoszącą 95 dB (*N* 95). Dla hałasu reprezentowanego przez widmo II (krzywa ciągła) otrzymamy *N* 100. Wynika stąd, że subiektywnie hałas II jest głośniejszy od hałasu I, mimo że maksymalna wartość poziomu ciśnienia akustycznego widma hałasu II jest o 11 dB niższa od wartości maksymalnej widma hałasu I.



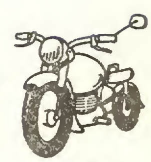
60-75



Rys. 6. Krzywe *N* oceny hałasu z naniesionymi 2 widmami hałasów *I* i *II*

Podstawową zaletą skali dB(*A*) jest prostota pomiaru. Z wyników pomiaru hałasu w dB(*A*) nie wynika jednak, w jakim zakresie częstości hałas jest subiektywnie najgłośniejszy. Informacja ta, bardzo istotna dla wstępnej oceny metod zmniejszenia hałasu, zawarta jest pośrednio w liczbach *N*. Dlatego też skalę dB(*A*) stosuje się powszechnie przy ocenie hałasu, natomiast skalę liczb *N* — przy opracowywaniu koncepcji wyboru metod zabezpieczeń przeciwdźwiękowych. Między poziomem hałasu *L* wyrażonym w dB (*A*) i wyrażonym w liczbach *N* istnieje związek

$$L_N \approx L_{dB(A)} - 5 \text{ dB}.$$



75-90

W celu uwzględnienia dodatkowych czynników, od których zależy subiektywne odczucie hałasu, do oceny hałasu w liczbach *N* stosuje się odpowiednie poprawki zestawione w tabeli.

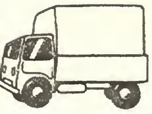
Osoby narażone na długotrwałe działanie hałasów są poddawane co jakiś czas kontrolom jakości słuchu.



80-90

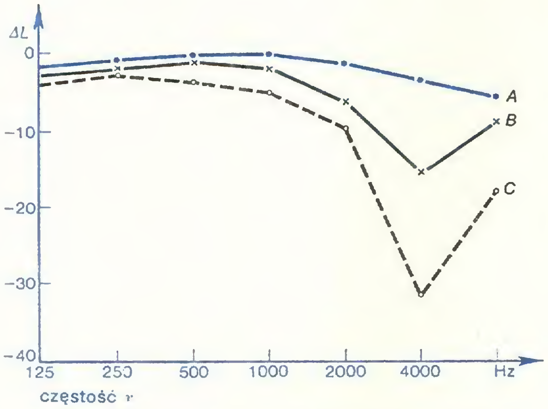
Poprawki ΔL do dopuszczalnych wartości liczb *N* oceny hałasu

Warunki		Poprawka ΔL
Pora dnia	6–22 h	0
	22–6 h	−10
Czas działania hałasu (% doby)	powyżej 50%	0
	12–50%	+5
	3–12%	+12



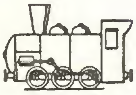
80-95

Są to tzw. pomiary audiometryczne polegające na wyznaczaniu różnicy między progiem słyszalności osoby badanej a uśrednionym progiem słyszalności osób o normalnym słuchu. Sygnały testujące o różnych częstościach i różnych poziomach doprowadzane są za pomocą słuchawek osobno do ucha prawego i lewego. Zadaniem osoby badanej jest zasygnalizowanie momentu usłyszenia dźwięku. Na rys. 7 pokazane są przykładowo trzy krzywe audiometryczne otrzymane po przebadaniu osób o normalnym słuchu (krzywa *A*) oraz o małym (krzywa *B*) i dużym (krzywa *C*) uszkodzeniu słuchu. Krzywe *B* i *C* obrazują najczęstsze przypadki uszkodzenia słuchu wywołanego działaniem



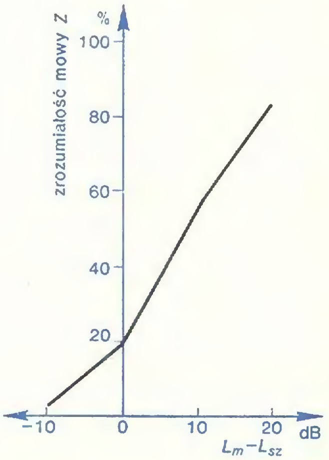
Rys. 7. Przykłady krzywych audiometrycznych; *A* — osoby o normalnym słuchu, *B* — osoby o małym uszkodzeniu słuchu, *C* — osoby o dużym uszkodzeniu słuchu. *L* jest różnicą między progiem słyszalności osoby badanej a uśrednionym progiem słyszalności osób o normalnym słuchu

hałasu. Jest to tzw. trauma, która objawia się selektywnym ubytkiem słuchu w zakresie częstości ok. 4000 Hz. Ponieważ próg słyszalności osób zdrowych zmienia się z wiekiem, dlatego jako odniesienie przy pomiarach audiometrycznych stosowane są krzywe odpowiednio skorygowane do wieku osoby badanej.



80-95

Jednym ze skutków ujemnego działania hałasu na człowieka jest utrudnienie zrozumiałości mowy ludzkiej wpływające w zasadniczy sposób na ograniczenie możliwości porozumienia się, co w efekcie prowadzi do zmniejszenia wydajności pracy.



Rys. 8. Zależność zrozumiałości mowy *Z* od różnicy między poziomem mowy L_m i poziomem szumu L_{sz}

Zrozumiałość mowy *Z*, wyrażona przez stosunek liczby jednostek fonetycznych odebranych prawidłowo do całkowitej liczby jednostek ocenianych, zależy bardzo silnie od różnicy poziomu mowy L_m i poziomu szumu L_{sz}, co ilustruje rys. 8.

Źródła hałasu

Źródła hałasu można podzielić na dwie podstawowe grupy: hałasy pochodzenia mechanicznego, wywołane drganiami mechanicznymi, i hałasy pochodzenia przepływowego (aerodynamiczne lub hydrodynamiczne), wywołane nieregularnościami przepływającej strugi. Nieregularnościami tymi mogą być wiry powstające przy przepływie burzliwym (turbulentnym) lub pulsacje ciśnienia występujące np. przy przepływach naddźwiękowych lub w procesie spalania.



85-95

Przykładem hałasów pochodzenia mechanicznego może być stuk maszyny do pisania, skrzypienie drzwi, audycja z głośnika, przykładem hałasów przepływowych — syk pary, hałasy instalacji wodociągowych, odgłos strzału. W praktyce występują często hałasy pochodzenia mieszanego, wywołane zarówno działaniem drgań mechanicznych jak i przepływem, np. hałas samochodu, odkurzacza czy piły tarczowej.

W procesie powstawania hałasu dużą rolę odgrywają układy rezonansowe, które w przypadku bezpośredniego sprzężenia ze źródłem mogą — w zależności od warunków współdziałania — silnie wzmacniać lub tłumić składowe pewnych częstości odpowiadające częstościom rezonansowym rezonatorów. W przypadku wzmocnienia — powstają w widmie częstości wyraźne maksima, które spowodować mogą znaczny wzrost głośności hałasu.

W zależności od źródła i wpływu układów rezonansowych widmo częstości hałasu może być różne.



90-100

Widmo hałasu mechanicznego kształtowane jest przez charakter drgań elementów mechanicznych jak również przez warunki jego promieniowania, czyli zamiany energii mechanicznej na energię akustyczną. Ostrość występujących maksimów w widmie częstości zależy od stopnia tłumienia układów rezonansowych. Im tłumienie jest większe, tym maksima mniej ostre.

Gdy przepływ strugi odbywa się w warunkach ruchu turbulentnego (przy przekroczeniu liczby Reynoldsa), wytwarzany hałas ma charakter szumu, czyli nie wykazuje wyraźnych maksimów w widmie częstości. Jednak w widmie częstości hałasu aerodynamicznego występują często ostre maksima, wywołane bądź własnościami rezonansowymi układu przepływowego (rezonans rur i komór), bądź wzajemnym oddziaływaniem pola aerodynamicznego z polem akustycznym. Oddziaływanie to prowadzi do akustycznego sprzężenia zwrotnego, które przy zgodności faz powoduje tworzenie się regularnych wirów, będących źródłem hałasu o odpowiedniej częstości. Jednym z częściej występujących wypadków generacji tego typu hałasu jest przepływ powietrza w rurociągu, wewnątrz którego znajduje się ostrze. Wówczas na tle szumu wynikającego z przepływu pojawia się wyraźne maksimum, którego częstość można wyliczyć ze wzoru Strouhala:

$$\nu = Sh\frac{v_p}{a},$$

gdzie *Sh* jest liczbą Strouhala, która dla prędkości przepływu $v_p < 50$ m/s zawiera się w granicach 0,15–0,3, zaś a — wymiarem poprzecznym przeszkody.

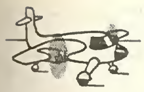


80-95

Ostre maksima w widmie częstości występują także przy wpływie z przewodu, gdy wypływający gaz gwałtownie się rozpręża. Powstaje wówczas struktura komórkowa strumienia na zewnątrz przewodu (il. 179, tabl. 46), która powoduje generację hałasu z wyraźnym maksimum przy jednej częstości wyrażonej zależnością:

$$\nu = \frac{c}{3d\sqrt{\beta - \beta_k}},$$

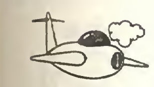
gdzie c jest prędkością rozchodzenia się dźwięku, d — średnicą wylotu rurociągu, β — stosunkiem ciśnienia występującego przy wylocie rurociągu do ciśnienia panującego na zewnątrz, $\beta_k = 1{,}89$ jest wartością krytyczną tego stosunku, przy której dla powietrza w temperaturze 20°C prędkość przepływu równa jest prędkości rozchodzenia się dźwięku.



95-115

Hałas, jak każdy dźwięk, rozchodzi się w ośrodkach płynnych (w gazach i cieczach) w postaci fal podłużnych oraz w ciałach stałych w postaci fal podłużnych lub poprzecznych. Hałasy rozchodzące się w ciałach stałych zwane dźwiękami materiałowymi przenoszą się dość łatwo przez sztywne konstrukcje i w efekcie zmniejszają skuteczność działania przegród ograniczających propagację dźwięków powietrznych. Przykładem wpływu propagacji dźwięków materiałowych jest przenoszenie się drgań silników od wind przez konstrukcję budynku, przenoszenie się drgań silnika autobusu przez karoserię. Ponieważ tłumienie dźwięków materiałowych wzrasta proporcjonalnie do częstości, największy ich wpływ występuje w zakresie częstości niskich, a szczególnie infradźwięków.

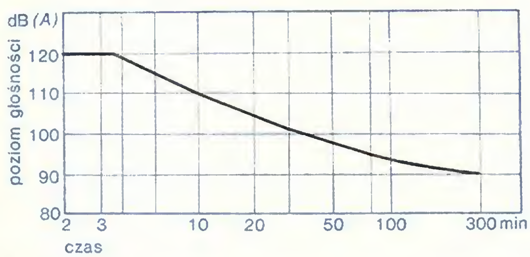
Środowisko akustyczne



95-110

W zależności od warunków w jakich przebywa człowiek narażony na działanie hałasu, hałasy dzieli się na: przemysłowe, komunikacyjne i mieszkaniowe. Poziomy hałasów wytwarzanych przez niektóre urządzenia przemysłowe i przez środki komunikacji oraz poziomy hałasów mieszkaniowych zaznaczone są (w decybelach) przy odpowiednich rysuneczkach na marginesach.

Część społeczeństwa narażona jest w sposób systematyczny w warunkach pracy na działanie hałasów przemysłowych; charakteryzuje je na ogół duże natężenie i długotrwałe działanie. Są one często przyczyną trwałych ubytków słuchu i poważnych schorzeń. Na podstawie badań przyjęto, że największą dopuszczalną wartością poziomu hałasu stacjonarnego działającego na człowieka w sposób długotrwały jest 90 dB(*A*) lub *N* 85. Poziom hałasów trwających krócej może przykraczać te wartości (rys. 9).



Rys. 9. Zależność dopuszczalnych wartości poziomów hałasu od czasu jego trwania



110-130

Obecnie, coraz większy problem stanowią hałasy komunikacyjne, gdyż działają na ludność w sposób masowy. Do hałasów komunikacyjnych należą hałasy wewnętrzne — działające na obsługę pojazdu i pasażerów, i hałasy zewnętrzne, szczególnie uciążliwe w aglomeracjach miejskich.

Bardzo uciążliwe są hałasy odczuwane w mieszkaniach, zakłócające spokojny tryb życia mieszkańców zarówno w dzień jak i w nocy. Poziom hałasów mieszkaniowych nie powinien przekraczać w dzień 40 dB(*A*), natomiast w nocy 30 dB(*A*). W pewnych sytuacjach dopuszcza się poziom o 5 dB wyższy.

Głównymi źródłami hałasów mieszkaniowych są hałasy komunikacyjne, hałasy przemysłowe oraz hałasy wewnątrzbudynkowe (hałasy instalacji wodnej i centralnego ogrzewania, windy, zsypu na śmieci, hałasy pochodzące od sąsiadów i punktów usługowych jak warsztaty, restauracje, sklepy).

Tłumienie hałasu



75-90

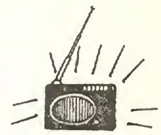
Zmniejszenie hałasu oddziałującego na aparat słuchowy człowieka można uzyskać następującymi sposobami: ograniczając generacje hałasu, ograniczając propagacje hałasu, stosując indywidualne zabezpieczenia aparatu słuchowego.



65-75

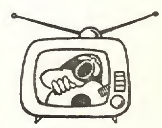
Ograniczenie generacji hałasu uzyskuje się dzięki konstruowaniu urządzeń cichobieżnych, co stanowi najbardziej skuteczny sposób zmniejszenia oddziaływania hałasu na człowieka.

Zmniejszenie hałasu mechanicznego uzyskuje się dzięki odpowiednim konstrukcjom i tolerancji wykonania. Ponadto hałas zmniejsza się ograniczając powierzchnie promieniujące fale akustyczne, stosując materiały o dużym tłumieniu wewnętrznym, w celu ograniczenia przenoszenia się dźwięków materiałowych oraz dąży się do zmniejszenia ujemnego wpływu rezonansów.



80-100

Zmniejszenie hałasu aerodynamicznego można uzyskać przez zmniejszenie prędkości przepływu oraz ograniczenie dużych zmian ciśnienia. Wymagania te prowadzą jednak często do zmniejszenia sprawności urządzeń przepływowych, co ogranicza ich stosowanie. Dlatego też stosuje się również inne środki polegające na doborze kształtu elementów wirujących i kanałów przepływowych, a przede wszystkim na unikaniu ostrych krawędzi i załamań. Przez dobór odpowiedniego kształtu urządzenia lub jego elementów można uzyskać ograniczenie wpływu częstości rezonansowych jak również zmniejszenie skuteczności promieniowania hałasu.



90-110

W celu ograniczenia propagacji hałasu stosuje się przegrody przeciwdźwiękowe. W przemyśle są to zazwyczaj ekrany, obudowy maszyn i urządzeń lub kabiny przeciwdźwiękowe dla personelu. W budownictwie przegrodami izolacyjnymi są ściany, stropy, okna i drzwi. Przegrody izolacyjne muszą zapewniać odpowiednią izolacyjność dla dźwięków powietrznych i dla dźwięków materiałowych.

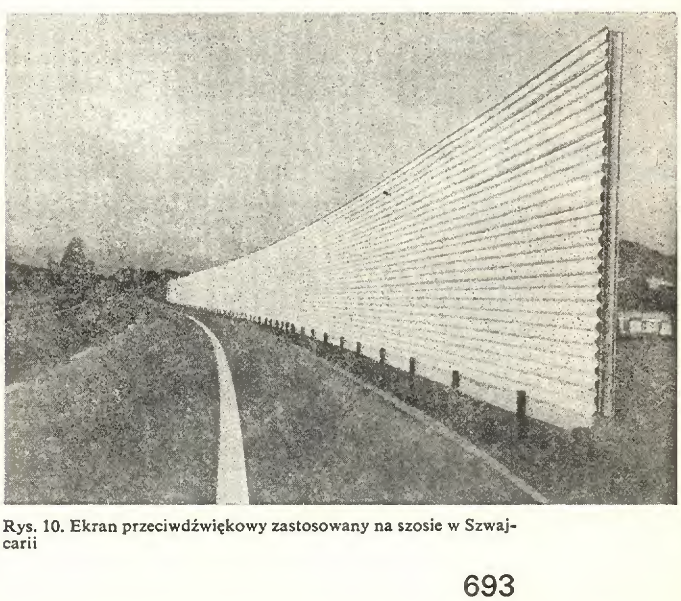
Izolacyjność dla dźwięków powietrznych przegród jednorodnych wyrażona w decybelach jest proporcjonalna do częstości, a także do masy przegrody na jednostkę powierzchni. Stawia to wymagania stosowania przegród ciężkich. Lekkie konstrukcje o dobrych własnościach izolacyjnych można uzyskać stosując przegrody wielowarstwowe.



65-75

W celu polepszenia izolacyjności dla dźwięków materiałowych unika się sztywnych połączeń konstrukcyjnych, które tworzą tzw. mostki akustyczne. W tym celu maszyny i urządzenia instaluje się na podkładach antywibracyjnych lub amortyzatorach. W budownictwie stosowane są specjalne konstrukcje, wśród których bardzo dobre własności wykazują tzw. pływające podłogi.

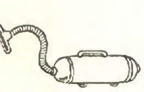
Ograniczenie propagacji hałasu uzyskuje się również stosując ekrany przeciwdźwiękowe. Ekranem nazywamy umieszczoną na drodze fali cienką przeszko-



Rys. 10. Ekran przeciwdźwiękowy zastosowany na szosie w Szwajcarii

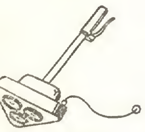
dę o wymiarach większych od długości fali, za którą powstaje cień akustyczny. Efektywność cienia akustycznego jest zmniejszona wskutek ugięcia fal, które docierają za przeszkodę. Wielkością charakteryzującą działanie ekranu jest jego efektywność wyrażona zależnością:

$$IL = 10\lg(E_0/E_u)\ \mathrm{dB},$$



75-85

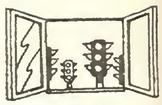
gdzie E_u jest energią akustyczną docierającą do wybranego punktu położonego za ekranem po jego zainstalowaniu, E_0 — energią docierającą do tego samego punktu przed zainstalowaniem ekranu. Ponieważ skuteczność IL ekranu wyrażona w decybelach jest proporcjonalna do częstości, działanie ekranów jest tym skuteczniejsze, im wyższa jest częstość fal akustycznych.



80-90

Ekrany są stosowane w halach przemysłowych w pobliżu źródeł dźwięku lub w pobliżu stanowisk pracy, a także w pobliżu tras komunikacyjnych w celu zmniejszenia poziomu hałasu docierającego do budynków (rys. 13). Efektywność działania jest rzędu 5–15 dB.

W przestrzeni ograniczonej lub w pomieszczeniach zamkniętych fale akustyczne trafiające na powierzchnie ograniczające odbijają się powiększając tym samym energię akustyczną we wnętrzu. Według teorii pogłosowej Sabine'a natężenie dźwięku fal odbitych w pomieszczeniu wyrażone jest zależnością:



75-90

$$I = \frac{4W}{\alpha_{śr}S}(1-\alpha_{śr}),$$

gdzie $\alpha_{śr}$ jest średnim współczynnikiem pochłaniania powierzchni ograniczającej pomieszczenie, V — jego objętością, S — całkowitym polem powierzchni ograniczających. W celu obniżenia wpływu fal odbitych wprowadza się do pomieszczenia materiały dźwiękochłonne, które powiększają wartości $\alpha_{śr}$. Materiały te bywają mocowane na suficie i na ścianach lub zawieszane w postaci tzw. pochłaniaczy przestrzennych. W praktyce powyższą metodą uzyskuje się obniżenie poziomu hałasu o 5–10 dB.

Gdy nie jest możliwe dostateczne obniżenie hałasu przez ograniczenie jego generacji i propagacji, zachodzi konieczność chronienia słuchu przez stosowanie indywidualnych wkładek do uszu lub stosowanie słuchawek przeciwdźwiękowych. Urządzenia te wykazują lepsze własności dla zakresu częstości wysokich, przy których mogą obniżać poziom hałasu działającego na aparat słuchowy człowieka w granicach 20–30 dB. Należy jednak pamiętać, że hałas szczególnie w zakresie niskich częstości, oddziałuje na cały organizm, w związku z tym stosowanie zabezpieczeń indywidualnych jest tylko półśrodkiem.



80-95

BIOFIZYKA I FIZYKA MEDYCZNA

Przedmiot i problemy biofizyki molekularnej · Organizacja procesów życiowych komórki · Błony komórkowe · Białka · Kwasy nukleinowe · Molekularne podstawy skurczu mięśnia · Biomechanika mięśni · Biocybernetyka · Fizyka medyczna · Modelowanie matematyczne procesów biologicznych ·

Przedmiot i problemy biofizyki molekularnej

Kazimierz Wierzchowski

Szybki rozwój nauk przyrodniczych w minionym ćwierćwieczu doprowadził do uformowania się molekularnego nurtu badań w biologii (biologia molekularna), wyłaniając w naturalny sposób pytanie, czy fizyka jako nauka o formach materii, ich ruchu i wzajemnym oddziaływaniu może również wyjaśnić właściwości biologicznej formy istnienia materii i tym samym przyczynić się do wyjaśnienia zjawiska życia. Opis budowy i organizacji procesów życiowych komórki (→ Organizacja procesów życiowych komórki), stworzony przez nauki biologiczne pokazuje, że forma, w jakiej występuje materia w układach biologicznych, radykalnie różni się wieloma swoistymi cechami od form materii nieożywionej.

Z punktu widzenia budowy chemicznej komórka jest wysoce niejednorodnym układem złożonym z kwazikrystalicznych ciał zbudowanych z liniowych makrocząsteczek organicznych oraz z cieczy stanowiącej roztwór wodny związków organicznych i nieorganicznych. W układzie tym zachodzą zorganizowane w przestrzeni i czasie setki reakcji chemicznych, katalizowanych swoiście przez zespoły katalizatorów — białek enzymatycznych. Ich ostatecznym wynikiem jest cykliczna reprodukcja całego układu wg własnego planu genetycznego zakodowanego w strukturze DNA. Procesy te mogą przebiegać tylko w warunkach ciągłej wymiany materii i energii z otoczeniem. Komórka jest więc — z termodynamicznego punktu widzenia — układem otwartym, przy tym samosterowanym i samoreprodukującym się. Pewne odchylenia od planu genetycznego spowodowane spontanicznymi fluktuacjami regulacji aparatu genetycznego lub jego uszkodzeniami przez czynniki zewnętrzne mogą prowadzić do pojawienia się komórek o zmienionych nieco właściwościach (mutanty). Lepsze przystosowanie mutantów do warunków środowiska może z kolei być przyczyną dodatniej selekcji wyróżniających je cech w populacji. Istotną cechą elementarnej formy istnienia materii ożywionej, jaką jest komórka biologiczna, jest więc również zdolność do ewolucji.

komórka jako układ otwarty

Z fizycznego punktu widzenia komórka stanowi zatem układ otwarty złożony z dużej liczby wzajemnie oddziałujących różnych klas elementów, spośród których wiele reprezentowanych jest przez stosunkowo małą ich liczbę. Oddziaływanie elementów zbioru prowadzi do ich uporządkowania, tj. do organizacji w przestrzeni i czasie. Zachowanie się takiego układu niewątpliwie podlega prawom dynamiki. Fizyka nie dysponuje dotychczas metodami pozwalającymi na ilościowy (matematyczny) opis zachowania się takich układów i na przewidywanie właściwości makroskopowych układu oparte na znajomości jego budowy mikroskopowej. Uważa się, że jedynie na gruncie termodynamiki procesów nieodwracalnych, rozpatrującej zachowanie się układów otwartych w stanie odbiegającym od stanu równowagi, można przewidywać warunki powstania organizacji czasowo-przestrzennej reagujących ze sobą chemicznie cząsteczek, a więc uzyskiwać pewien wgląd w powiązania molekularnego i makroskopowego poziomu organizacji układów i mechanizm ewolucji makrocząsteczkowych układów chemicznych w samoreprodukujące się układy prekomórkowe. Wskazują na to teoretyczne badania M. Eigena oraz I. Prigogine'a, wyróżnione nagrodą Nobla (1977 r.)

W ogólnym opisie budowy molekularnej i funkcjonowania komórki stworzonym przez nauki biologiczne, cechy samosterowności, samoreprodukcji i zdolności do ewolucji związane są w dużym stopniu z właściwościami chemicznymi i fizycznymi kilku klas liniowych polimerów stanowiących podstawowe elementy budowy komórek: białek, kwasów nukleinowych, fosfolipidów i polisacharydów. W środowisku wodnym cząsteczki ich mają zdolność do samoorganizacji i uporządkowane przestrzennie formy konformacyjne i supercząsteczkowe układy zw. organellami komórkowymi oraz do specyficznych oddziaływań fizycznych ze związkami nieorganicznymi i organicznymi uczestniczącymi pośrednio lub bezpośrednio w reakcjach biochemicznych.

organelle komórkowe

Wyjaśnienie tych szczególnych właściwości biopolimerów z punktu widzenia ich budowy chemicznej stanowi niewątpliwie klucz do poznania molekularnych mechanizmów procesów komórkowych przebiegających przy ich udziale. Dzięki metodom fizycznym możliwy był szybki rozwój biologii molekularnej: wyjaśnienie na początku lat 50-ych zasady przestrzennej organizacji białek (L. Pauling, nagroda Nobla w 1954 r.) i DNA (F. Crick, i J. Watson, nagroda Nobla 1962 r.). Fizyka w coraz większym stopniu uczestniczy w badaniach z tego zakresu. W ten sposób wyodrębnia się i kształtuje biofizyka molekularna, nauka opisująca doświadczalnie i teoretycznie

biofizyka molekularna

strukturę i właściwości makrocząsteczek biopolimerów w nieuporządkowanym i w różnych uporządkowanych konformacyjnie stanach termodynamicznych, przejścia między stanami, wzajemne oddziaływania i oddziaływania z małocząsteczkowymi składnikami środowiska wodnego komórki oraz organizację w supercząsteczkowe organelle komórkowe. Na tej podstawie poszukuje się wyjaśnienia elementarnych mechanizmów funkcji spełnianych przez biopolimery oraz procesów ich regulacji.

W dalszym ciągu artykułu przedstawimy podstawowe zagadnienia i problemy badawcze tak rozumianej biofizyki molekularnej.

Stan nieuporządkowany makrocząsteczek w roztworach

Stan nieuporządkowany makrocząsteczek w roztworach cechuje zdolność zmian organizacji przestrzennej tworzących je łańcuchów pod wpływem czynników zmieniających ich energię wewnętrzną. Makroskopowe cechy ich budowy: wielkość i kształt, są więc miarą tej organizacji oraz zachodzących w niej zmian pod wpływem czynników zewnętrznych.

Badania kształtu makrocząsteczek

badania roztworów

Informacji o wielkości i kształcie makrocząsteczek w roztworach dostarczają: badania makroskopowych właściwości roztworów, zależnych od parametrów budowy znajdujących się w nich cząsteczek, np. badanie lepkości lub ciśnienia osmotycznego; bezpośrednie pomiary wielkości charakteryzujących procesy transportu cząsteczek: swobodną dyfuzję i sedymentację w polu sił odśrodkowych; badania rozpraszania światła w szerokim zakresie kątów obserwacji oraz rozpraszania promieniowania rentgenowskiego pod małymi kątami. Zależność mierzonych wielkości od parametrów kształtu makrocząsteczek jest na ogół bardzo złożona. W rezultacie opisujące te zależności funkcje można otrzymać w postaci analitycznej tylko dla uproszczonych modeli cząsteczek. Najczęściej stosowane przybliżenia sprowadzają kształt makrocząsteczek do sferoid o różnych promieniach obrotu i stosunkach osi, jak kula, elipsoidy obrotowe o różnym stosunku osi oraz sferoidy o trzech różnych osiach.

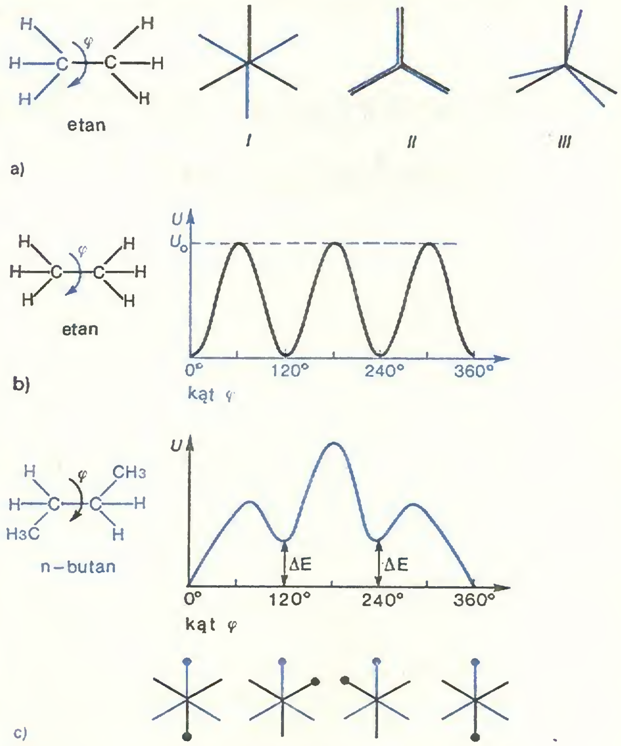
rozmiary kłębka

Kształt makrocząsteczek łańcuchowych w roztworach, w warunkach, w których nie przyjmują one uporządkowanych konformacji, jest zbliżony do bezładnie zwiniętego kłębka nici. Rozmiary jego zależą od budowy chemicznej polimeru, rodzaju rozpuszczalnika i temperatury. W roztworach wodnych polimerów biologicznych, takich jak białka i kwasy nukleinowe, które mają budowę polielektrolitów, rozmiary cząsteczek zależą również od stężenia jonów wodorowych i stężenia elektrolitu (od siły jonowej roztworu). Badania zmian rozmiarów kłębka w zależności od tych czynników wskazują na jego luźną organizację i dużą giętkość łańcucha.

Giętkość łańcuchów polimerów liniowych

Giętkość liniowych cząsteczek polimerów wynika z ich budowy chemicznej, umożliwiającej przyjmowanie przez nie wielu różniących się geometrią form izomerycznych (giętkość termodynamiczna), które mogą przechodzić szybko jedna w drugą pod wpływem drgań cieplnych (giętkość kinetyczna). Cząsteczka chemiczna nie jest układem sztywnym. Zachodzą w niej różnego rodzaju drgania atomów lub grup atomów bez zmiany jej zasadniczej budowy elektronowej, tzn. konfiguracji. Towarzyszą im niewielkie zmiany długości wiązań i wielkości kątów między nimi (kątów torsyjnych). Drgania skrętne wywołujące zmianę względnego położenia grup atomów połączonych wiązaniem pojedynczym, czyli zmianę konformacji cząsteczki, przedstawia się jako zmiany kąta torsyjnego (inaczej kąta dwuściennego, wewnętrznej rotacji, konformacyjnego). Nie zachodzą one swobodnie, ponieważ między nie związanymi bezpośrednio atomami obu grup występują oddziaływania typu van der Waalsa i oddziaływanie elektrostatyczne, których charakter (przyciągające lub odpychające) i energia zależy od efektywnego promienia, odległości i rodzaju atomów. W rezultacie cząsteczka może przyjmować izomeryczne postacie przestrzenne (konformacje) odpowiadające minimom funkcji energii opisującej obrót. Izomerię rotacyjną najprostszych cząsteczek węglowodorów: etanu i *n*-butanu ilustruje rys. 1. Jeżeli energie poszczególnych trwałych izomerów są takie

konfiguracja i konformacja



Rys. 1. Izomery rotacyjne: a) Względne położenie wiązań C—H obu grup CH_3 etanu CH_3—CH_3 (rzut na płaszczyznę prostopadłą do wiązania C—C) w izomerach trans (*I*), cis (*II*) i skręconym (*III*); cząsteczka etanu posiada trójkrotną oś symetrii wzdłuż wiązania C—C, tak więc pełny obrót grupy CH_3 generuje na przemian trzy izomery *trans* o najniższej energii oraz trzy izomery cis o najwyższej energii. b) Funkcja potencjalna opisująca obrót ma postać $U = \frac{1}{2}U_0\,(1 - \cos 3\varphi)$, gdzie U_0 jest to bezwzględna wysokość bariery energetycznej oddzielającej stany trans. Dla etanu wysokość bariery między izomerami cis i trans wynosi 12 kJ/mol. c) Izomeria rotacyjna *n*-butanu; wysokość bariery między izomerami *trans* i skręconym wynosi 3,7 kJ/mol. Grupa CH_3 oznaczona jest na rzucie kropką

same (jak w przypadku etanu), występują one oczywiście z jednakowym prawdopodobieństwem. W przypadku *n*-butanu energie izomerów *trans* i skręconego są różne i ich udział w populacji cząsteczek określony jest różnicą ich energii ΔE. Wysokość bariery energetycznej ΔE oddzielającej sąsiadujące stany konformacyjne determinuje szybkość procesu zmiany konformacji.

izomery geometryczne

W wypadku obrotów wokół pojedynczych wiązań ΔE jest rzędu kilkunastu J/mol. Wówczas zgodnie z przewidywaniami teorii reakcji chemicznych czas trwania procesu zmiany konformacji jest rzędu 10^{-10} s. Podobnie drgania skrętne polimeru liniowego wywołują izomeryzację geometryczną jego łańcucha. Ze względu na dużą liczbę rotacyjnych stopni swobody

łańcuch główny polimeru może przyjąć ogromną liczbę różnych przypadkowych ułożeń w przestrzeni odpowiadających różnym jego izomerom geometrycznym (konformerom), z których żaden nie jest szczególnie uprzywilejowany. Na przykład cząsteczka polietylenu $CH_3(CH_2)_nCH_3$ złożona z $n = 100$ merów ma 3^{100} możliwych izomerów geometrycznych, ponieważ obrót wokół każdego z wiązań C—C prowadzi do powstania trzech izomerów. Oczywiście w wypadku makrocząsteczek białek (→ Białka) czy kwasów nukleinowych (→ Kwasy nukleinowe), o bardziej złożonej budowie niż polietylen, liczba wiązań przypadających na jedno ogniwo łańcucha, wokół których mogą występować obroty, jest znacznie większa. Reszta peptydowa może przyjmować w zależności od budowy chemicznej do 30, a reszta nukleotydowa — aż do 200 różnych izomerycznych konformacji. Zwiększa to radykalnie liczbę teoretycznie możliwych form izomerycznych łańcucha w porównaniu z polimerami syntetycznymi o prostej na ogół budowie chemicznej. Formy te różnią się jednak znacznie energią, tak że tylko niektóre z nich realizują się w normalnych warunkach w roztworach wodnych polipeptydów i polinukleotydów.

Statystyczny opis kształtu makrocząsteczek

Rozróżnienie tak dużej liczby izomerów makrocząsteczek metodami doświadczalnymi nie jest możliwe. Pozwalają one tylko na określenie średniego kształtu cząsteczki. Zatem opis właściwości konformacyjnych nieuporządkowanych polimerów liniowych wymaga zastosowania metod statystycznych. Rozważenie procesu izomeryzacji polietylenu (rys. 2) wyjaśnia dążność cząsteczki do zwijania się w tzw. kłębek statystyczny. W roztworze pod wpływem sił osmozy kłębek ulega rozwinięciu w stosunku do stanu, jaki osiągnąłby w próżni. Miarą stopnia rozwinięcia kłębka, a więc jego rozmiarów, może być średnia odległość h między końcami zwiniętego łańcucha, tj. między pierwszym i ostatnim jego atomem. W statystycznej teorii izomerii geometrycznej polimerów jest ona określona za pomocą funkcji rozkładu $W(h)dh$, równej prawdopodobieństwu znalezienia odległości między końcami łańcucha w przedziale h i $h+dh$; oczywiście $\int_0^\infty W(h)dh = 1$. Średn kwadrat odległości h, równy iloczynowi liczby ogniw łańcucha n i kwadratu długości b ogniwa: $\langle h^2 \rangle = nb^2$, jest proporcjonalny do wyznaczonego doświadczalnie (np. na podstawie pomiaru lepkości lub badania rozproszenia światła) średniego kwadratu promienia obrotu całej cząsteczki. W wypadku „idealnego" polimeru, mającego swobodę obrotów (tzn. obroty mogą występować bez zmian energii) funkcja $W(h)$ ma postać zbliżoną do funkcji Gaussa. Rozkład ogniw jego łańcucha jest początkowo sferycznie symetryczny, jednakże bardzo szybko przyjmuje postać zbliżoną do sferoidy o trzech różnych osiach bezwładności.

kłębek statystyczny

oddziaływanie kooperatywne

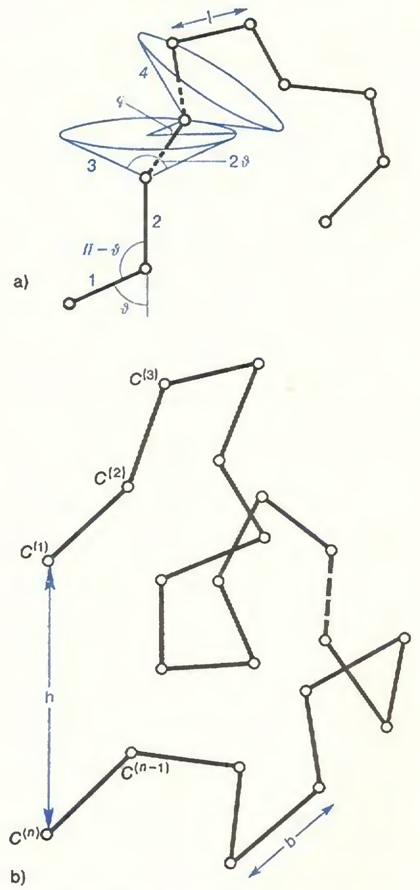
W polimerach rzeczywistych obroty są oczywiście zahamowane przez oddziaływania wewnątrzcząsteczkowe. Oddziaływania te są również przyczyną korelacji kątów rotacyjnych sąsiadujących i blisko położonych ogniw, czyli mają charakter oddziaływań kooperatywnych. W zwiniętej nici polimeru występują także oddziaływania dalekiego zasięgu między ogniwami odległych części łańcucha. W rezultacie część przestrzeni w obrębie kłębka zajęta przez dane ogniwo jest niedostępna dla pozostałych ogniw łańcucha. Jest to tzw. objętość wyłączona, zwiększająca oczywiście rozmiary rzeczywistego kłębka w stosunku do kłębka „idealnego". W polielektrolitach odpychanie elektrostatyczne ładunków jednoimiennych wywołuje dodatkowe zwiększenie objętości kłębka. Uwzględnienie w funkcji $W(h)$ tych czynników jest konieczne do poprawnego opisu jego konformacji.

model budowy polimeru

Zaproponowanie fizycznego modelu zachowania się rzeczywistego polimeru wymaga oczywiście znajomości geometrii budowy jego ogniw (długości wiązań, kątów walencyjnych), a także znajomości energii potencjalnej jako funkcji obrotów ogniw wokół łączących je wiązań oraz energii oddziaływań elektrostatycznych zarówno wewnątrz łańcucha, jak i z otaczającymi go jonami elektrolitu obecnego w roztworze. Parametry geometrii ogniw łańcucha określa się na podstawie wartości odpowiednich parametrów cząsteczek monomerów lub ich analogów, wyznaczanych w kryształach tych związków metodą rentgenografii strukturalnej. Zakres zmienności kątów rotacyjnych w pobliżu ich najbardziej prawdopodobnych wartości, względne energie poszczególnych konformacji oraz bariery energetyczne obrotów określa się badając zależne od konformacji właściwości spektroskopowe takich modelowych cząsteczek w funkcji temperatury.

Można je również oszacować teoretycznie obliczając energie całkowite różnych konformacji obracającego się ogniwa polimeru przy użyciu półempirycznych metod mechaniki kwantowej (→ Chemia kwantowa). Punktem wyjściowym do obliczenia energii oddziaływań elektrostatycznych wewnątrz łańcucha oraz oddziaływań z chmurą otaczających go w roztworze ładunków jest teoria elektrolitów Debye'a–Huckla,

polietylen — przykład izomeryzacji rotacyjnej



Rys. 2. Izomeryzacja rotacyjna polietylenu $CH_3(CH_2)_nCH_3$: a) rotacja wokół wiązania $C^{(2)}$—$C^{(3)}$ kolejnego wiązania $C^{(3)}$—$C^{(4)}$; przy ustalonym położeniu wiązań $C^{(1)}$—$C^{(2)}$ i $C^{(2)}$—$C^{(3)}$ trzecie wiązanie $C^{(3)}$—$C^{(4)}$ może przyjąć dowolne położenie na powierzchni stożka o kącie rozwarcia 2θ ($\pi-\theta$ — kąt walencyjny między wiązaniami C—C bliski tetraedrycznemu, tj. 109°28'). Czwarte z kolei wiązanie $C^{(4)}$—$C^{(5)}$ może przyjąć dowolne położenie na powierzchni stożków opisanych wokół każdego z położeń wiązania trzeciego. Tak więc w miarę oddalania się wiązań wzdłuż łańcucha polietylenu rośnie dowolność orientacji jego wiązań w przestrzeni. Liniowy łańcuch dąży do przyjęcia postaci kłębka statystycznego, bezładnie zwiniętego (zobacz rys. b), którego rozmiary można wyrazić za pomocą odległości h między jego końcami

oparta na liniowym przybliżeniu rozkładu gęstości ładunku. Gęstość powierzchniowa ładunku elektrostatycznego na cząsteczkach silnie zjonizowanych polielektrolitów jest jednak często tak duża, że występują efekty nieliniowe, których uwzględnienie jest bardzo trudne. Teoria roztworów wodnych stężonych elektrolitów i polielektrolitów nie jest jeszcze dostatecznie rozwinięta.

teoria Flory'ego

Dotychczas opracowano kilka statystyczno-termodynamicznych i statystyczno-mechanicznych teorii właściwości nieuporządkowanych makrocząsteczek w roztworach rozcieńczonych (w roztworach rozcieńczonych można pominąć oddziaływania międzycząsteczkowe) — interpretują one poprawnie wyniki badań doświadczalnych.

Bardzo wiele wniosły badania P. J. Flory'ego wyróżnione w 1974 r. nagrodą Nobla. W teorii Flory'ego przyjmuje się, że łańcuch rzeczywistego polimeru jest złożony z segmentów o wirtualnej długości l, obejmujących dwa lub więcej sąsiadujących ogniw, między którymi zachodzą oddziaływania kooperatywne. Względne obroty tak zdefiniowanych N segmentów traktuje się jako swobodne, czyli średni kwadrat odległości h równa się:

$$\langle h^2 \rangle = \alpha^2 N l^2.$$

Współczynnik α w tym wyrażeniu opisuje efekt objętości wyłącznej. W odpowiednio dobranym rozpuszczalniku, sprzyjającym wystąpieniu oddziaływań dalekiego zasięgu, i w odpowiedniej temperaturze wzmożone oddziaływania wewnątrzcząsteczkowe kompensują efekt objętości wyłączonej. W warunkach tych, w tzw. punkcie θ, współczynnik α jest równy. Innymi słowy, drugi współczynnik wirialny ciśnienia osmotycznego jest wówczas równy zeru, roztwór zachowuje się jak roztwór idealny i ciśnienie osmotyczne jest opisywane przez równanie Van't Hoffa. Występuje więc tutaj analogia do punktu Boyle'a dla gazów rzeczywistych. Wyniki teoretycznych obliczeń należy zatem porównywać z wynikami doświadczalnymi otrzymanymi w punkcie θ. Teoretyczne obliczenie współczynnika α jest bardzo trudne. Przybliżone metody jego wyznaczania pozwalają na sprowadzenie wyników otrzymanych dla roztworów nieidealnych do warunków w punkcie θ.

przejście cząsteczek w konformacje kwazikrystaliczne

Podstawowym problemem fizyki roztworów biopolimerów jest wyjaśnienie, w jaki sposób ich cząsteczki przechodzą ze stanu kłębka statystycznego w wysoce uporządkowane kwazikrystaliczne konformacje właściwe ich funkcjonalnemu stanowi *in vivo*. Dotyczy to zarówno układów *in vitro* jak i przebiegu tego procesu w komórce równolegle z syntezą łańcuchów. Cząsteczki w postaci kłębka statystycznego nie mogą krystalizować. Muszą przyjąć przedtem jedną z dostępnych im s^n konformacji o najniższej w danych warunkach energii swobodnej (s — liczba możliwych izomerów geometrycznych pojedynczego ogniwa, n — liczba ogniw). Osiągnięcie przez którąś z cząsteczek w roztworze w sposób przypadkowy konformacji zbliżonej geometrycznie do jej konformacji obserwowanej w krysztale nie oznacza bynajmniej, że konformacja ta wyróżnia się niższą energią swobodną. Dopiero w warunkach sprzyjających odpowiedniej organizacji cząsteczek rozpuszczalnika w warstwie solwatacyjnej (warstwie częściowo uporządkowanego rozpuszczalnika wokół cząsteczki związku rozpuszczonego) i wystąpieniu silniejszych, a także ukierunkowanych oddziaływań wewnątrzcząsteczkowych między sąsiadującymi ogniwami łańcucha, następuje termodynamiczna stabilizacja takiej konformacji.

Problemy te nie mieszczą się już w teorii izomerii geometrycznej polimerów. Próby opisu tych problemów wymagają znajomości konformacji łańcuchów w stanach uporządkowanych makrocząsteczek oraz uwzględnienia energii i kierunkowości oddziaływań międzycząsteczkowych, a także kinetyki izomeryzacji łańcuchów.

Stan uporządkowany makrocząsteczek

Badanie struktury metodami dyfrakcji promieniowania rentgenowskiego

Zdolność cząsteczek białek, kwasów nukleinowych i polisacharydów do krystalizacji z roztworów wodnych lub skupiania się w parakrystaliczne silnie uwodnione włókniste formy stwarza możliwość badania ich przestrzennie uporządkowanych konformacji metodami dyfrakcji promieniowania rentgenowskiego (→ Badanie struktury kryształów). Wyznaczenie na tej drodze współrzędnych atomów tworzących makrocząsteczkę jest jednak znacznie bardziej złożonym zadaniem niż w przypadku małych cząsteczek (→ Osiągnięcia krystalografii białek). Z analizy dyfraktogramów można otrzymać uproszczoną funkcję rozkładu gęstości elektronowej w przestrzeni trójwymiarowej, która posiada maksima w miejscach znajdowania się atomów tworzących cząsteczkę (z wyjątkiem atomów wodoru, których zdolność rozpraszająca jest zbyt niska). Metoda wyznaczania parametrów geometrii makrocząsteczki polega na porównywaniu doświadczalnego rozkładu gęstości elektronowej z rozkładami obliczonymi na podstawie szeregu kolejno uściślanych modeli jej budowy aż do uzyskania dostatecznie dobrej zgodności między obu rozkładami. Wyznaczenie współrzędnych atomów cząsteczki możliwe jest więc tylko wówczas, gdy można zbudować jej model, czyli gdy znana jest jej budowa chemiczna (rodzaj i kolejność występowania poszczególnych ogniw w łańcuchu), kąty konformacyjne i długości wiązań łańcucha głównego i łańcuchów bocznych poszczególnych ogniw oraz struktura podjednostkowa (w białkach złożonych i w wielołańcuchowych formach włóknistych). W przypadku, gdy nie jest znana kolejność występowania (sekwencja) ogniw w łańcuchu polimeru, udaje się tylko określić zasadnicze cechy przestrzennego rozkładu atomów w makrocząsteczce: jego symetrię oraz ułożenie przestrzenne łańcucha głównego.

wyznaczanie geometrii makrocząsteczki

analiza dyfraktogramów DNA

Przykładem takiego postępowania była analiza dyfraktogramów DNA wówczas, gdy nieznana była jeszcze sekwencja występowania zasad we włóknach tego związku. Zaproponowany na podstawie takiej analizy w 1951 r. przez dwóch uczonych F. Cricka i J. Watsona i obowiązujący po dzień dzisiejszy model organizacji przestrzennej cząsteczki DNA w postaci podwójnego heliksu (→ Kwasy nukleinowe) najbardziej odpowiadał symetrii i gęstości doświadczalnego rozkładu elektronów. Model ten znalazł bezpośrednie potwierdzenie dopiero po dwudziestu pięciu latach, gdy wykazano analogiczne upakowanie i konformację cząsteczek dwunukleotydów ApT i GpC w kryształach.

Analiza rentgenograficzna struktury makrocząsteczek napotyka również wiele trudności i ograniczeń wynikających z ich wielkości. Funkcja rozkładu gęstości elektronowej nie odzwierciedla statycznej konformacji pojedynczej cząsteczki, lecz jest superpozycją rozkładów odpowiadających identycznie położonym cząsteczkom w komórce elementarnej kryształu. W dużych cząsteczkach interferencja między rozproszonymi wiązkami promieniowania pochodzącymi od różnych atomów powoduje spadek natężenia refleksów na dyfraktogramach (natężenie refleksów jest proporcjonalne do pierwiastka kwadratowego z liczby atomów, a nie wprost do ich liczby, jak w wypadku małych cząsteczek). Gdy bada się cząsteczki o masie $M > 10^5$ daltonów liczba obserwowanych refleksów jest zbyt mała na to, aby z analizy dyfraktogramów określić współrzędne atomów. (Dalton — stosowana w biochemii nazwa jednostki masy atomowej u; 1u jest równa $^1/_{12}$ masy jądra izotopu węgla ^{12}C). Makrocząsteczki w krysztale mogą różnić się szczegółami ułożenia łańcuchów na ich

(dalton)

powierzchni w wyniku oddziaływania z innymi cząsteczkami tworzącymi kryształ, z cząsteczkami wody krystalicznej oraz jonami elektrolitu obecnego w roztworach. Powoduje to oczywiście rozmycie obrazów dyfrakcyjnych i obniżenie dokładności analizy. Innym, istotnym problemem jest wyznaczenie przesunięć fazowych między padającą i rozproszonymi wiązkami promieniowania; znajomość przesunięć fazowych jest konieczna dla wyznaczenia współrzędnych atomów. Gdy bada się małe cząsteczki zbudowane z nie więcej niż 100 atomów, wówczas określenie przesunięcia fazowego polega na wyborze odpowiedniej funkcji rozkładu gęstości. Natomiast gdy bada się makrocząsteczki zbudowane zazwyczaj z przynajmniej 1000 atomów, wtedy takie bezpośrednie określenie fazy jest niemożliwe. Wprowadzenie do cząsteczki silnie rozpraszającego, ciężkiego atomu daje dopiero odpowiedni punkt odniesienia. W wypadku cząsteczek białek zawierających związane strukturalnie atomy ciężkich pierwiastków, takich jak żelazo (hemoglobina, miogiobina, cytochromy, wiele enzymów utleniająco-redukujących itp.) problem rozwiązany został przez przyrodę. Jednakże w wypadku innych białek wprowadzenie takiego atomu w określone miejsce cząsteczki jest zadaniem niełatwym.

rozmycie obrazów dyfrakcyjnych

rola ciężkiego atomu

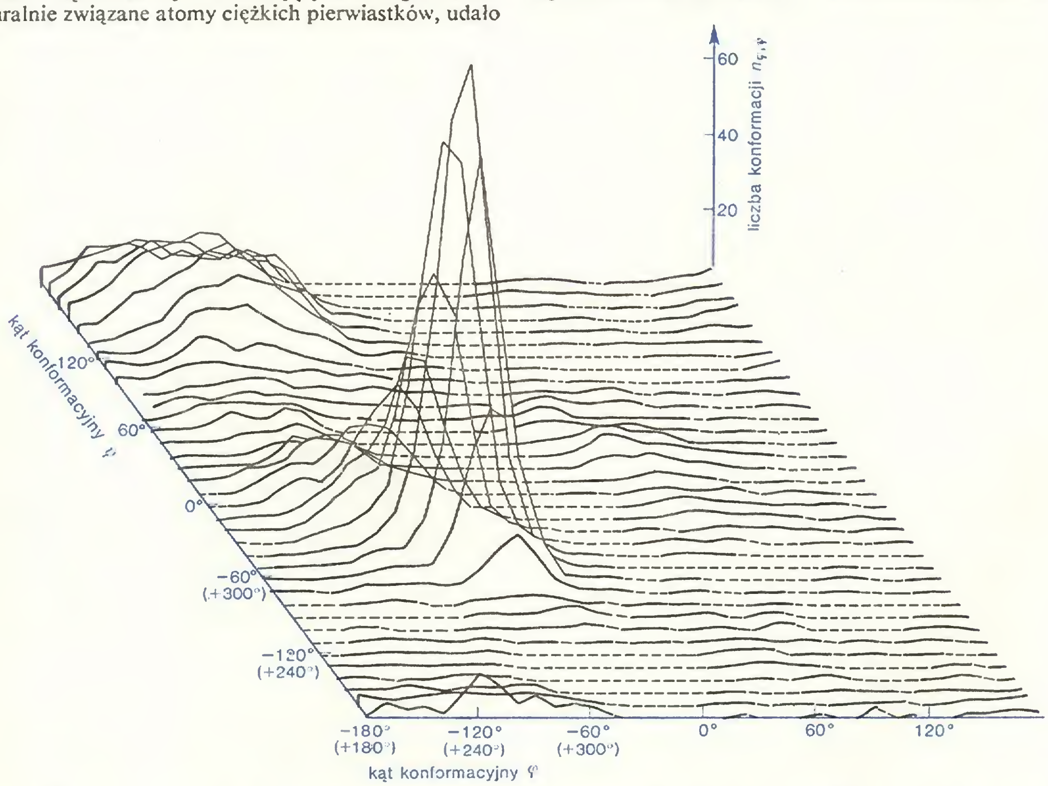
Badania rentgenograficzne struktury makrocząsteczek są bardzo trudne i pracochłonne. Nawet po wprowadzeniu automatyzacji pomiarów i obliczeń za pomocą komputerów rozwiązanie problemu struktury cząsteczki białka średniej wielkości wymaga niejednokrotnie kilkuletniej pracy całego zespołu badaczy. Z tch względów, mimo że rentgenografia stwarza ogromne możliwości poznawcze, nie jest ona metodą powszechnie stosowaną do badania struktury makrocząsteczek, tym bardziej, że wielu związków nie daje się otrzymać w stanie krystalicznym. Dotychczas zbadana została struktura tylko kilkudziesięciu spośród tysięcy białek występujących w przyrodzie. Wśród nich tylko dla kilkunastu, o stosunkowo małej masie cząsteczkowej i zawierających na ogół strukturalnie związane atomy ciężkich pierwiastków, udało się określić współrzędne atomów, dla pozostałych białek określono tylko ułożenie łańcucha peptydowego w przestrzeni. Niemniej jednak wyniki krystalograficznych badań biopolimerów metodami dyfrakcji promieniowania rentgenowskiego stanowią zasadnicze źródło wiedzy o ich uporządkowanych przestrzennie konformacjach. Zaproponowane na ich podstawie modele budowy różnych chemicznie klas liniowych makrocząsteczek są powszechnie stosowane jako podstawa interpretacji zachowania się ich w roztworach i *in vivo*.

Formy przestrzennego uporządkowania biopolimerów

forma heliksu

Najbardziej rozpowszechnioną formą uporządkowania przestrzennego polipeptydów, polinukleotydów i polisacharydów jest ułożenie głównego łańcucha polimeru wzdłuż linii śrubowej, tj. w postaci heliksu. Wobec ograniczonej swobody konformacyjnej tworzących te polimery ogniw zapewnia ona maksymalne upakowanie atomów odpowiadające minimum energii układu. Łańcuchy boczne zależnie od swojej budowy chemicznej, wzajemnych oddziaływań oraz periodyczności występowania wzdłuż łańcucha głównego modyfikują parametry geometryczne heliksu i umożliwiają powstawanie bardziej złożonych form helikalnych, takich jak np. podwójny heliks DNA lub miozyny czy potrójny heliks kolagenu lub hybrydu DNA-RNA. Ogólnie rzecz biorąc łańcuchy przyjmują formy helikalne w wyniku występowania sił bliskiego zasięgu między sąsiadującymi ze sobą ogniwami łańcuchów (wiązania wodorowe, oddziaływania van der Waalsa, oddziaływania elektrostatyczne polarnych lub zjonizowanych grup). Natomiast oddziaływania dalszego zasięgu między odległymi (wzdłuż osi łańcucha) jego elementami powodują powstawanie form globularnych białek i kwasów nukleinowych (tRNA, RNA fagów, informacyjny RNA — o budowie chemicznej

formy globularne



Rys. 3. Rozkład konformacji wiązania peptydowego w 15 globularnych białkach otrzymany na podstawie badań krystalograficznych: $n_{\varphi,\psi}$ liczba konformacji o danej wartości kątów konformacyjnych (torsyjnych) φ i ψ w przedziale $x = 10°$; linie przerywane odpowiadają obszarom przestrzeni konformacyjnej, w których nie stwierdzono występowania wiązań peptydowych; maksima na wykresie odpowiadają konformacjom typu heliksu α (wg F. M. Pohl, *Chemistry of Macromolecules*, London 1974)

łańcucha nie sprzyjającej przyjęciu przez niego regularnej formy helikalnej). W cząsteczkach globularnych łańcuchy mogą przyjmować na pewnych odcinkach helikalne lub inne właściwe danej klasie makrocząsteczek konformacje (np. w przypadku cząsteczek białek poza formami helikalnymi, czyli strukturami α, są to struktury β). Wiązania wodorowe i jonowe dodatkowo spinają lokalnie uporządkowaną globulę i w ten sposób przyjmuje ona kształt właściwy dla łańcucha (łańcuchów) o danej budowie chemicznej. Innymi słowy — indywidualną cząsteczkę cechuje określony zespół kątów konformacyjnych (współrzędnych) odpowiadających jednemu punktowi wielowymiarowej przestrzeni konformacyjnej. W danej klasie polimerów odchylenia od najbardziej prawdopodobnych wartości kątów konformacyjnych walściwych budowie poszczególnych ich ogniw są jednak stosunkowo niewielkie. Ilustruje to rys. 3, na którym przedstawiony jest rozkład kątów konformacyjnych φ i ψ łańcucha polipeptydowego znalezionych w kryształach 15 białek globularnych.

struktury β

Teoretyczne badania konformacji

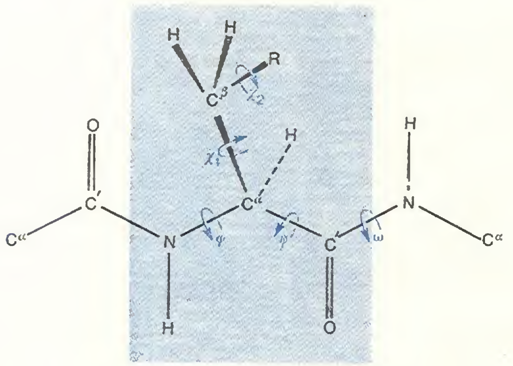
Doświadczalnym badaniom struktury cząsteczek biopolimerów w kryształach towarzyszą szybko rozwijające się w ostatnich latach badania teoretyczne. Ich celem jest przewidywanie konformacji makrocząsteczki o najniższych wartościach energii potencjalnej na podstawie znajomości budowy chemicznej łańcucha polimeru oraz oddziaływań między tworzącymi go atomami. Zadanie sprowadza się do znalezienia funkcji opisującej zależność energii potencjalnej układu od współrzędnych wszystkich atomów cząsteczki w dopuszczalnym zakresie zmian ich wartości.

Podstawą tego typu obliczeń jest model budowy chemicznej łańcucha przyjęty w teorii izomerii geometrycznej. Dopuszczalne stereochemicznie konformacje cząsteczki definiuje się za pomocą kątów torsyjnych opisujących obroty ogniw wokół wiązań łańcucha głównego polimeru i grup bocznych w stosunku do osi łańcucha; podanie ich wartości dla wszystkich kolejnych ogniw w łańcuchu daje opis konformacji makrocząsteczki. Jednakże wykonanie obliczeń dla tak dużych układów jest jeszcze technicznie niemożliwe, zwłaszcza że należałoby w nich uwzględniać również oddziaływania między odległymi segmentami łańcucha oraz między poszczególnymi segmentami i cząsteczkami roztworu. Korzystając jednak z faktu, że oddziaływania określające postać funkcji i energii potencjalnej obrotu wokół danego wiązania w łańcuchu głównym zależą w pierwszym przybliżeniu tylko od wzajemnego położenia sąsiadujących ze sobą ogniw, można obliczyć energie potencjalne segmentów łańcucha, w którym występuje korelacja kątów obrotu. W ten sposób przewiduje się najbardziej prawdopodobne konformacje podstawowych ogniw łańcuchów. Najwięcej takich obliczeń wykonano dotychczas dla białek. Płaskość wiązania peptydowego pozwala opisać izomerię geometryczną łańcucha za pomocą kątów torsyjnych φ i ψ (rys. 4). Tak więc segmentami łańcucha, których rotacje można traktować jako praktycznie swobodne, są dwupetydy zdefiniowane przez parę kątów φ_i i ψ_i oraz rodzaj łańcucha bocznego R. Energia konformacyjna takich jednostek zależy w zasadzie tylko od ich chemicznej budowy, tzn. od długości wiązań kowalencyjnych i kątów między nimi. Parametry geometryczne tej budowy muszą być wyznaczone z niezależnych badań krystalograficznych dwupeptydów lub odpowiednich modelowych związków. Graficzne przedstawienie energii potencjalnej układu w funkcji kątów φ i ψ przy optymalnej wartości kąta rotacyjnego χ_2 grupy bocznej R stanowi mapę konformacyjną dwupeptydu, z której można odczytać najbardziej prawdopodobne jego konformacje, odpowiadające najniższym wartościom energii. Energię układu można obliczyć w sposób przybliżony metodami klasycznymi rozdzielając funkcję energii potencjalnej na składowe pochodzące od różnego typu oddziaływań międzyatomowych w cząsteczce (wiązania wodorowe, van der Waalsa, elektrostatyczne, dipol–dipol indukowany), dla których zależność energii oddziaływania od parametrów budowy cząsteczki można podać w postaci analitycznej.

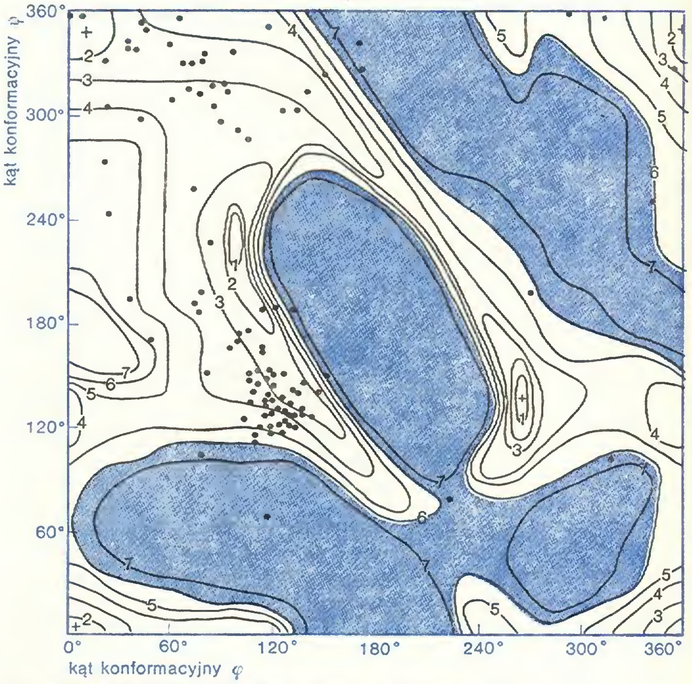
opis konformacji makrocząsteczki

mapa konformacyjna

Jednakże od czasu opracowania w połowie lat 60-ych metod mechaniki kwantowej przystosowanych do badań konformacyjnych „dużych układów" (metody CNDO-2, PCILO, uwzględniające wszystkie elektrony walencyjne π i σ) obliczenia te są wykonywane głównie przy ich użyciu (→ Chemia kwantowa). Mapa konformacyjna grupy alanylowej, obliczona metodą PCILO, przedstawiona na rys. 5,



Rys. 4. Opis konformacji łańcucha polipeptydowego. Grupa dwupeptydowa, w tym przypadku alanylowa, wewnątrz której występuje korelacja obrotów wokół wiązań N—C (kąt rotacyjny φ) i C^{α}—C′ (kąt rotacyjny ψ) podana jest na niebieskim tle. Obrót wokół wiązań N—C′ (kąt rotacyjny $\omega = 180°$) nie zachodzi w normalnych warunkach ze względu na dużą wysokość bariery energetycznej dla obrotu ($\Delta E \approx 80$ kJ/mol)



Rys. 5. Mapa konformacyjna reszty alanylowej białek, wyniki obliczeń otrzymano metodą PCILO; przy liniach izoenergetycznych podane są wartości energii (od 1 do 7) odniesionej do ogólnego minimum energii cząsteczki, wyrażone w kcal/mol (1 kcal ≈ ≈ 4kJ, 7 kcal ≈ 29 kJ); znak + oznacza lokalne minimum; obszar konformacji o energii do 24 kJ/mol powyżej ogólnego minimum jest zaznaczony na niebiesko. Na mapę naniesiono punkty odpowiadające konformacjom grupy alanylowej stwierdzonym w kilku globularnych białkach na podstawie badań krystalograficznych (wg A. Pullman)

pokazuje dopuszczalne jej konformacje w przedziale energii 0–25 kJ/mol. Naniesione zostały na nią punkty odpowiadające znalezionym doświadczalnie parom kątów φ i ψ dla grupy alanylowej w krystalicznych białkach globularnych. Leżą one wprawdzie w obrębie przewidzianego teoretycznie obszaru konformacyjnej stabilności grupy alanylowej, jednakże ich statystyczny rozkład w cząsteczkach białek nie odpowiada tym przewidywaniom. Jest to wynik oddziaływań dalszego zasięgu, których w dotychczas wykonywanych obliczeniach nie brano świadomie pod uwagę. Podejmowane są obecnie prace mające na celu ich uwzględnienie, a także — i to przede wszystkim — uwzględnienie wpływu uwodnienia i obecności jonów w warstwie solwatacyjnej łańcuchów.

opis konformacji DNA

Podobne teoretyczne badania konformacji „niezależnych" segmentów łańcuchów prowadzi się i dla innych klas biopolimerów — kwasów nukleinowych, polisacharydów, fosfolipidów. W przypadku kwasów nukleinowych nie wyszły one poza obliczenia map konformacyjnych poszczególnych jej fragmentów ze względu na znacznie większą liczbę stopni swobody ogniw nukleotydowych; głównym przedmiotem zainteresowania są dopuszczalne ułożenia przestrzenne łańcuchów bocznych względem łańcucha głównego, decydujące o geometrii podwójnych heliksów oraz konformacje pięcioczłonowych pierścieni rybozy i dezoksyrybozy.

Badania konformacji makrocząsteczek w roztworach

Ustalona na podstawie danych rentgenograficznych uśredniona, statyczna konformacja makrocząsteczki w krysztale odpowiada określonemu jej stanowi termodynamicznemu. Czy jednak stan ten jest zachowany przez cząsteczki biopolimerów w roztworach wodnych w warunkach odpowiadających trwałości termodynamicznej ich uporządkowanych konformacji? Ogromna liczba geometrycznie różnych ułożeń, które może przyjąć łańcuch polimeru, nasuwa z kolei pytanie, czy utworzona w krysztale lub roztworze konformacja makrocząsteczki odpowiada globalnemu minimum energii swobodnej układu, czy też ma ona szereg stanów konformacyjnych odpowiadających lokalnym minimom energii swobodnej i oddzielonych od siebie wysokimi barierami energetycznymi, tak że w określonych warunkach realizuje się niekoniecznie stan najbardziej prawdopodobny. Odpowiedzi na te pytania biofizyka poszukuje na drodze badania konformacji makrocząsteczek w roztworach, sił odpowiedzialnych za ich uporządkowanie, trwałości termodynamicznej (denaturacja) i samoorganizacji (renaturacja) cząsteczek ze stanu kłębka statystycznego. Znajdują w nich zastosowanie praktycznie wszystkie metody badania kształtu i konformacji cząsteczek, jakimi dysponuje współczesna chemia i fizyka.

badanie kształtu i giętkości

Przyjęcie przez łańcuch polimeru zdefiniowanej przestrzennie konformacji znajduje swoje odbicie w zewnętrznym kształcie cząsteczki oraz w jej większej sztywności w porównaniu z formą nieuporządkowaną. Toteż badanie kształtów i giętkości makrocząsteczek metodami stosowanymi w badaniach ich form nieuporządkowanych umożliwia w sposób pośredni ocenę charakteru i stopnia uporządkowania ich łańcuchów. Informacji o konformacji łańcuchów dostarczają badania spektroskopowe właściwości uwarunkowanych symetrią i względnym położeniem grup chromoforowych (grup o charakterystycznych cechach widmowych). Na przykład formy helikalne polipeptydów i polinukleotydów cechuje ze względu na właściwą im symetrię budowy duża aktywność optyczna różniąca się znakiem i wielkością od aktywności optycznej ich elementów budowy — aminokwasów i nukleotydów (różne widma dyspersji światła spolaryzowanego liniowo lub widma absorpcji światła spolaryzowanego kołowo). Inne jeszcze właściwości optyczne mają warstwowe struktury β polipeptydów. W helikalnych, dwułańcuchowych konformacjach polinukleotydów warstwowo ułożone pary zasad absorbują znacznie słabiej promieniowanie nadfioletowe niż w bezładnie zwiniętych łańcuchach (tzw. efekt hypochromowy). Podobnie wiele informacji dotyczących zarówno uporządkowania większych obszarów makrocząsteczek jak i ich fragmentów uzyskuje się badając właściwości spektroskopowe określonych atomów lub grup atomów (chromoforów) i wykorzystując zależność ich stanów spinowych (spekstroskopia jądrowego rezonansu magnetycznego i elektronowego rezonansu paramagnetycznego, zjawisko Mössbauera), oscylacyjnych (widma absorpcyjne w podczerwieni, widma Ramana) i elektronowych (absorpcyjna i emisyjna spektroskopia w nadfioletowej i widzialnej części widma) od natury, odległości i orientacji sąsiadujących z nimi innych atomów i ich grup. Metody te pozwalają na obserwację zmian zachodzących w otoczeniu chromoforów w procesach organizacji i dezorganizacji cząsteczek oraz w wyniku oddziaływań ich z innymi cząsteczkami w roztworze. Zakres zastosowania metod spektroskopowych stale wzrasta, m.in. dzięki rozwojowi metod badawczych i ich podstaw teoretycznych oraz opanowaniu techniki wprowadzania w określone miejsca makrocząsteczek małych cząsteczek (wiązanych kowalencyjnie lub w postaci specyficznych kompleksów) o szczególnie czułych na zmiany w otoczeniu i o wyróżniających się właściwościach spektroskopowych. Spełniają one funkcję sond stanu konformacyjnego makrocząsteczki. Żadna jednak z metod badania konformacji makrocząsteczek w roztworach nie pozwala na wyznaczenie parametrów geometrii łańcucha.

badania aktywności optycznej

badanie właściwości spektroskopowych chromoforów

badanie fragmentów

Złożoność budowy chemicznej i wielkość cząsteczek biopolimerów narzuca oczywistą strategię badań właściwości ich stanu uporządkowanego w roztworach, polegającą na redukcji wielkości badanych układów. Bada się zatem określone większe fragmenty natywnych cząsteczek, proste elementy ich budowy jak monomery lub dimery oraz cząsteczki tworzące łańcuchy boczne. Przedmiotem badań są również syntetyczne modele naturalnych łańcuchów o prostszej chemicznie i odpowiednio zaprogramowanej budowie i stopniu polimeryzacji. Badania właściwości konformacyjnych fragmentów i modelowych cząsteczek pozwalają na ocenę roli łańcuchów bocznych monomerów o różnej chemicznej budowie (grupy alifatyczne i aromatyczne aminokwasów, zasady purynowe i pirymidynowe nukleotydów) w formowaniu się określonych konformacji polimerów. Dostarczają również danych dotyczących fizycznego charakteru i wielkości energii oddziaływań określających geometrię i trwałość tych konformacji.

Siły odpowiedzialne za organizację łańcuchów makrocząsteczek

Poznanie sił odpowiedzialnych za organizację giętkich łańcuchów polimerów w uporządkowane konformacje ma zasadnicze znaczenie w zrozumieniu właściwości ich stanu uporządkowanego.

Ogólnie rzecz biorąc oddziaływania van der Waalsa i elektrostatyczne między atomami fragmentów budowy makrocząsteczek odpowiedzialne są zarówno za konformacje łańcuchów głównych jak i ich przestrzenne upakowanie w uporządkowanych formach makrocząsteczek. Charakterystyczne dla danej makrocząsteczki rozmieszczenie różnych łańcuchów bocznych wzdłuż łańcucha głównego determinuje wystąpienie nieco silniejszych, kooperatywnych oddziaływań przyciągających między określonymi fragmentami jej budowy i przyjęcie przez nią najbardziej prawdopodobnej energetycznie i termodynamicznie formy uporządkowania przestrzennego. Oddziaływania te

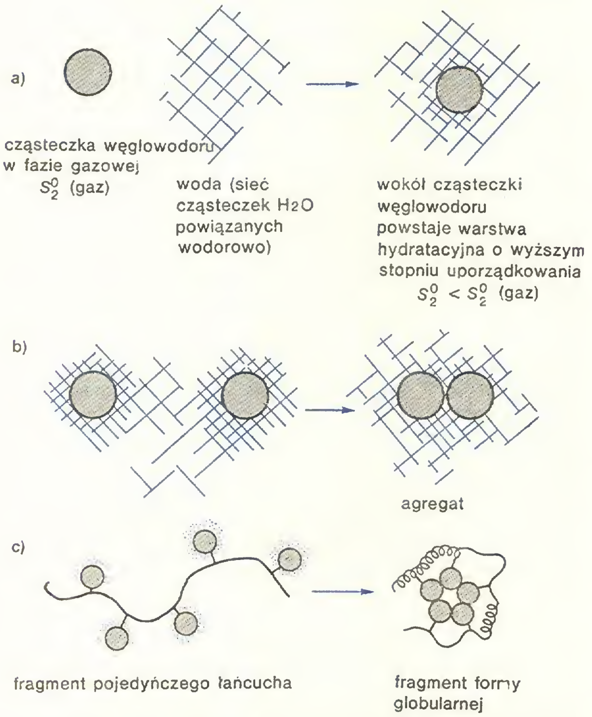
wiązania wodorowe

mają w tym sensie charakter kierunkowy. Wyróżnia się wśród nich przede wszystkim oddziaływania prowadzące do utworzenia tzw. wiązań wodorowych (—A···H—B—) między akceptorowymi (A) i donorowymi (B—H) ugrupowaniami atomów, zawierającymi atomy tlenu, azotu i siarki, z których jeden związany jest kowalencyjnie z atomem wodoru.

W przyjętych modelach konformacji łańcuchów polipeptydowych i polinukleotydowych wiązaniom wodorowym przypisuje się rolę oddziaływań służących wzajemnemu rozpoznawaniu się i przestrzennej orientacji segmentów cząsteczki sąsiadujących ze sobą w tym samym łańcuchu (np. heliks α polipeptydów), w przeciwległym łańcuchu (np. podwójny heliks DNA) lub oddalonych od siebie (formy globularne). Szczegółowa charakterystyka stereochemicznych i termodynamicznych warunków ich powstawania w układach modelowych oraz określenie ich energii w zależności od typu struktury, w której występują, jest przedmiotem licznych prac.

oddziaływanie z cząsteczkami wody

Specyficzne właściwości konformacyjne biopolimerów ujawniają się wyłącznie w roztworach wodnych, toteż poznanie hydratacji makrocząsteczek (wiązania się z cząsteczkami wody) i roli wody w formowaniu się i stabilizacji uporządkowanych struktur zajmuje szczególnie istotne miejsce we współczesnych badaniach. Utworzeniu specyficznych, wewnątrzcząsteczkowych wiązań wodorowych towarzyszy usunięcie cząsteczek wody związanej podobnymi wiązaniami



Rys. 6. Molekularny i termodynamiczny model oddziaływań hydrofobowych między grupami alifatycznymi w roztworze wodnym: a) Schemat rozpuszczania cząsteczek węglowodoru alifatycznego w wodzie; ograniczona rozpuszczalność węglowodoru alifatycznego w H_2O, tzn. $\Delta G^0_{rozp} = \Delta H^0_{rozp} - T(\Delta S^0_{konf} + \Delta S^0_{hydrofob} \geqslant 0$ (gdzie G — energia swobodna Gibbsa, H — entalpia) wynika z obniżenia się entropii układu $\Delta S^0_{hydrofob} = S_2^0 - S_2^0$ (gaz) < 0 na skutek wzrostu uporządkowania cząsteczek wody wokół cząsteczek węglowodoru ($\Delta H^0_{rozp} < 0$ jest to entalpia rozpuszczania, $\Delta S^0_{konf} \approx 0$ — zmiana entropii konfiguracyjnej węglowodoru, S_2^0 i S_2^0 (gaz) — cząstkowe entropie związku w roztworze oraz w fazie gazowej). b) Schemat agregacji cząsteczek węglowodoru oraz c) grup alkilowych łańcuchów bocznych liniowego polimeru. Agregacja jest procesem termodynamicznie korzystnym, tzn. $\Delta G^0_{agr} = \Delta H^0_{agr} - T\Delta S^0_{agr} < 0$, ponieważ towarzyszy mu duży przyrost entropii układu $\Delta S^0_{agr} = \Delta S^0_{konf} + \Delta(\Delta S^0_{hydrofob}) > 0$ spowodowany usunięciem cząsteczek wody z warstwy solwatacyjnej otaczającej agregujące grupy i cechującej się wyższym stopniem uporządkowania molekularnego niż sam rozpuszczalnik

wodorowymi z grupami akceptorowymi (A) i donorowymi (B—H) polimerów. Bilans energetyczny procesów zerwania istniejących wiązań wodorowych A···H—O—H i B—H···OH_2 i utworzenia nowych wiązań A···H—B wnosi swój wkład do zmiany energii swobodnej układu. Procesy te są jednak ciągle niewystarczająco zbadane termochemicznie i zinterpretowane termodynamicznie. Udział wody w powstawaniu natywnych konformacji biopolimerów przejawia się jednak przede wszystkim w skłonności niepolarnych ich fragmentów budowy do skupiania się z wytworzeniem bezpośrednich kontaktów i usunięciem cząsteczek wody z ich powierzchni (oddziaływania hydrofobowe). Procesowi temu towarzyszy znaczny wzrost entropii układu, który decyduje o obniżeniu jego energii swobodnej, stanowiąc zatem siłę napędową reorganizacji struktury molekularnej (rys. 6).

oddziaływanie hydrofobowe

W cząsteczkach białek globularnych oddziaływania hydrofobowe powodują skupianie się we wnętrzu globuli segmentów łańcuchów polipeptydowych, zbudowanych z aminokwasów mających grupy węglowodorowe jako łańcuchy boczne. W cząsteczkach kwasów nukleinowych oddziaływania hydrofobowe są jedną z przyczyn warstwowej organizacji par zasad we wnętrzu podwójnego heliksu. Odpowiedzialne są one również za dwuwarstwową organizację fosfolipidowych błon komórkowych (→ Błony komórkowe). Zaproponowane dotychczas molekularne modele oddziaływań hydrofobowych, a także oparta na nich termodynamiczno-statystyczna interpretacja właściwości roztworów zawierających cząsteczki organiczne o budowie zbliżonej do elementów budowy biopolimerów, tylko w sposób przybliżony wyjaśniają doświadczalnie obserwowane ich właściwości. Przyczyną takiego stanu rzeczy jest niezadowalający postęp badań termochemicznych i badań struktury roztworów wodnych. Brak jest nawet poprawnej molekularnej teorii stanu ciekłego wody.

wpływ stabilizujący jonów elektrolitu

Trwałość uporządkowanych konformacji biopolimerów, których łańcuchy mają właściwości polielektrolitów, zależy oczywiście od stanu oddziaływań elektrostatycznych w układzie makrocząsteczka–jony elektrolitów obecnych w roztworze. Elektrostatyczne odpychanie się ujemnie naładowanych grup fosfoestrowych sąsiadujących ze sobą w cząsteczkach kwasów nukleinowych obniża stabilność konformacji podwójnego heliksu. Obecne w roztworze kationy silnych elektrolitów, takich jak NaCl, KCl, $MgCl_2$, skupiają się wokół ujemnie naładowanych centrów i ekranują je. Ponieważ gęstość powierzchniowa ujemnego ładunku podwójnego heliksu DNA jest większa niż pojedynczej nici polinukleotydowej, kationy są silniej wiązane przez formy uporządkowane kwasów nukleinowych, wywierając wpływ stabilizujący je. Analogicznie jest stabilizowana obecnością kationów warstwowa budowa fosfolipidowych błon biologicznych. *In vivo* oprócz kationów metali Na^+, K^+, Ca^{+2}, Mg^{+2} itd. stabilizujący wpływ na DNA wywierają również dodatnio naładowane grupy poliamin i białek zasadowych obecnych w chromatynie.

Ponieważ energia oddziaływań elektrostatycznych jest odwrotnie proporcjonalna do przenikalności dielektrycznej ε ośrodka rozdzielającego ładunki elektryczne, a wartość ε pozbawionego wody apolarnego wnętrza globuli białkowej jest znacznie mniejsza niż wody, zatem energia elektrostatyczna we wnętrzu globuli jest przynajmniej o jeden rząd wielkości większa niż na jej powierzchni. Toteż przyciągające oddziaływania elektrostatyczne między dodatnio i ujemnie naładowanymi ogniwami — aminokwasami (wiązania jonowe) odgrywają istotną rolę w fałdowaniu łańcuchów polipeptydowych i stabilizacji globularnych makrocząsteczek.

W wielu białkach spełniających funkcje transportowe (np. hemoglobina, cytochromy) lub katalityczne (np. enzymy uczestniczące w reakcjach utleniania i redukcji) jony metali ciężkich, takich jak Fe, Mo, Co, Mn, Zn, specyficznie i trwale związane w warun-

kach fizjologicznych, odgrywają rolę centrów wiążących i katalitycznych. Zmiany stanu utlenienia jonu metalu wywołują w nich zmiany konformacyjne, przystosowujące cząsteczki do udziału w kolejnym etapie cyklu transportowego lub katalitycznego.

Niespecyficzne i specyficzne oddziaływania jonów metali z cząsteczkami biopolimerów kontrolujące ich konformacje, a tym samym i funkcje w komórce, są od wielu lat przedmiotem systematycznych badań doświadczalnych i teoretycznych, obecnie dalekich jeszcze od zakończenia.

Dynamika uporządkowanych konformacji biopolimerów

Stan termodynamiczny i trwałość uporządkowanych konformacji makrocząsteczek określone są sumą bardzo wielu słabych oddziaływań wewnątrzcząsteczkowych i międzycząsteczkowych ze składnikami roztworu wodnego. W rezultacie w cząsteczkach występują termiczne fluktuacje konformacyjne wywołujące odchylenia kątów konformacyjnych od wartości najbardziej prawdopodobnych w „statycznej” konformacji w stanie krystalicznym. Fluktuacje te mają najprawdopodobniej charakter drgań własnych układu, określonych budową chemiczną łańcucha i jego przestrzenną organizacją. Nadają one makrocząsteczkom właściwości układów stochastycznych. O występowaniu fluktuacji świadczą przede wszystkim obserwacje wskazujące na to, że pewne grupy funkcjonalne białek i kwasów nukleinowych, ukryte — wg danych krystalograficznych — w ich wnętrzu, są dostępne przez krótkie, 10^{-3}–10^{-8} s, okresy czasu dla różnych reakcji chemicznych (np. wymiana atomów wodoru i deuteru), bądź dla oddziaływań fizycznych z cząsteczkami obecnymi w roztworze (np. gaszenie fluorescencji określonych grup przez tlen). Poznanie „widma fluktuacji konformacyjnych” makrocząsteczek, tzn. zależności częstości fluktuacji od energii aktywacji, oraz przypisanie poszczególnym częstościom określonych zmian w budowie makrocząsteczek będzie miało istotne znaczenie dla wyjaśnienia mechanizmów specyficznego rozpoznawania przez makrocząsteczki innych makrocząsteczek, a także małych cząsteczek, w trakcie spełniania właściwych im funkcji w komórce. Dotychczas jednak jeszcze bardzo niewiele wiadomo o tych fluktuacjach z doświadczalnego punktu widzenia. Rozwój technik szybkich pomiarów kinetycznych w przedziałach czasu mikro- i nanosekundowych połączonych ze spektroskopowymi badaniami zmian konformacyjnych rokuje nadzieje na postęp w doświadczalnych badaniach dynamiki budowy makrocząsteczek. Podejmowane są również próby teoretycznego jej opisu.

fluktuacje konformacyjne

fluktuacje a mechanizm rozpoznawania

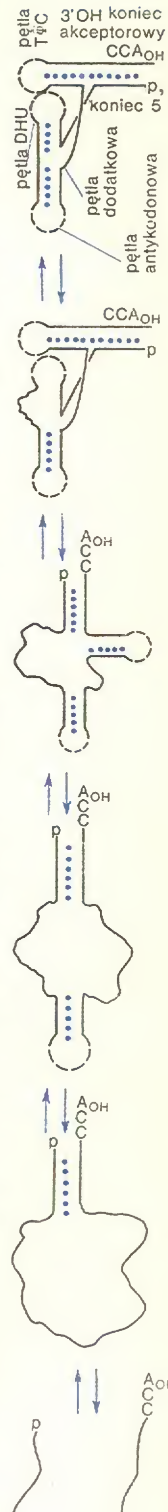
Zmiany stanu uporządkowania makrocząsteczek

Doświadczalne badania przebiegu procesów dezorganizacji przestrzennej budowy (denaturacji), samoorganizacji i przejścia cząsteczek białek i kwasów nukleinowych oraz ich syntetycznych modeli z jednej uporządkowanej konformacji w drugą dowodzą, że zmiana stanu konformacyjnego makrocząsteczki jest procesem odwracalnym (jeżeli nie zachodzą w niej zmiany chemiczne), kooperatywnym i ma charakter przejścia fazowego II rodzaju. Oznacza to, że zmiany strukturalne w makrocząsteczce zachodzą w sposób ciągły, towarzyszy im ciągła zmiana entalpii i skokowa zmiana pojemności cieplnej układu. Poznanie termodynamiki, molekularnych mechanizmów i kinetyki tych procesów dla różnych klas biopolimerów i indywidualnych makrocząsteczek w poszczególnych klasach oraz ich statystyczno-termodynamiczna interpretacja ma na celu otrzymanie odpowiedzi na dwa zasadnicze pytania: w jaki sposób można przewidywać, opierając się na znajomości budowy chemicznej cząsteczek, ich uporządkowane konformacje? Jakie są elementarne fizyczne mechanizmy ich działania w komórce i jak one wynikają z ich właściwości konformacyjnych? Dzisiaj biofizyka nie jest jeszcze w stanie udzielić zadowalającej odpowiedzi na te pytania. Systematyczne badania w tej dziedzinie zostały podjęte stosunkowo niedawno. Jednakże wyniki ich pozwalają już na sformułowanie szeregu ogólniejszych wniosków i hipotez.

mechanizm samoorganizacji

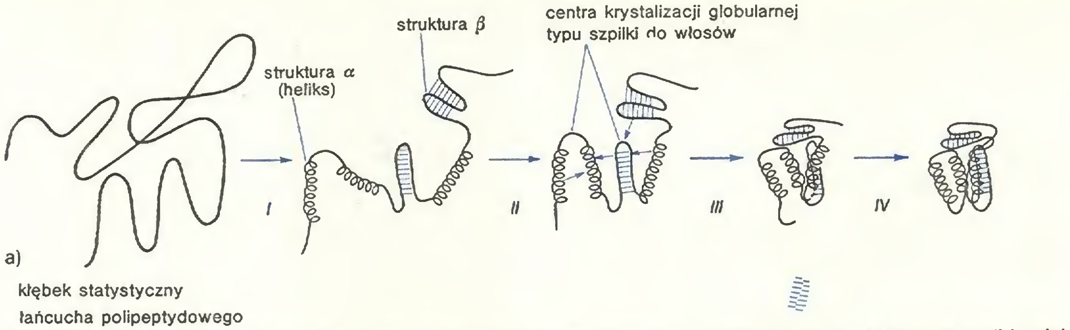
Szczególnie wiele informacji dotyczących mechanizmu samoorganizacji makrocząsteczek dostarczają obserwacje spektroskopowe procesu relaksacji uporządkowanych konformacji, wytrąconych z równowagi termodynamicznej przez skokową zmianę temperatury lub ciśnienia w roztworze. Okazuje się, że czas potrzebny do przyjęcia przez bezładnie zwinięty kłębek określonej konformacji jest bardzo krótki, zależy jednak oczywiście od wielkości, charakteru i złożoności budowy chemicznej cząsteczki. Na przykład jest on rzędu 10^{-8} s dla formowania się helikalnych konformacji najprostszych polipeptydów zbudowanych tylko z jednego rodzaju aminokwasów. Natomiast samoorganizacja łańcuchów naturalnych polipeptydów w formy globularne o złożonej przestrzennej organizacji wymaga od 0,01 do kilku sekund. Tak krótki czas, w którym giętki łańcuch polimeru przyjmuje określoną konformację spośród ogromnej liczby izomerycznych jej odmian, realizujących się w stanie kłębka statystycznego, świadczy o tym, że proces organizacji musi przebiegać przez ograniczoną liczbę stanów pośrednich. Można łatwo obliczyć, że gdyby proces fałdowania łańcucha polipeptydowego zbudowanego ze 150 aminokwasów miał przebiegać przez „sprawdzanie” w przypadkowy sposób wszystkich dostępnych mu konformacji trwałoby to wówczas 10^{26} lat! Na pojawienie się stanów przejściowych w procesach formowania się globularnych cząsteczek białek i tRNA wskazują obserwacje spektroskopowe, analiza profili przejść konformacyjnych oraz pomiary czasów relaksacji. W białkach globularnych obserwuje się występowanie dwu i więcej czasów relaksacji w obrębie przejścia kłębek $\rightleftharpoons$ globula. Interpretacja ich na podstawie analogicznych doświadczeń z polipeptydami o konformacjach heliksu α i warstwowej struktury β prowadzi do wniosku, że pierwsze, szybkie stadium fałdowania łańcucha polipeptydowego białka polega na utworzeniu helikalnych fragmentów. Dużo powolniejsze stadium, ograniczające szybkość powstawania globuli, wiąże się z formowaniem struktury β (rys. 8). Podobnie w przypadku tRNA powstanie globularnej formy cząsteczki rozpoczyna się od utworzenia szeregu odcinków podwójnego heliksu między komplementarnymi odcinkami tego samego łańcucha (rys. 7). Najpowolniejszym stadium procesu powstawania form helikalnych i warstwowych jest utworzenie zarodka konformacji właściwej dla danej cząsteczki (powstanie wiązań wodorowych między sąsiadującymi resztami peptydowymi lub między komplementarnymi parami zasad w polinukleotydach), od którego poczynając następuje szybki, kooperatywny proces formowania się tej konformacji w całym łańcuchu.

Przedstawiony wyżej model procesu samoorganizacji białek globularnych stanowi podstawę prób teoretycznego wyjaśnienia konformacji białek, których struktura w kryształach została ustalona metodami dyfrakcji promieniowania rentgenowskiego. Termodynamiczno-statystyczna interpretacja przejść konformacyjnych (zob. niżej) w odpowiednio dobranych modelowych polipeptydach pozwala na określenie parametrów termodynamicznych charakteryzujących pra-



Rys. 7. Samoorganizacja cząsteczki tRNA: powstawanie formy globularnej poprzedzone jest formowaniem się struktur typu szpilki od włosów o budowie podwójnego heliksu między komplementarnymi fragmentami łańcucha polirybonukleotydowego (p jest symbolem reszty estrowej kwasu ortofosforowego)

wdopodobieństwo występowania reszt poszczególnych aminokwasów w konformacjach heliksu α, struktury β i mniej regularnych konformacjach warunkujących zaginanie się łańcuchów. Następnie stosując metody termodynamiki statystycznej oblicza się dla kooperatywnego. Wśród biopolimerów nie stwierdzono dotychczas antykooperatywności, tj. $\sigma > 1$. Dla układów o złożonej budowie nie można otrzymać rozwiązań w postaci analitycznej ze względu na ogromne trudności matematyczne. Dotyczy to zwłasz-



Rys. 8. Samoorganizacja makrocząsteczek globularnego białka: *I* proces powstawania zarodków struktur α-helikalnych, pętli i zagięć łańcucha polipeptydowego w wyniku lokalnych oddziaływań między sąsiadującymi ogniwami; *II* w wyniku oddziaływań między oddalonymi fragmentami łańcucha formują się centra krystalizacji globularnej typu szpilki od włosów; *III* powstanie przejściowej struktury globularnej w wyniku połączenia się kilku centrów krystalizacji; *IV* ostateczne uformowanie się konformacji globuli po adjustacji kątów konformacyjnych łańcucha i wzajemnych orientacji segmentów do wartości odpowiadających ogólnemu minimum energii układu

polipeptydu o znanej budowie łańcucha prawdopodobieństwo przyjmowania przez jego segmenty konformacji odpowiadających tym formom uporządkowania. Otrzymuje się dość dobrą zgodność przewidywań teoretycznych z danymi krystalograficznymi.

Statystyczno-termodynamiczna teoria przejść konformacyjnych

Statystyczno-termodynamiczna teoria przejść konformacyjnych została opracowana dla wielu jedno-, dwu- i trójwymiarowych układów makrocząsteczkowych i superczasteczkowych o właściwościach kooperatywnych. Opisuje ona przejścia między różnymi konformacyjnie uporządkowanymi stanami łańcuchów, między danym stanem uporządkowanym i i stanem kłębka statystycznego, a także przejścia polegające na zmianie wzajemnej orientacji lub stopnia polimeryzacji makrocząsteczek tworzących funkcjonalne kompleksy nadmolekularne zbudowane z wielu identycznych lub różnych podjednostek.

funkcja rozdziału

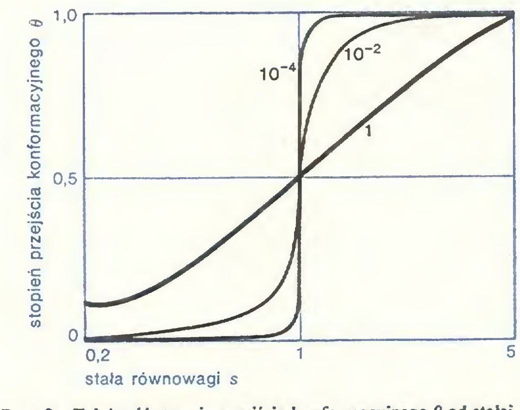
Jeżeli znana jest liczba *i* dostępnych makrocząsteczce lub układowi makrocząsteczek stanów konformacyjnych o energii E_i i statystycznej wadze g_i, można obliczyć funkcję rozdziału $Q = \sum_i g_i e^{-E_i/kT}$. Wiąże ona molekularne i termodynamiczne właściwości układu, ponieważ energia swobodna układu G zależy następująco od jej wartości: $G = -RT\ln Q$. Posługując się rachunkiem macierzowym wyprowadza się funkcję analityczną wiążącą doświadczalnie obserwowany stopień przejścia θ z parametrami termodynamicznymi układu — stałymi równowagi, i z temperaturą (ewentualnie innym parametrem charakteryzującym czynnik przesuwający tę równowagę). W najprostszym wypadku przejścia konformacyjnego realizującego się w dużym, jednorodnym i jednowymiarowym układzie (długi łańcuch zbudowany z identycznych ogniw, dla którego można pominąć efekty brzegowe w procesie nukleacji, czyli tworzenie się zarodka konformacji), parametrami termodynamicznymi układu są: stała równowagi s procesu wzrostu uporządkowania po nukleacji oraz stała równowagi tego ostatniego procesu σs. Współczynnik σ określa kooperatywność układu. Zależy ona wykładniczo od zmiany energii swobodnej procesu nukleacji:

$$\sigma = e^{-G_{\text{nukl}}/kT}.$$

Gdy $G_{\text{nukl}} > 0$, to $\sigma < 1$, czyli układ jest kooperatywny, natomiast gdy $G_{\text{nukl}} = 0$, również $\sigma = 0$ i zmiany zachodzące w układzie nie mają charakteru cza układów trójwymiarowych, takich jak białka globularne. Udaje się jedynie oszacować metodami numerycznymi przybliżone wartości parametrów termodynamicznych uproszczonego modelu układu.

zastosowanie modelu Isinga

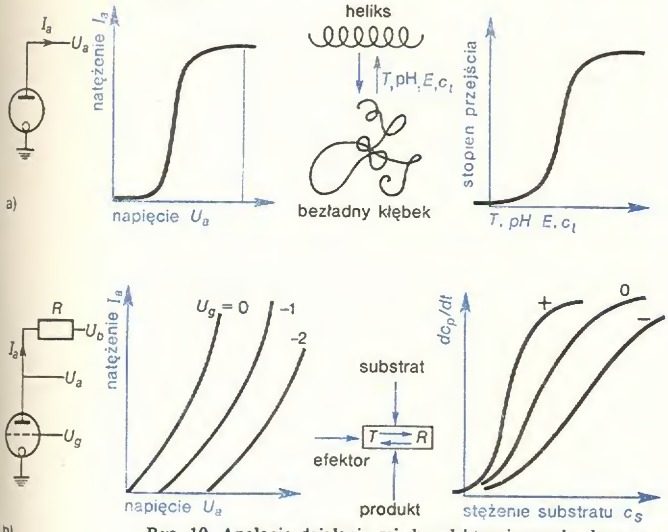
Wszystkie podstawowe właściwości termodynamiczne układu kooperatywnego wyjaśnia w prosty i jednocześnie ogólny sposób teoria przejść konformacyjnych w układzie jednowymiarowym, opracowana przez B. Zimma i J. Bragga. Oparta jest ona na modelu Isinga jednowymiarowej sieci spinów elektronowych ferromagnetyka, rozważanym w teorii przejść fazowych ferromagnetyk ⇌ paramagnetyk w punkcie Curie. Model ten zakłada występowanie elementów układu w dwóch tylko stanach *A* i *B* oraz kooperatywność ograniczoną do oddziaływań między bezpośrednio sąsiadującymi ze sobą elementami. Oznacza to, że *i*-ty element w sąsiedztwie elementu, np. w stanie *B* wykazuje tendencję do przyjęcia tego samego stanu. Model ten dobrze opisuje wiele rzeczywistych układów, np. przejście między dwiema helikalnymi konformacjami poliproliny I i II, polegające na izomeryzacji *cis–trans* wiązania peptydowego (wówczas wartości θ równe 1 i 0 odpowiadają formom I i II). Dla długiego, jednorodnego łańcucha spełniającego założenia takiego modelu wykres funkcji $\theta = f(s, \sigma s)$ pokazuje, że ostrość przejścia (nachylenie krzywej) zależy od wartości σ (rys. 9). Im mniejsza wartość tego współczynnika, tym większa jest liczba $N_0 = 1/\sqrt{\sigma}$ sąsiadujących ze sobą ogniw zmieniających jednocześnie swoją konformację (tzw.



Rys. 9. Zależność stopnia przejścia konformacyjnego θ od stałej równowagi s przy różnych wartościach współczynnika kooperatywności δ ($\delta = 1 \cdot 10^{-2}, 10^{-4}$) w jednowymiarowym modelu Isinga liniowego polimeru

długość kooperatywna) i tym samym ostrzejsze, bardziej kooperatywne przegięcie. Oznacza to, że w układach o silnej kooperatywności (σ rzędu 10^{-3}–10^{-4}) nawet niewielka zmiana równowagi $\partial \ln s/\partial x$ wywołana działaniem bodźca x na układ (zmiana ta jest proporcjonalna do $\sqrt{\sigma}$) pociąga za sobą rozległe zmiany konformacyjne. Makrocząsteczkowe układy kooperatywne mogą zatem spełniać analogiczne funkcje jak wyzwalające układy elektroniczne (rys. 10a). Jeżeli funkcja danego układu kooperatywnego zależy od jego stanu konformacyjnego (np. funkcje katalityczne białek enzymatycznych lub transportowe białek błon komórkowych), wówczas bodźce wyzwalające w nim przejścia między stanami o dużej i małej aktywności spełniają rolę regulatorów jego działania. Układy takie zachowują się analogicznie jak elektroniczne układy wzmacniające, takie jak np. trioda lampowa lub odpowiedni układ tranzystorowy (rys. 10b).

analogia z układami przełączającymi



Rys. 10. Analogia działania między elektronicznymi a kooperatywnymi układami makromolekularnymi: a) Przełączanie: schemat i charakterystyka prądowo-napięciowa diody lampowej oraz schemat i charakterystyka przejścia konformacyjnego wywołanego działaniem bodźca zewnętrznego (T oznacza temperaturę, pH ujemny logarytm stężenia jonów wodorowych, E pole elektryczne, c_1 — stężenie ligandu). b) Wzmacnianie: schemat i charakterystyka prądowo-napięciowa triody lampowej oraz schemat allosterycznej regulacji aktywności enzymu (T i R stany enzymu o niskiej i wysokiej aktywności katalitycznej) i wykres zależności szybkości reakcji od stężenia substratu (c_s) w obecności allosterycznego aktywatora (+) i inhibitora (−) katalizy

czas przebiegu przemiany

Czas przebiegu elementarnych procesów związanych z przemianami konformacyjnymi jest rzędu 10^{-10} s. Całkowity czas przebiegu przemiany obejmującej wiele ogniw makrocząsteczki jest oczywiście odpowiednio dłuższy, jednakże wystarczająco krótki w skali czasu procesów komórkowych dla zapewnienia sprawnej kontroli sterowania tymi procesami.

Małocząsteczkowe układy molekularne, w których zachodzą rzeczywiste przejścia fazowe I rodzaju typu ciecz–kryształ nie mają takich właściwości. Zmiany struktury, a więc i objętości oraz wszystkich potencjałów termodynamicznych, następują w nich skokowo. Wymagają one znacznie większych energii aktywacji i przebiegają wobec tego znacznie wolniej niż zmiany konformacyjne w kooperatywnych układach makrocząsteczkowych.

Kooperatywne właściwości biopolimerów a regulacja ich funkcji

Cząsteczki biopolimerów spełniają wszystkie warunki układów o dużym stopniu kooperatywności. Zmiany ich stanu konformacyjnego mogą zachodzić w warunkach fizjologicznych, jeśli chodzi o temperaturę, stężenie elektrolitów oraz stężenie jonów wodorowych. Rolę bodźca wyzwalającego zmianę stanu równowagi termodynamicznej makrocząsteczki, oprócz zmian temperatury i stężenia określonych jonów, może odgrywać wiele innych czynników fizycznych i chemicznych.

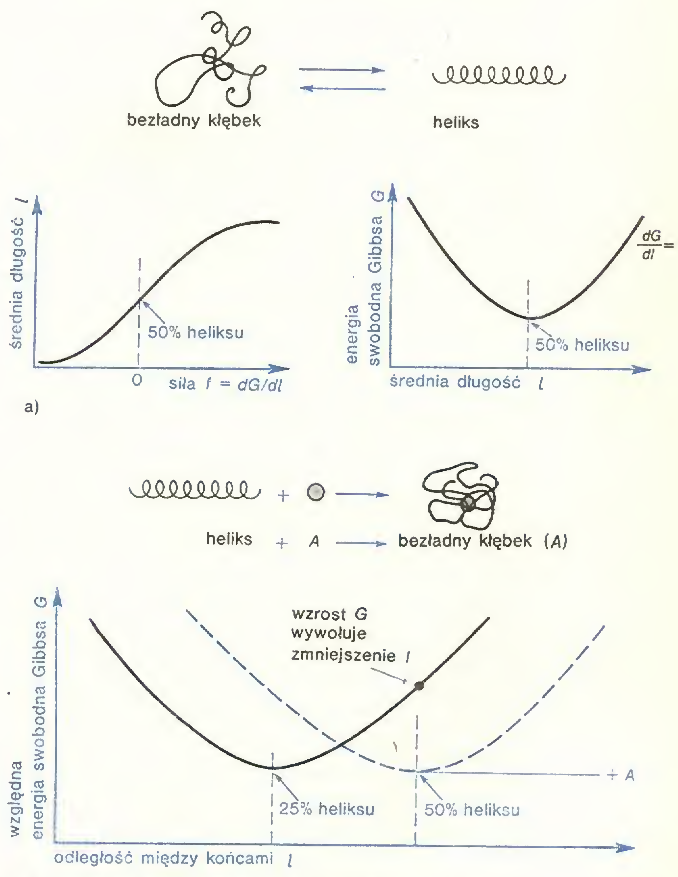
pole elektryczne jako bodziec

Energia cząsteczek polielektrolitów — jako układów elektrostatycznych — zależy od wielkości przyłożonego pola elektrycznego. W wielu peptydach obserwowano *in vitro* przejścia konformacje wywoływane działaniem pól o potencjale porównywalnym z potencjałem błon biologicznych *in vivo*.

czynniki chemiczne

Wiadomo również, że wiele małocząsteczkowych i makrocząsteczkowych ligandów pełniących w komórce różnorodne funkcje regulacyjne i kontrolne ma większe powinowactwo do jednej z pozostających w równowadze konformacji polimeru lub jego fragmentu i wiążąc się z nim przesuwa równowagę w kierunku tej konformacji. Na przykład aminy alifatyczne i białka zasadowe — histony, odgrywające istotną rolę w organizacji i funkcjonowaniu chromatyny, stabilizują podwójny heliks DNA. Inne białka tzw. rozplatające, które uczestniczą w procesie replikacji DNA fagów, bakterii i organizmów wyższych (np. białko, będące produktem genu 32 bakteriofaga T4), obniżają temperaturę przejścia konformacyjnego DNA heliks ⇌ kłębek. Wiązanie przez cząsteczki białek enzymatycznych cząsteczek substratów, koenzymów i regulatorów allosterycznych wywołuje zmiany stanu konformacyjnego pewnych obszarów globuli białkowej.

Przyczyną zmian stanu termodynamicznego makrocząsteczek w procesach oddziaływania z innymi cząs-



Rys. 11. Generacja i działanie siły w układzie kooperatywnym: a) generacja siły wywołana zmianą stanu równowagi konformacyjnej heliks ⇌ kłębek bezładny, b) po związaniu cząsteczki ligandu A o większym powinowactwie do kłębka wzrost swobodnej energii układu wywołuje jego kontrakcję $l_{heliks} > l_{kłębek}$

teczkami jest porównywalność energii oddziaływań wewnątrzcząsteczkowych i międzycząsteczkowych. Dalszą konsekwencją tego stanu rzeczy jest sprzężenie energetyczne różnych ligandów związanych z różnymi fragmentami makrocząsteczki. Innymi słowy, działanie danego bodźca na układ może wzmagać lub niweczyć jednoczesne działanie innego.

przemiany energetyczne wywołane przejściem konformacyjnym

Reakcja kooperatywnego układu na działanie zewnętrznego bodźca w formie przejścia konformacyjnego może z kolei być źródłem generacji siły: $f = -\delta G/\partial x$ mechanicznej (rys. 11), osmotycznej, chemicznej lub elektrycznej. Energia swobodna G makrocząsteczki w roztworze jest bowiem, ogólnie rzecz biorąc, funkcją jej rozmiarów liniowych, gradientu stężenia jonów i wiążących się z makrocząsteczką cząsteczek związków organicznych, a także powierzchniowego potencjału elektrycznego. Warunkiem pojawienia się takiej siły jest oczywiście anizotropowa budowa danej cząsteczki jak i większego, otaczającego ją układu, gdyż tylko wówczas działanie bodźca i reakcja układu może mieć charakter kierunkowy. Praca wykonywana działaniem tych sił w komórce, która jest układem silnie anizotropowym, jest związana z przemianą różnych form energii chemicznej związków wysokoenergetycznych, takich jak ATP (o wysokim potencjale chemicznym), stanowiących podstawowe źródło energii komórki, w inne formy energii chemicznej (reakcje enzymatyczne), a także w energię osmotyczną (aktywny transport przez błony komórkowe), mechaniczną (m.in. skurcz mięśnia) i elektryczną (potencjały błon komórkowych). Tak więc kooperatywne przejścia konformacyjne makrocząsteczek stanowią podstawę wielu elementarnych mechanizmów funkcjonowania układów komórkowych na poziomie molekularnym, a także funkcjonalnych powiązań układów o różnych funkcjach biochemicznych między sobą.

Regulacja układów enzymatycznych

Dobrze poznane pod względem biochemicznym i kinetycznym procesy regulacji działania enzymów i kompleksów enzymatycznych (→ Białka) dostarczają szczególnie dużo przykładów powiązań funkcjonalnych, opartych na kooperatywnych właściwościach cząsteczek białek. Zapewne najprostszym przykładem jest allosteryczna regulacja aktywności katalitycznej enzymów zbudowanych z pojedynczej cząsteczki białka. Wiązanie przez białko w miejscu odległym od jego centrum katalitycznego cząsteczki efektora reakcji, powstającej w innym procesie enzymatycznym lub transportowanej do komórki przez otaczającą ją błonę, wywołują zmianę konformacyjną w miejscu wiązania substratu, zwiększającą lub obniżającą szybkość reakcji enzymatycznej.

allosteryczna regulacja aktywności katalitycznej

Znacznie bardziej złożony jest kooperatywny mechanizm regulacji enzymów zbudowanych z kilku cząsteczek białek tworzących trwały termodynamicznie kompleks. W kompleksach takich mogą zachodzić zmiany wzajemnego położenia podjednostek, wywołane lokalnymi zmianami konformacji tych podjednostek. Różnym stanom funkcjonalnym takiego układu odpowiadają różne konfiguracje ułożenia podjednostek w kompleksie. W supercząsteczkowych kompleksach białkowych może realizować się sprzężenie dwóch różnych reakcji enzymatycznych. Na przykład, związanie substratu lub uwolnienie produktu reakcji w jednym z centrów katalitycznych może wywoływać zmianę konformacyjną w cząsteczce białka, która w wyniku konformacyjnego sprzężenia dalekiego zasięgu indukuje aktywację centrum katalitycznego innej reakcji biochemicznej. Przypuszcza się, że tego typu mechanizm może być odpowiedzialny za sprzężenie transportu elektronów i syntezy ATP z ADP w procesie oddychania w mitochondriach (→ Organizacja procesów życiowych komórki). Zmiana stanu utlenienia atomu żelaza w cytochromie C, jednym z białek tworzących łańcuch transportu elektronów, wywołuje w nim rozległą zmianę konformacyjną, która reguluje aktywność ATPazy — enzymu sąsiadującego z nim w kompleksie supercząsteczkowym. W ten sposób obywa się przemiana chemicznej energii substratów łańcucha oddechowego w energię ATP — uniwersalne źródło energii chemicznej w komórce.

supercząsteczkowe kompleksy białkowe

Regulacja funkcji błon komórkowych

przejścia konformacyjne w warstwach fosfolipidowych

Zbliżone swym charakterem do przejść fazowych ciekłych kryształów przejścia konformacyjne w podwójnych warstwach fosfolipidowych błon komórkowych (→ Błony komórkowe, Ciekłe kryształy) odgrywają prawdopodobnie istotną rolę w regulacji ich różnorodnych funkcji. Doświadczenia z modelowymi błonami pokazują, że w podwójnej warstwie lipidowej duża część energii swobodnej układu związana jest z oddziaływaniami elektrostatycznymi polarnych i zjonizowanych grup fosfoestrowych na jej powierzchni. Zmiany gęstości powierzchniowej ładunku, a tym samym i powierzchniowego potencjału elektrycznego błony, wywołane przesunięciem równowagi dysocjacji kwasowej tych grup oraz adsorpcją na ich powierzchni jonów metali dwuwartościowych (Ca^{++}), przesuwają równowagę fazową uporządkowania łańcuchów węglowodorowych we wnętrzu warstwy. Obniżenie gęstości ładunku zwiększa stan uporządkowania łańcuchów, podczas gdy jego wzrost wywołuje odwrotny efekt. Jony Na^{+} i K^{+} obniżają potencjał powierzchniowy błony zwiększając stopień dysocjacji kwasowej grup fosfoestrowych i obniżając temperaturę przejścia fazowego. Natomiast jony H^{+} i Ca^{++} ulegają adsorpcji na powierzchni błony, redukują jej ujemny ładunek powierzchniowy i zwiększają temperaturę przejścia. Tak więc przejścia fazowe wywołane w stałej temperaturze działaniem jednego czynnika, np. H^{+}, Na^{+}, zgodnie z kooperatywnymi właściwościami układu wywołują duże zmiany w równowadze wiązania przez błonę wapnia Ca^{++}. Błona może zatem być kontrolowanym rezerwuarem jonów wapnia, odgrywających istotną rolę regulacyjną w wielu procesach, np. w regulacji skurczu mięśnia oraz w funkcjonalnym sprzężeniu rodopsyny z zakończe niami synaptycznymi układu nerwowego w narządzie wzroku.

cząsteczki białek w błonach

W naturalnych błonach biologicznych cząsteczki białek, spełniające różnorodne funkcje enzymatyczne, transportowe i sygnalizacyjne, są „zanurzone" w warstwach lipidowych. Zapewnia to wzajemne sprzężenie konformacyjne warstwy lipidowej i białek oraz białek między sobą, a tym samym powiązanie funkcjonalnej błony i cząsteczek białek.

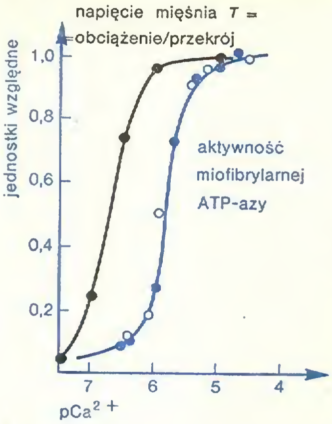
Regulacja skurczu mięśnia

Innym, bardziej złożonym przykładem powiązań funkcjonalnych zachodzących dzięki kooperatywnym zmianom w molekularnej organizacji układu jest włókno mięśniowe. W tym wypadku kooperatywne oddziaływania elementów jego budowy obejmują całą komórkę. Złożony proces przemiany energii chemicznej w energię mechaniczną w trakcie skurczu włókien mięśniowych przejawia się jako widoczne pod mikroskopem przesuwanie się cienkich (aktynowych) filamentów względem nieruchomych filamentów grubych (miozynowych) ku centralnej części sarkomeru (→ Molekularne podstawy skurczu mięśnia). Proces inicjowany jest przez jony Ca^{+2} uwalniane z błony siateczki sarkoplazmatycznej komórki mięśniowej w wyniku pobudzenia przez impuls nerwowy. Zarówno napięcie mięśnia jak i szybkość hydrolizy cząsteczki ATP (źródła chemicznej energii dla pracy mięśnia), związanej z miozyną, krytycznie zależą od stężenia jonów wapnia w sposób typowy dla procesów kooperatywnych (rys. 12).

rola jonów wapnia

Kooperatywność procesu makroskopowego jest wynikiem ciągu następujących po sobie kooperatyw-

nych zmian w molekularnej organizacji włókna mięśniowego, inicjowanych przyłączeniem jonów Ca^{+2} do jednej z podjednostek troponiny — białka regulującego skurcz, składnika filamentów cienkich. Związanie



Rys. 12. Napięcie mięśnia i aktywność miofibrylarnej ATP-azy w zależności od stężenia jonów Ca^{+2} (pCa^{+2} ujemny logarytm stężenia); wg A. Weber *Functional Linkage in Biomolecular Systems*, New York 1975

jonów wapnia wywołuje przejście konformacyjne w troponinie, które pociąga za sobą zmianę położenia innego białka regulującego, obecnego w filamencie cienkim — tropomiozyny, względem tworzących rdzeń filamentu czesteczek aktyny. W ten sposób, w wyniku przyłączenia jonów Ca^{2+} do jednej cząsteczki troponiny zostają odsłonięte miejsca reagowania z miozyną w siedmiu sąsiadujących ze sobą monomerach łaktyny. Łącząc się z globularnymi zakończeniami cząsteczek miozyny w grubym filamencie, stanowiącym część tzw. poprzecznych mostków miozynowych, aktyna przyspiesza hydrolizę ATP w centrum katalitycznym miozyny zlokalizowanym w mostku. W trakcie hydrolizy zachodzi zmiana konformacji zarówno globularnej jak i liniowej części cząsteczki miozyny tworzącej mostek, a jej wynikiem jest przesunięcie cienkiego filamentu względem grubego. Także uporządkowanie cząsteczek miozyny w grubym filamencie, przejawiające się w helikalnym ułożeniu ich globularnych zakończeń wzdłuż osi filamentu, ma charakter kooperatywny i kontrolowane jest przez stężenie elektrolitu.

fizyczny model skurczu mięśnia

Mimo stosunkowo dobrze poznanej molekularnej organizacji włókien mięśniowych oraz zachodzących w niej w trakcie skurczu zmian, ciągle hipotetyczny jest charakter sił powodujących przesuwanie się filamentu aktynowego. Wiele przesłanek doświadczalnych wskazuje na istotne znaczenie w tym procesie oddziaływań elektrostatycznych między cyklicznie generowanymi centrami na powierzchni stykających się filamentów w trakcie „włączania" aktyny przez jony Ca^{+2} oraz hydrolizy ATP w zakończeniach mostków poprzecznych miozyny (powstają wówczas aniony dwufosforanu adenozyny i kwasu fosforowego). Również zmiany w elastyczności konformacji mostków związane z hydrolizą ATP niewątpliwie odgrywają w nim rolę. Dzięki elastyczności budowy filamentów pojawienie się siły odpychającej, skierowanej poprzecznie do nich, musi powodować izometryczne napięcie mięśnia wzdłuż osi filamentów i jego skurcz w wyniku przesunięcia się filamentów cienkich. Elektrostatyczne oddziaływania odpychające między filamentami oraz elastyczność ich budowy leżą u podstaw wielu modeli skurczu mięśnia, opisujących fizjologiczne i termodynamiczne właściwości mięśni. W molekularnym mechanizmie skurczu mięśnia dopatrujemy się obecnie ogólniejszych praw rządzących ukierunkowanym ruchem makrocząsteczek w cytoplazmie.

L. A. Blumenfeld *Problemy fizyki biologicznej*, Warszawa 1978; W. N. Cwietkow i in. *Struktura makrocząsteczek w roztworach*, Warszawa 1968; W. Hoppe i in. *Biophysik*, Berlin 1977; H. Morawetz *Fizykochemia roztworów makrocząsteczek*, Warszawa 1970; F. O. Schmitt i in. *Functional linkage in biomolecular systems*, Nowy Jork 1975; M. W. Wolkensztein *Biologia molekularna*, Warszawa 1969; M. W. Wolkensztein *Molekularnaja biofizika*, Moskwa 1975.

Organizacja procesów życiowych komórki

Tadeusz Kłopotowski

Planeta nasza zamieszkała jest przez ogromną liczbę różnych gatunków organizmów. Różnią się one między sobą nie tylko wielkością i kształtem, lecz również odmiennymi warunkami życia oraz stopniem złożoności ich budowy i procesów fizjologicznych.

bakteriofagi

Zróżnicowanie świata żywego rozpoczyna się od form tak prostych jak bakteriofagi i wirusy, które zawierają jedynie kwasy nukleinowe (RNA lub DNA) i są otoczone warstwą identycznych cząsteczek białkowych oraz mają struktury niezbędne do wniknięcia lub wstrzyknięcia samych tylko kwasów nukleinowych do komórek. Bakteriofagi i wirusy nie wykazują żadnej przemiany materii poza komórkami. DNA bakteriofagu po wniknięciu do komórki jest włączany w określone mu miejsce w chromosomie albo też zawarte w tym chromosomie geny podlegają procesowi uaktywnienia. W tym drugim wypadku następują kolejno zmiany w metabolizmie komórki: produkcja swoistych białek regulatorowych i enzymów, replikacja DNA pasożyta, uformowanie nowych bakteriofagów, które po rozpadzie komórki wydobywają się do otoczenia. Po napotkaniu nowych wrażliwych komórek cykl ten się powtarza od wniknięcia, przez moment decyzji — integracja lub ekspresja informacji — do stabilizacji lub mnożenia się.

bakterie

Najprostszymi organizmami, czyli jednostkami zdolnymi do samodzielnej wegetacji, są bakterie. Są to twory jednokomórkowe, zdolne do samodzielnego metabolizmu i reprodukcji bez związku z innymi komórkami. Istnieje ogromna liczba gatunków bakterii. Różnią się one swoją budową genetyczną, a przez to strukturą poszczególnych elementów budowy, przebiegiem dróg metabolicznych i wymogami życiowymi. Bakterie autotroficzne nie wymagają do życia żadnych związków organicznych z zewnątrz, inne zwane prototroficznymi, wymagają jedynie glukozy. Niektóre gatunki bakterii żyjące w organizmach roślinnych lub zwierzęcych wymagają do swego wzrostu również i innych związków organicznych z zewnątrz, np. witamin i aminokwasów. Ewolucja wykształciła gatunki bakterii o różnych możliwościach przystosowywania się do zmiennych warunków zewnętrznych: do głodu względnego — przez zatrzymanie jednych, a uruchomienie drugich szlaków metabolicznych, do życia w ciemności lub wykorzystywania energii świetlnej, do życia w warunkach tlenowych lub beztlenowych. Zakres zdolności adaptacyjnych jest cechą charakterystyczną danego gatunku i jest zdeterminowany genetycznie. Ponieważ materiał genetyczny, DNA, jest narażony na mutacje, czyli przypadkowe zmiany składu chemicznego, wszystkie gatunki, i to nie tylko bakterii, wykształciły w procesie ewolucji systemy enzymatyczne naprawy tych zmian.

prokarionty

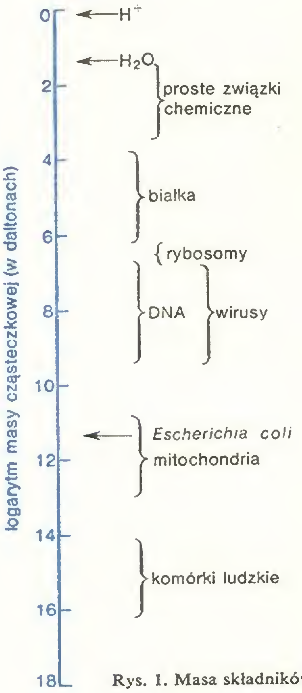
Informacja genetyczna dla pełnego cyklu reprodukcji bakterii oraz różnych procesów metabolicznych i adaptacyjnych zawarta jest w pojedynczych

kolistych cząsteczkach DNA. Na tej podstawie wszystkie zbadane pod tym względem bakterie zostały nazwane organizmami prokariotycznymi (prokariontami).

eukarionty

Stojące wyżej w rozwoju świata żywego organizmy eukariotyczne (eukarionty) mają w każdej komórce po kilka lub nawet kilkadziesiąt różnych chromosomów, będących linearnymi cząsteczkami DNA, zwykle w kompleksach z białkami. Stanowią one zasadniczą część jąder komórkowych, tworów otoczonych błonami o wybiórczej przepuszczalności. Najprostszymi eukariontami są drożdże i niektóre grzyby zdolne do bytu jednokomórkowego. Wszystkie organizmy wielokomórkowe, zarówno rośliny jak i zwierzęta, są eukariontami.

Komórki organizmów wyższych są zróżnicowane i tworzą odrębne tkanki, czyli zespoły komórek o takiej samej budowie i funkcji. Zróżnicowanie komórek polega na tym, że mają one nie wszystkie funkcje zakodowane w ich DNA, ale tylko ich część. W porównaniu z komórkami prokariotycznymi mają one ograniczone zdolności adaptacyjne, a wiele ich funkcji nie służy ich własnej wegetacji, lecz celom niezbędnym do trwania, rozwoju lub reprodukcji całego organizmu. Kombinacje tkanek tworzą wyspecjalizowane narządy, takie jak: wątroba, serce, mózg.



Rys. 1. Masa składników komórkowych i komórek

Złożoność świata żywego i rozmaitość struktury i funkcji jego podstawowych elementów, czyli komórek, powoduje oczywiste trudności we wszelkich próbach zwięzłego przedstawienia systemów organizacji procesów życiowych. Trudności te są tym większe, że tylko niektóre z tych procesów zostały poznane na tyle, by można było przedstawić choćby najważniejsze uczestniczące w nich cząsteczki, a także ich oddziaływania. A jeśli już do tego doszło, to poznane procesy mają często ograniczony zasięg występowania w świecie żywym (rys. 1).

Strukturalne elementy komórek

błony komórkowe

Każda komórka jest tworem oddzielonym od środowiska zewnęrznego lub innych komórek układem błon (→ Błony komórkowe). Funkcja tych błon polega na zabezpieczeniu stałości środowiska wewnętrznego komórki przez ograniczenie swobodnej dyfuzji składników znajdujących się w jej wnętrzu na zewnątrz, jak również związków chemicznych, znajdujących się w otoczeniu, do wnętrza komórki. Ponadto, błony te stanowią siedzibę układów transportujących selektywnie związki chemiczne oraz receptorów służących do rozpoznawania zmian środowiska lub innych, bardziej swoistych bodźców.

Poszczególne błony danej komórki mają różną budowę chemiczną i mało jeszcze wiadomo o poprzecznych powiązaniach między nimi. U niektórych bakterii błona najbardziej zewnętrzna jest oddzielona od tzw. sztywnej ściany przestrzenią peryplazmatyczną. Zachowuje ona własne środowisko osmotyczne, lecz jest względnie łatwo dostępna dla związków chemicznych pochodzących z zewnątrz. W przestrzeni peryplazmatycznej nie zachodzą żadne reakcje wymagające energii. Zawiera ona enzymy hydrolizujące estry lub naturalne polimery, a także białka uczestniczące w aktywnym transporcie, który dopiero w dalszych etapach wymaga nakładu energii.

Najpowszechniejszym elementem funkcjonalnym błon komórkowych są liczne, lecz mało poznane kompleksy transportujące selektywnie proste jony lub nie zjonizowane związki chemiczne. Poznano ich właściwości kinetyczne, ale tylko w nielicznych wypadkach udało się zidentyfikować białka uczestniczące w tych procesach. Stwierdzono np., że enzymy fosforylujące cukry są silnie związane z błonami bakterii i udowodniono, że fosforylacja ta jest niezbędna do pobierania prostych cukrów z zewnątrz. W błonach komórek zwierzęcych znajduje się enzym, umożliwiający dostarczanie energii do procesu wymiany kationów wbrew gradientom stężeń.

W błonach bakterii zlokalizowane są enzymy układu oddechowego niezbędnego w gospodarce energetycznej komórki oraz specjalne struktury przetwarzające energię metaboliczną na ruch rzęsek. Dzięki temu mogą się one kierować w stronę większych stężeń potrzebnych im związków chemicznych. Zdolność tę nazwano chemotaksją. Układ czuciowy służący do rozpoznawania gradientów tych związków lub związków toksycznych jest zlokalizowany w błonach, lecz jedynymi poznanymi dotąd składnikami tego układu są wspomniane wyżej białka przestrzeni peryplazmatycznej, uczestniczące również w aktywnym transporcie.

cytoplazma

Wnętrze komórki stanowi cytoplazma. W płynnym składniku cytoplazmy, zwanym ostatnio cytosolem, zachodzą procesy metaboliczne. W cytoplazmie znajdują się elementy strukturalne, takie jak jądro oraz inne, zwane ogólnie organellami. Jądro i organelle mają własne błony o selektywnej przepuszczalności, która zapewnia im autonomię środowiska wewnętrznego.

jądro

Zasadniczym składnikiem jąder są chromosomy, które są nośnikami informacji genetycznej. Komórki eukariotyczne mają w jądrach podwójne komplety chromosomów odziedziczone po rodzicach. Dzięki temu każdy ich gen występuje w dwu wersjach rodzicielskich. W danej komórce tylko niektóre pary genów są aktywne, tzn. podlegają procesowi transkrypcji, której produktem jest komplementarny mRNA (→ Kwasy nukleinowe).

Aktywność genów jest regulowana w zależności od bodźców wewnętrznych lub zewnętrznych, które mogą być m.in. przenoszone przez rekacje hormonów ze swoistymi receptorami na powierzchni komórek. W jądrach znajdują się składniki układu replikacji chromosomów i ich segregacji, zapewniające komórkom potomnym otrzymanie takich samych kompletów chromosomów. Tylko niektóre bardzo wyspecjalizowane komórki, np. krwinki czerwone ssaków, nie mają jąder.

mitochondria

Organellami powszechnie występującymi w organizmach eukariotycznych są mitochondria. Ich główną, a może nawet jedyną rolą w komórce jest przetwa-

rzanie energii chemicznej metabolitów w energię wysokoenergetycznych pirofosforanowych wiązań ATP. Elektrony uwolnione z utlenianych związków są przenoszone wzdłuż łańcucha swoistych białek, w tym cytochromów (→ Białka), o coraz to wyższych potencjałach utleniania i redukcji. Przypuszcza się, że przestrzenne usytuowanie tych białek względem siebie sprzyja procesowi transportu elektronów. Są one zlokalizowane na ułożonych równolegle przegrodach nadających obrazom uzyskanym za pomocą mikroskopu elektronowego charakterystyczny wygląd. Ważnym elementem mitochondriów jest ich błona, gdyż jej potencjał jest najistotniejszym elementem w procesach przekształcania energii. W komórce bakterii funkcję tę spełnia błona komórkowa.

Mitochondria zawierają własne koliste cząsteczki DNA o wielkości odpowiadającej ok. jednej setnej chromosomu bakterii. Treść genetyczna tego DNA najlepiej jest poznana u drożdży piekarskich, będących najprostszymi organizmami eukariotycznymi.

Mitochondria mają odrębny układ biosyntezy białka. Jego składniki i właściwości przypominają bakteryjny układ biosyntezy białka. Podobieństwo to dotyczy np. wielkości rybosomów oraz wrażliwości biosyntezy na niektóre antybiotyki. Dało ono asumpt hipotezie, że we wczesnych etapach ewolucji doszło do symbiozy jakichś prabakterii i innych komórek. W dalszych jej etapach powstały komórki eukariotyczne z mitochondriami jako reliktami tych prabakterii.

chloroplasty

Komórki roślinne zawierają organelle nazywane chloroplastami, które są miejscem, gdzie przebiegają procesy fotosyntezy. Zasadniczym ich składnikiem są barwniki, takie jak chlorofil, umożliwiające wykorzystanie energii słonecznej. Podobnie jak mitochondria, chloroplasty zawierają DNA i mają układ biosyntezy białka podobny do bakteryjnego. Podobna jest też hipoteza ich powstania w dawnych etapach ewolucji.

cytosol

Wiele ważnych procesów metabolicznych zachodzi poza omówionymi już elementami strukturalnymi komórki, w jej płynnym składniku — cytosolu. Są to zarówno procesy kataboliczne, z których najważniejszym jest glikoliza czyli przemiana glukozy poprzez estry fosforowe do pirogronianu, jak i liczne procesy anaboliczne, polegające na syntezie elementów struktury komórkowej z prostych metabolitów.

Najważniejszym procesem anabolicznym jest biosynteza białka. Głównymi czynnikami w tym procesie są informacyjne RNA, tzw. mRNA (→ Kwasy nukleinowe), będące rybonukleinowymi wersjami genów, rybosomy, tj. makrocząsteczkowe kompleksy swoistego RNA z kilkudziesięcioma różnymi białkami, kilkadziesiąt różnych tRNA, z których każdy jest związany z właściwym mu aminokwasem, wiele białek o swoistych funkcjach oraz ATP i GTP (guanozynotrójfosforan) jako źródła energii dla tworzenia wiązań peptydowych. W trakcie procesu translacji są syntetyzowane białka wg programu zakodowanego w mRNA.

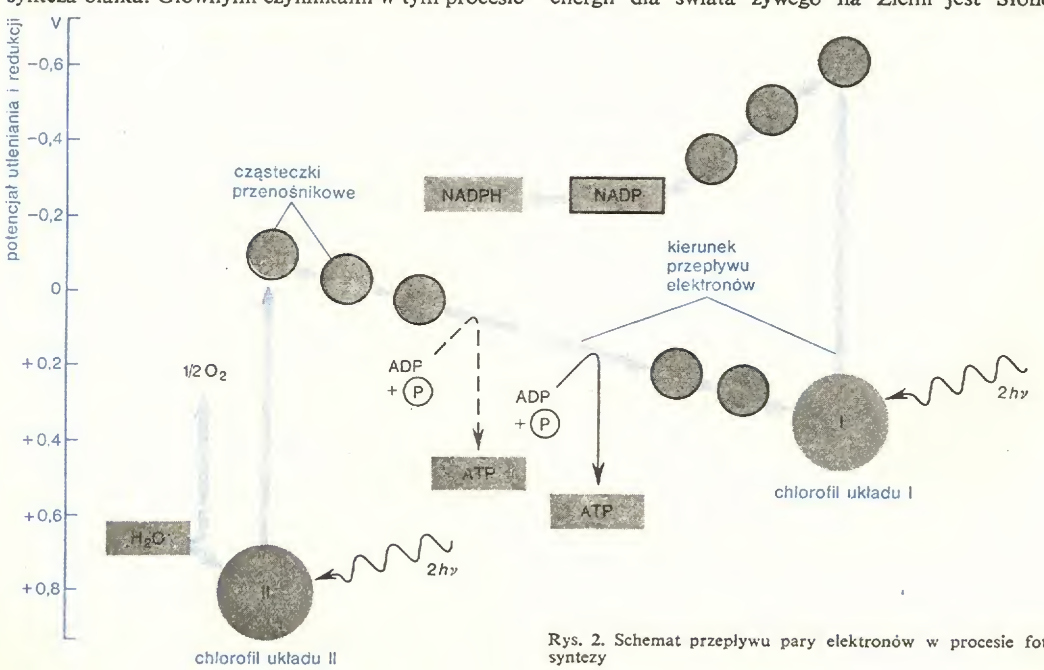
U bakterii proces ten zachodzi w całej cytoplazmie i nie wykazuje żadnej szczególnej lokalizacji. W komórkach eukariotycznych natomiast znaczna część biosyntezy białka zlokalizowana jest na powierzchni błoniastego labiryntu stanowiącego tzw. siateczkę endoplazmatyczną. Ważną jej funkcją jest wytwarzanie błony jądrowej. Przypuszcza się, że związek biosyntezy białka z siateczką endoplazmatyczną umożliwia odpowiednie rozprowadzenie mRNA, a także nowo utworzonych białek, szczególnie tych, które mają być przemieszczone w kierunku powierzchni komórki — ku jej błonom lub na zewnątrz. Białka te zostają przeniesione do innego błoniastego układu zwanego aparatem Golgiego. W jego woreczkowatych elementach (wakuolach) białka ulegają modyfikacjom, np. glikozylacji przez przyłączenie reszt cukrowych. Wakuole są zróżnicowane czynnościowo: jedne służą do transportu nowo zsyntetyzowanych białek do miejsca ich przeznaczenia — tzn. do błony lub na zewnątrz komórki, inne transportują stamtąd związki chemiczne i jednocześnie odpowiednio je przekształcają. Szczególnymi wakuolami są lizosomy, zawierające enzymy, których ostatecznym przeznaczeniem może być zniszczenie komórki po jej obumarciu. Inną funkcję mają peroksysomy, zawierające enzymy katalizujące utlenianie związków chemicznych przy użyciu nadtlenku wodoru.

siateczka endoplazmatyczna

aparat Golgiego

Podstawowe procesy metaboliczne

Do przebiegu procesów życiowych wszystkich komórek jest potrzebna energia. Pierwotnym źródłem energii dla świata żywego na Ziemi jest Słońce.



Rys. 2. Schemat przepływu pary elektronów w procesie fotosyntezy

fotosynteza

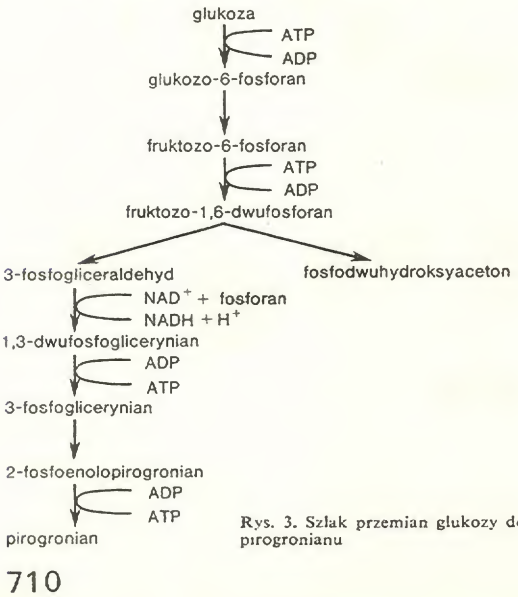
Energia jego promieniowania jest zamieniana w chloroplastach roślin i niektórych bakterii na energię wysokoenergetycznych wiązań pirofosforanowych w cząsteczce ATP, będących głównymi nośnikami biologicznej postaci energii (jest to tzw. proces fotosyntezy). W chloroplastach znajdują się 2 układy złożone z chlorofilu oraz układów przenośników elektronów, różnych dla każdego z tych układów (rys. 2). Absorpcja fotonu przez chlorofil układu *II* umożliwia hydrolizę wody i dostarczenie wzbudzonego elektronu do łańcucha utleniająco-redukujących przenośników. Jego przenoszenie przez kolejne ogniwa tego łańcucha, aż do chlorofilu układu *I* powoduje wytworzenie dwu wysokoenergetycznych wiązań pirofosforanowych. Absorpcja następnego fotonu przez chlorofil układu *I* sprawia, że wzbudzony elektron dostaje się do innego łańcucha przenośników. To przenoszenie elektronów, zgodnie z ich potencjałem utleniania-redukcji prowadzi do redukcji cząsteczki NADP, związku będącego koenzymem wielu enzymów utleniająco-redukujących (→ Białka). Cząsteczka NADP może ulec utlenieniu w innym układzie przenośników utleniająco-redukujących, a część zawartej w niej energii może przejść w energię trzech wysokoenergetycznych wiązań pirofosforanowych ATP.

Zmagazynowanie energii słonecznej w ATP umożliwia przebieg endoergicznej reakcji aktywacji węglanu i przyłączenia go do rybulozodwufosforanu. W wyniku tej reakcji powstają dwie cząsteczki fosfoglicerynianu. Ich przemiany regenerują rybulozodwufosforan, który może reagować z następną cząsteczką aktywnego węglanu. Sześciokrotne powtórzenie tego cyklu prowadzi w ogólnym bilansie przemian do powstania jednej cząsteczki glukozy.

glikoliza

Glukoza ulega glikolizie, dzięki czemu staje się źródłem zredukowanego węgla, tlenu i wodoru dla prawie wszystkich procesów biosyntetycznych roślin. W ciemnościach oraz w tych częściach roślin, które nie mają chloroplastów, jest ona również źródłem energii. Tę podwójną funkcję dostarczania energii oraz węgla i wodoru — spełnia glukoza również i dla przeważającej większości innych organizmów żywych.

Schemat wykorzystania glukozy ma niewiele wariantów w świecie żywym. Przemiana glukozy występująca najpowszechniej, została zbadana przez H. Embdena, O. Meyerhofa i J. Parnasa (rys. 3). Po przyłączeniu fosforanu cząsteczka tego cukru ulega kolejnym przekształceniom, w których wyniku powstają dwie cząsteczki pirogronianu, a część uwolnionej energii zostaje zachowana w postaci wysokoenergetycznych wiązań ATP i zredukowanych cząsteczek NAD, innego koenzymu reakcji utleniania — redukcji.



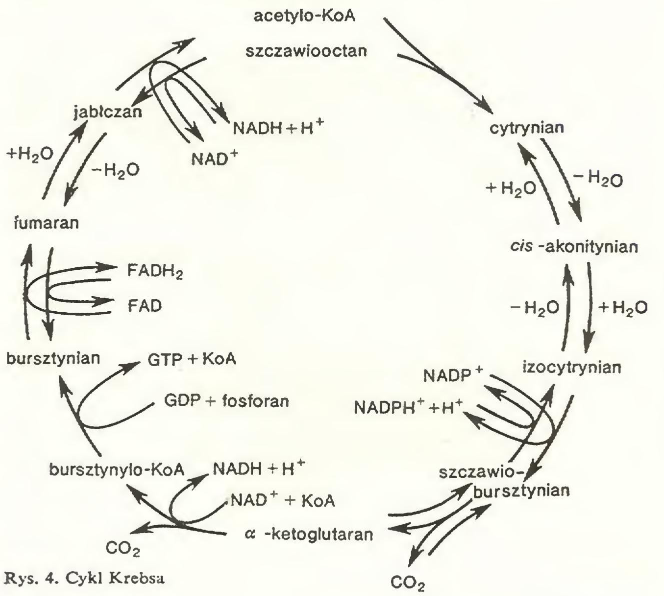
Rys. 3. Szlak przemian glukozy do pirogronianu

Odmienny rodzaj przemian glukozy prowadzi przez estry fosforanowe mające 3, 4, 5, 6 lub 7 atomów węgla. Są to w zasadzie te same reakcje, które uczestniczą w procesie asymilacji węglanu i syntezie glukozy. U organizmów niezdolnych do fotosyntezy ten cykl fosfopentozowy przemian glukozy służy głównie do produkcji fosforybozy, potrzebnej do syntezy kwasów nukleinowych oraz zredukowanego NADP niezbędnego dla redukcyjnych procesów biosyntetycznych, takich jak wiązanie atmosferycznego azotu u niektórych bakterii lub synteza kwasów tłuszczowych u wszystkich organizmów.

Jednakże najważniejszym i najpowszechniejszym sposobem odzyskiwania energii z glukozy jest proces przemian pirogronianu. Ulega on utlenieniu i odłączaniu CO_2 z udziałem kilku koenzymów i oczywiście swoistego enzymu. Produktami tej reakcji są acetylo-koenzym A, zredukowany NAD i dwutlenek węgla. Acetylo-koenzym A reaguje ze szczawiooctanem.

cykl Krebsa

Powstający z tej reakcji cytrynian jest metabolizowany w dalszych etapach cyklu Krebsa. Polega on na dziesięciu reakcjach, w wyniku których ze szczawiooctanu i acetylo-KoA powstają dwie cząsteczki dwutlenku węgla, szczawiooctan oraz 13 równoważników wysokoenergetycznego wiązania pirofosforanowego w postaci, m.in., zredukowanych koenzymów NAD, NADP i FAD. Przebieg tych reakcji zob. rys. 4.



Rys. 4. Cykl Krebsa

utleniająca fosforylacja

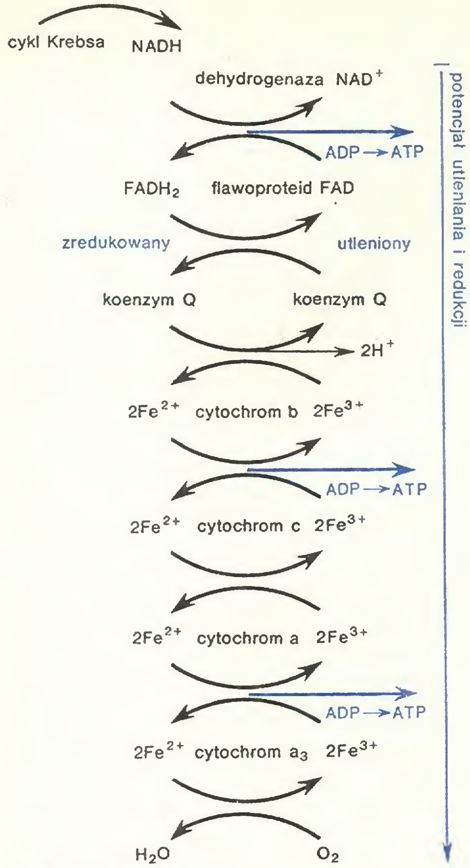
W procesie utleniającej fosforylacji zredukowane koenzymy ulegają utlenieniu i przekazują elektrony do układu utleniająco-redukującego złożonego z kilku białek (jest to schematycznie przedstawione na rys. 5). Kolejne ich przeniesienie do ogniwa o coraz to wyższym potencjale utleniania-redukcji powodują wydalanie protonów na zewnątrz błony (mitochondriów, chloroplastów lub komórek bakteryjnych). Wytwarza to potencjał, który jest wykorzystywany albo bezpośrednio do procesów transportu przez błony albo też do reakcji nieorganicznego fosforanu z ADP. Na skutek tej reakcji powstają nowe wysokoenergetyczne wiązania pirofosforanowe. Przedstawia to schematycznie rys. 6.

metabolizm energetyczny

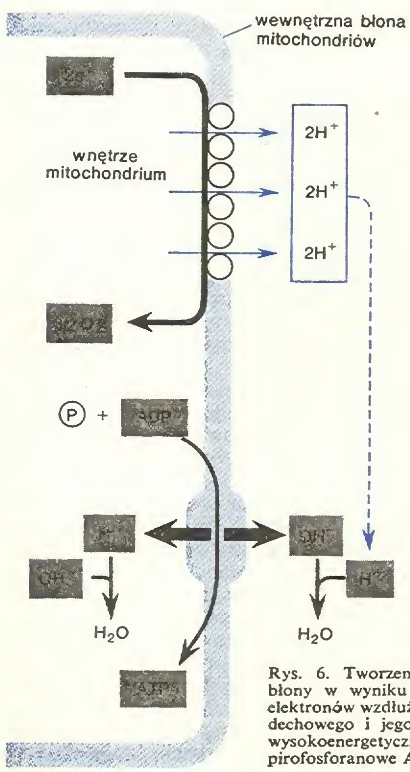
ATP jest głównym i najbardziej bezpośrednim źródłem energii w procesach metabolicznych komórki, w tym dla wszystkich procesów nieredukcyjnych. W wyniku tych reakcji ulega on rozpadowi na ADP i fosforan lub AMP i pirofosforan. W niektórych reakcjach bierze udział GTP, lecz jego wiązania wysokoenergetyczne pochodzą przeważnie z ATP.

Jak już wspomniano, w wielu procesach redukcyjnych, takich jak wiązanie azotu atmosferycznego

lub synteza kwasów tłuszczowych dawcą energii jest zredukowany koenzym NADP.



Rys. 5. Łańcuch oddechowy



Rys. 6. Tworzenie potencjału błony w wyniku przenoszenia elektronów wzdłuż łańcucha oddechowego i jego zamiana na wysokoenergetyczne wiązania pirofosforanowe ATP

Cechy metabolizmu energetycznego są prawie uniwersalne dla świata żywego. Natomiast wykorzystanie energii, a raczej cele, na które jest wydatkowana, są różne i zależą od natury komórek i warunków ich bytu. Rośliny i bakterie fotosyntetyzujące wykorzystują mineralne źródła azotu i siarki. Syntetyzują one wszystkie związki organiczne niezbędne do konstrukcji komórek, a także do pewnych reakcji na środowisko zewnętrzne. Natomiast wiele bakterii i grzybów oraz wszystkie organizmy zwierzęce wymagają do życia związków organicznych — glukozy, niektórych aminokwasów oraz witamin. Dzięki ograniczaniu liczby procesów biosyntetycznych zwierzęta zyskiwały w coraz większym stopniu mechanizmy zwiększające ich niezależność od środowiska. Tak więc, wszystkie zwierzęta zużywają wiele energii na ruch (→ Molekularne podstawy skurczu mięśnia), a stałocieplne również na utrzymanie temperatury ciała.

enzymy

Cały metabolizm komórkowy, w tym i energetyczny, jest rezultatem reakcji katalizowanych przez białka o swoistym działaniu na inne związki chemiczne, czyli enzymy (→ Białka). Wszystkie reakcje chemiczne tu omawiane, czy wspominane, są katalizowane przez swoiste enzymy, nie wymieniane tylko dlatego, aby zachować jasność tekstu.

Każda jednostkowa reakcja metabolizmu jest katalizowana przez odrębny enzym. Rzadkimi odstępstwami od tej reguły są reakcje katalizowane przez kilka enzymów o podobnej funkcji katalitycznej oraz zupełnie wyjątkowe sytuacje, kiedy ten sam enzym katalizuje dwie różne reakcje.

Metabolizm komórkowy musi być koordynowany tak, by najbardziej celowo wykorzystać dostępne produkty. W przeciwnym wypadku komórka byłaby wyeliminowana w drodze ewolucji przez inne, bardziej sprawne. Niezrównoważona akumulacja niektórych pośrednich produktów metabolizmu prowadzi bowiem nie tylko do marnotrawienia energii, ale również i do pewnych efektów toksycznych. Akumulacja galaktozo-1-fosforanu np. zatrzymuje całość metabolizmu bakterii, a u dzieci z defektami galaktozoepimerazy prowadzi do ogólnego niedorozwoju, niekiedy ślepoty lub nawet śmierci.

Rezultat pracy enzymu zależy od stężenia jego substratów. Ciąg reakcji enzymatycznych, np. biosynteza jakiegoś aminokwasu, przebiega z wydajnością zdeterminowaną przez enzym o najniższej aktywności. Jednakże ewolucja wykształciła systemy regulacyjne, działające na zasadzie sprzężenia zwrotnego (inhibicja zwrotna). Ich działanie sprawia, że raczej zapotrzebowanie na produkt końcowy szlaku metabolicznego niż stężenia jego substratów decydują o sumarycznej wydajności. Otóż enzymy kluczowe dla wydajności szlaków metabolicznych, są to z reguły enzymy katalizujące pierwsze reakcje tych szlaków, mają one centra niezależne od centrów aktywnych, które wykazują powinowactwo do pewnych produktów końcowych szlaków metabolicznych, nazywanych efektorami inhibicji zwrotnej. Na przykład pierwszy enzym biosyntezy histydyny — syntetaza fosforybozylo-ATP ma oprócz centrum aktywnego, w którym reagują kosubstraty reakcji ATP i fosforybozylopirofosforan, inne centrum, do którego wykazuje powinowactwo L-histydyna. Przyłączenie L-histydyny powoduje taką zmianę konformacji enzymu, że centrum aktywne ulega przestrzennej deformacji, co prowadzi do obniżenia jego zdolności katalizowania syntezy fosforybozylo-ATP. Enzymy, a ogólniej rzecz biorąc — białka, które mogą zmienić swą konformację i aktywność w wyniku współdziałania z efektorami biologicznie korzystnymi, nazwano allosterycznymi.

inhibicja zwrotna

Odrębny mechanizm regulacyjny działa dzięki procesowi syntezy enzymów. Struktura poszczególnych enzymów jest zakodowana w DNA chromosomu każdej komórki. Odpowiedni odcinek DNA nazywa się genem struktury.

regulacja transkrypcji

W bezpośrednim sąsiedztwie genu struktury znajdują się inne geny, które określają jego ekspresję (aktywność). Taka funkcjonalna jednostka genomu

operon

bakterii, która może zawierać większą liczbę genów struktury nazywa się operonem. Geny, które decydują o ilości wytwarzanych enzymów znajdują się zawsze w jednym końcu operonu i ich nadrzędne działanie rozciąga się na cały operon.

Znane są dwa rodzaje współdziałających ze sobą

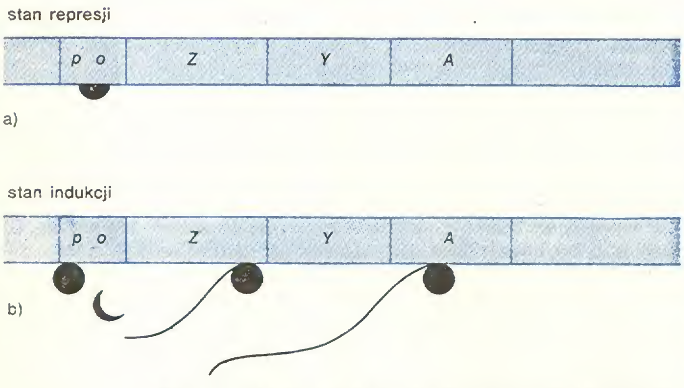
promotor i operator

nadrzędnych genów operonu. Promotor jest genem warunkującym rozpoczęcie transkrypcji genów struktury przez polimerazę RNA. Usunięcie lub deformacja struktury promotora przez mutację prowadzi do braku transkrypcji podległych genów struktury, co jest niezbędnym warunkiem w procesie syntezy odpowiednich enzymów.

Drugi z tych genów to operator, znajdujący się między promotorem a genem struktury. Przyłączenie do operatora swoistego białka o odpowiedniej konformacji allosterycznej powoduje wstrzymanie trans-

represor

krypcji. To białko, zwane represorem, jest kodowane przez gen nie należący do operonu. Jest ono produkowane stale z taką samą szybkością. Powoduje to stały naturalny brak transkrypcji danego operonu. Dopiero gdy w komórce znajdą się związki chemiczne zmieniające konformację represora i powodujące jego niezdolność do łączenia się z operatorem, może dojść do ekspresji genów struktury tego operonu. W ten sposób jest regulowana ekspresja operonu laktozy niektórych bakterii. Pojawienie się w komórkach laktozy prowadzi do powstania metabolitów, np. galaktozy, które łącząc się z represorem tego operonu zmieniają jego konformację, przez co traci on zdolność do wiązania się z operatorem. W rezultacie tego następuje ekspresja genów struktury operonu w postaci tworzonego swoistego mRNA. W układzie biosyntezy białka zachodzi proces translacji informacji zawartej w mRNA i dochodzi do wytworzenia odpowiednich enzymów. W wypadku operonu *lac* są to (rys. 7): β-galaktozydaza (gen *Z*), permeaza (gen *Y*) i transacetylaza (gen *A*). Biologiczny sens tego mechanizmu jest oczywisty; do syntezy enzymów katabolizmu laktozy dochodzi tylko wówczas, gdy laktoza znajdzie się w komórce.



Rys. 7. Schemat regulacji operonu laktozy *Escherichia coli*. Promotor i operator oznaczono literami *p* i *o*, zaś geny struktury, kodujące trzy białka operonu literami *Z*, *Y* i *A*. Aktywny represor przedstawiono w kształcie półkola, zaś represor allosterycznie zmieniony przez przyłączenie efektora ma na rysunku kształt półksiężyca. Cząsteczki polimerazy RNA przestawiono jako koła, zaś mRNA jako linie krzywe. a) Stan represji jest skutkiem przyłączenia represora do operatora. Uniemożliwia to funkcję polimerazy RNA. b) Do indukcji, czyli uaktywnienia operonu dochodzi, gdy operator jest wolny, gdyż allosterycznie zmieniony represor nie może się do niego przyłączyć. Wówczas kolejne cząsteczki polimerazy RNA przyłączają się do promotora, a następnie posuwając się wzdłuż DNA operonu syntetyzują mRNA, mające strukturę komplementarną do jednej z nici genów struktury

represja kataboliczna

Jednakże, kiedy komórka jest dostatecznie zaopatrzona w prostsze źródło węgla i energii, jakim jest glukoza, indukcja enzymów katabolizmu laktozy ulega ograniczeniu przez inny mechanizm, zwany kataboliczną represją. Działa on na ekspresję wielu układów enzymatycznych, których funkcja polega na katabolizmie różnych związków, np. aminokwasów, w produkty mogące służyć jako źródła węgla i energii.

represja glukozowa

Pośrednim efektorem represji katabolicznej jest glukoza. Jej obecność w komórce bakterii powoduje zmniejszenie wytwarzania cyklicznego adenozynomonofosforanu (cAMP) przez związaną z błonami komórkowymi cyklazę adenylową. Szczegółowy mechanizm tego wpływu nie jest znany i nie wiadomo, co jest efektorem działającym na cyklazę. Jedyną znaną funkcją cyklicznego AMP u bakterii jest tworzenie kompleksu ze swoistym białkiem, zwanym CRP. Kompleks ten ma powinowactwo do promotorów i jego związanie z DNA ułatwia wiązania się polimerazy RNA z promotorem, a dzięki temu zwiększenie transkrypcji operonu. Dlatego właśnie brak cyklicznego AMP w obecności glukozy zmniejsza transkrypcję operonów katabolicznych.

Reasumując, ekspresja operonu laktozy jest regulowana zarówno w zależności od obecności substratów dla enzymów kodowanych przez ten operon, jak i od zapotrzebowania na produkty przemiany tych substratów.

Poznano już kilka operonów bakteryjnych. Każdy z nich różni się od innych rodzajem oddziaływania białek regulacyjnych na zapoczątkowanie procesu transkrypcji przez polimerazę RNA. Od operonu laktozy najbardziej się różnią te, których ekspresja wymaga obecności swoistych białkowych aktywatorów. Ich zdolność do aktywacji procesu transkrypcji zależy, podobnie jak w wypadku represora operonu — laktozy, od konformacji allosterycznej. Uzyskują ją przez wiązanie odpowiednich niskocząsteczkowych efektorów. Różnica między nimi polega na tym, że związany represor zwiększa ekspresję swego operonu przez odłączenie się od niego, związany zaś z efektorem aktywator zwiększa ekspresję odpowiedniego operonu przez przyłączenie się do jego operatora.

Poznanie operonu laktozy uwidoczniło jeden ze sposobów, w jaki zachodzi regulacja metabolizmu w zależności od potrzeb komórki. Jednakże nawet rozszerzony model operonu stosuje się jedynie do najprostszych organizmów, tj. bakteriofagów i bakterii. Już w drożdżach regulacja ta przebiega inaczej i na razie nie jest na tyle poznana, by można ją tu przedstawić z podobną ilością niewątpliwych szczegółów.

represja azotowa

Represja glukozowa nie jest jedynym przykładem nadrzędnego mechanizmu regulacji syntezy enzymów u bakterii. Wykryto, że jon amonowy działa jako pośredni efektor represji azotowej. Reguluje ona wydajność procesów prowadzących do wytwarzania jonu amonowego z innych źródeł, jak np. azotanu lub azotu atmosferycznego. Bezpośrednim efektorem represji azotowej jest w niektórych bakteriach białko, a mianowicie enzym, syntetaza glutaminy.

Wpływ regulacyjny na syntezę niektórych enzymów może wywierać skład gazowy atmosfery. Tlen hamuje w nieznany sposób syntezę enzymów redukcji azotanu i azotu atmosferycznego, a indukuje syntezę enzymów utleniająco-redukujących, która też, z drugiej strony, podlega katabolicznej represji. Natomiast efekt Pasteura, czyli zahamowanie glikolizy przez tlen nie zachodzi wskutek wpływu tlenu na syntezę enzymów glikolizy.

regulacja metabolizmu u organizmów wyższych

W komórkach organizmów wyższych, szczególnie zwierząt, procesy regulacyjne są bardziej złożone. Przyczyną tego jest chyba fakt, że komórki tych organizmów mają mniejszą samodzielność i przebieg ich metabolizmu jest kontrolowany z punktu widzenia interesu całego organizmu. W związku z tym komórki zwierząt są ze sobą powiązane systemem informacji metabolicznej i czynnościowej, tzn. za pomocą systemów, których elementami są hormony (→ Białka). Molekularny mechanizm ich działania jest bardzo słabo poznany. Receptory poszczególnych hormonów znajdują się tylko w tych komórkach, które mają na

nie reagować. Receptory hormonów sterydowych mogą występować wewnątrz komórek, natomiast receptory innych hormonów, np. insuliny czy adrenaliny występują tylko w błonach komórkowych. Ich stymulacja przez różne hormony zwiększa aktywność cyklazy adenylowej i ilość produkowanego przez nią cyklicznego AMP. W komórkach zwierzęcych cykliczny AMP jest aktywatorem kinaz białkowych, czyli enzymów przyłączających grupy fosforanowe do białka. Przypuszcza się, że powstające w wyniku ich działania pochodne są dalszym ogniwem łańcucha informacji międzykomórkowej i one, w nieznany jeszcze sposób, kierują przebiegiem odpowiednich reakcji metabolicznych. Tych kilka informacji o regulacji metabolizmu w komórkach zwierzęcych nie stanowi obrazu tej regulacji, a raczej ilustruje jej złożoność i niepełność aktualnego stanu wiedzy.

Wiadomo niewątpliwie, że dwa podstawowe mechanizmy regulacyjne, a mianowicie inhibicja zwrotna oraz indukcja enzymatyczna zachodzą w komórkach zwierzęcych. Stwierdzono np. że aktywność pierwszego enzymu syntezy puryn wątroby gołębia podlega inhibicji przez niektóre nukleozydotrójfosforany. Synteza enzymu wątroby szczura przyłączającego grupę hydroksylową do policyklicznych węglowodorów może być wywoływana przez substraty tych enzymów, ale również przez niektóre hormony.

Jednym z mechanizmów występujących zarówno u bakterii, jak i u organizmów wyższych jest zależność syntezy rybosomalnego RNA od biosyntezy białka. Brak choćby jednego aminokwasu u bakterii powoduje zatrzymanie translacji. Wskutek tego dochodzi do przemiany GTP w pochodną z grupą pirofosforanową — ppGpp. Związek ten zwiększa transkrypcję szeregu operonów kodujących enzymy procesów biosyntetyzujących aminokwasy, a zmniejsza transkrypcję genów kodujących rybosomalny RNA. Wykazano, że w fibroblastach, niedojrzałych komórkach zwierzęcych, zachodzi podobna zależność syntezy rybosomalnego RNA od sprawnego przebiegu procesu translacji.

Uogólniając sprawę mechanizmów regulacyjnych u zwierząt, można stwierdzić, że występują u nich mechanizmy podobne do tych, które poznano u bakterii i innych prostych makroorganizmów, jak również działają jeszcze inne mechanizmy, koordynujące metabolizm wielu komórek w celu zapewnienia sprawności całego organizmu.

Mechanizmy zachowania gatunków i ich ewolucji

podział komórki

Procesy metabolizmu umożliwiają komórkom uzyskanie energii i syntezę składników niezbędnych do ich wzrostu. Jego kulminacją jest podział komórki, w którego wyniku powstają dwie komórki potomne. Warunkiem żywotności tych komórek jest otrzymanie informacji genetycznej, determinującej całość cyklu życiowego. Zapewnia to zachodząca uprzednio replikacja materiału genetycznego w komórce macierzystej.

Proces ten ma zasadniczo inny przebieg u organizmów protokariotycznych, posiadających pojedynczy kolisty chromosom, niż u wyższych, eukariotycznych organizmów, mających liczne chromosomy, zawierające liniowe DNA związane z białkami i to zamknięte w jądrze, które ma swoje błony i które ma również swój cykl podziałowy (→ Kwasy nukleinowe).

reperacja uszkodzeń DNA

Proces replikacji przebiegający w sposób idealny zapewniałby powstanie dwu identycznych komórek. Jednakże zasady wchodzące w skład DNA podlegają wielu reakcjom fizycznym lub w pewnych warunkach nawet chemicznym, co powoduje zmiany zdolności komplementacji zasad drugiego łańcucha. Te komplementacyjne niezgodności są rozpoznawane przez swoiste enzymy, które również dokonują korekt, np. przez wycięcie zmienionych zasad. Powstałą lukę wypełnia z zachowaniem zasady komplementarności inny enzym. Komórki dysponują kilkoma niezależnymi systemami naprawy zmian w DNA, lecz mimo to nie wszystkie zmiany w DNA, nazywane mutacjami, mogą być naprawione. Prowadzi to do odmienności odpowiednich elementów informacji genetycznej w komórkach potomnych.

mutacje

Mutacje są przypadkowymi zmianami w budowie pojedynczych kodonów lub całych ich serii. Zwykle odbija się to niekorzystnie na zdolności do metabolizmu lub innych bardziej złożonych funkcji komórek. Na przykład zmiana w kodonie aminokwasu ważnego dla centrum aktywnego jakiegoś enzymu biosyntezy biotyny, koenzymu szeregu enzymów przyłączających węglan do innych związków, powoduje, że enzym ten jest nieaktywny, przez co komórka jest niezdolna do produkcji biotyny i jej żywotność zależy od możliwości uzyskania tej biotyny z zewnątrz. Inne geny — kodujące inne enzymy — biosyntezy biotyny — mogą ulegać mutacjom, które nie umniejszą już zdolności życiowych komórek, gdyż enzymy te i tak są bezużyteczne. Długi ciąg mutacji może przekształcić produkty tych genów w inne enzymy. Ich użyteczność zapewni komórkom pewną przewagę w walce o byt. Ciągłe występowanie mutacji może na przestrzeni tysięcy pokoleń doprowadzić do powstania nowych odmian gatunków, a nawet zupełnie nowych gatunków.

Ważnym czynnikiem w procesie ewolucji są procesy przekazywania materiału genetycznego poszczególnym osobnikom.

transformacja

Bakterie mają wiele mechanizmów przekazywania materiału genetycznego. Transformacja polega na przenikaniu z roztworu DNA pochodzącego z rozpadłych bakterii. Ulega on rekombinowaniu, jeżeli wykazuje dostateczny stopień homologii z DNA komórki, w której się znalazł.

transdukcja

Inny proces, nazwany transdukcją polega na tym, że niektóre bakteriofagi przenoszą DNA bakterii, w których się namnożyły, do bakterii, w których się znalazły. Jeszcze inny proces, zwany koniugacją, wymaga obecności plazmidów.

koniugacja

Są to koliste cząsteczki DNA zdolne do samodzielnej replikacji wewnątrz bakterii oraz do włączania się do chromosomu bakterii. Nadają one komórkom cechy, określane jako męskie, gdyż umożliwiają przenoszenie DNA do komórek nie zawierających tego plazmidu.

modyfikacja i restrykcja

We wszystkich trzech wypadkach los przeniesionego DNA zależy od tego, czy nie ulegnie on zniszczeniu przez nukleazy. Wszystkie bakterie mają enzymy metylujące niektóre sekwencje DNA oraz endonukleazy rozpoznające, czy sekwencje te są odpowiednio zmodyfikowane przez metylację. Niezmetylowane sekwencje są przecinane, co umożliwia działalność egzonukleaz i hydrolizę fragmentu DNA, rozpoznanego jako obcy materiał genetyczny. Zjawisko to nosi nazwę restrykcji. Utrudnia ona wymianę informacji genetycznej między różnymi gatunkami bakterii, a nawet różnymi liniami tego samego gatunku bakterii.

rekombinacja

Przeniesiony materiał genetyczny u bakterii ulega procesowi nazwanemu rekombinacją. Ten wieloetapowy proces polega na tym, że homologiczne, czyli zasadniczo podobne fragmenty dwuniciowego DNA rozpoznają się w sposób, który do dzisiaj jest jedną z największych zagadek biologii molekularnej. Rozpoznanie to umożliwia równoległe ułożenie homologicznych odcinków. Następnie dochodzi do wzajemnej wymiany odcinków o przypadkowej, prawdopodobnie, długości. Enzymy biorące udział w rekombinacji są również dotąd nieznane.

Wszystkie trzy procesy — przenoszenia, restrykcji i homologicznej rekombinacji — ograniczają wymianę informacji genetycznej do osobników blisko spokrewnionych. Powoduje to, że tylko mutacje akumulowane w poszczególnych liniach genealogicznych da-

nego gatunku mogą się sumować. Powstające rekombinanty podlegają naturalnej selekcji, na skutek czego przeżywają odmiany najlepiej dostosowane do warunków otoczenia.

W ostatnich latach wykryto, że wiele bakterii i ich plazmidów ma szczególne odcinki DNA, nazwane sekwencjami insercyjnymi lub czynnikami IS. Ich właściwością jest zdolność do przemieszczania się z jednego miejsca w chromosomie na drugie, znajdujące się albo w tym samym chromosomie lub w innej cząsteczce DNA tej samej komórki. Co więcej, przemieszczeniom tym podlegają pary identycznych czynników IS (a poznano już kilka ich rodzajów), wraz z odcinkiem DNA, zawartym między nimi. Te tzw. transpozony są jakościowo różnym czynnikiem ewolucyjnym. Ich przenoszenie nie odbywa się bowiem na zasadzie homologicznej wymiany, lecz włączenia w dowolne miejsce zupełnie nowego i dodatkowego materiału genetycznego.

transpozony

zmienność u eukariontów

Nie ma dotąd dowodów, by transformacja czy transdukcja, mogły zachodzić u organizmów eukariotycznych. Niemniej jednak, międzyosobniczy a wewnątrzgatunkowy proces tworzenia nowych kombinacji informacji genetycznej zachodzi dzięki zróżnicowaniu płciowemu. Komórki wyższych organizmów mają podwójne komplety chromosomów. Jedynie komórki płciowe: plemniki i jaja, mają pojedyncze komplety. Nowy organizm, powstający w wyniku zapłodnienia jaja, ma po jednym chromosomie z każdej pary od jednego z rodziców. Stanowi to zupełnie inną kombinację cech odróżniającą go od rodziców. Ponieważ każdy plemnik czy jajo jest inną praktycznie kombinacją chromosomów odziedziczonych po rodzicach rodziców, a ogólna ich liczba u człowieka wynosi 23 pary, zróżnicowanie potomstwa tych samych rodziców jest bardzo duże. Powiększa je fakt, że podczas mejozy, czyli redukcyjnego podziału jądra w czasie wytwarzania komórek płciowych, dochodzi do rekombinacji między homologicznymi chromosomami.

Przytoczone tu dane dowodzą, że zróżnicowanie płciowe umożliwia organizmom eukariotycznym ogromną zmienność osobniczą. Częstość mutacji u organizmów wyższych jest równie duża, jak u bakterii. Powstają więc podobne warunki do procesów ewolucyjnych przez selekcję osobników lepiej dostosowanych do środowiska. Jednakże w skali czasu są one powolniejsze ze względu na długość życia pokolenia i rozmiary populacji organizmów wyższych. Ewolucję opóźnia również ich większa niezależność od zmienności środowiska. Potrzeba ewolucji organizmów wyższych, szczególnie człowieka, jest zagadnieniem wykraczającym poza biologię.

Błony komórkowe

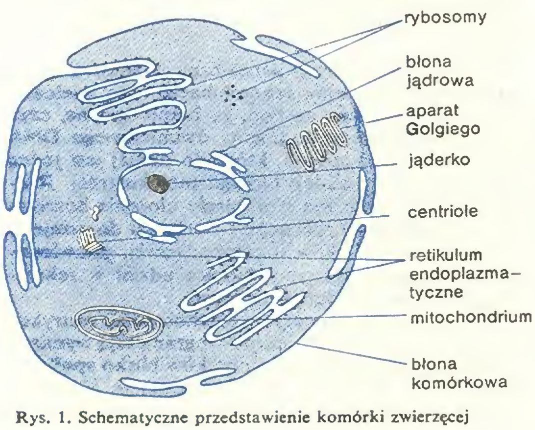
Stanisław Przestalski

Występowanie błon i ich znaczenie

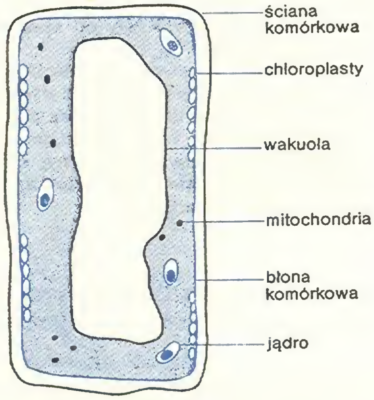
Błona, albo membrana, jest to warstwa substancji oddzielająca dwa ośrodki od siebie. Błona różni się od obydwu ośrodków cechami chemicznymi, strukturalnymi i funkcjonalnymi. Na ogół błona jest cienka w porównaniu z ośrodkami, które rozdziela.

błony komórkowe (plazmatyczne)

Wśród różnorodnych błon, z jakimi spotykamy się w przyrodzie nieożywionej i ożywionej, na szczególną uwagę zasługują błony komórkowe. Błony te, zwane również błonami plazmatycznymi, otaczają żywe komórki wszystkich rodzajów: komórki zwierzęce, roślinne i bakteryjne. Komórki zwierzęce otoczone są jedynie błonami plazmatycznymi (rys. 1). Natomiast komórki roślinne i bakteryjne, niezależnie od błony plazmatycznej, pokryte są dodatkowo osłoną, zwaną ścianą komórkową (rys. 2); ściana komórkowa odgrywa przede wszystkim rolę osłony mechanicznej i nie będzie dalej rozpatrywana. Błona plazmatyczna oddziela wnętrze komórki (protoplazmę) od ośrodka zewnętrznego i umożliwia zachowanie jej indywidualności zapobiegając wymieszaniu zawartości komórki z ośrodkiem. Błona umożliwia równocześnie konatakt ze światem zewnętrznym dzięki swojej szczególnej właściwości polegającej na zdolności do wybiórczego przepuszczania substancji.



Rys. 1. Schematyczne przedstawienie komórki zwierzęcej



Rys. 2. Schematyczne przedstawienie komórki roślinnej

Grubość błony komórkowej jest rzędu 10 nm, podczas gdy liniowe rozmiary większości żywych komórek zawierają się w granicach od $0{,}5 \cdot 10^3$ nm do 10^4 nm, a więc są od dwóch do trzech rzędów wielkości większe od grubości błony.

wprowadzenie pojęcia błony

Pojęcie błony komórkowej zostało zaproponowane przez C. Nagāliego w 1855 r. na podstawie doświadczeń nad przenikaniem do wnętrza komórek pewnych barwników. Od tego czasu przez wiele lat ścierały się poglądy zwolenników pojęcia „błona komórkowa" z poglądami jego przeciwników. Ci ostatni utrzymywali, że błona komórkowa nie istnieje, a powierzchnię komórki charakteryzuje jedynie zagęszczona warstwa protoplazmy. Obecnie istnienie błony komórkowej nie budzi na ogół wątpliwości. Przyjmuje się, że błona plazmatyczna jest tworem o ściśle określonej strukturze i o charakterystycznych właściwościach. Za istnieniem błony komórkowej przemawia wiele argumentów. A więc warstwy powierzchniowe żywych komórek zawierają przeważającą (do 80%) część suchej masy komórek. Przekłu-

argumenty za istnieniem błony

cie powierzchni komórki może spowodować wypływ protoplazmy na zewnątrz. Po wypłynięciu zawartości komórki pozostaje często otoczka utożsamiana przeważnie z błoną (chociaż najprawdopodobniej to nie jest sama błona). Można również izolować błonę plazmatyczną uszkadzając komórkę innymi sposobami. Umieszczając np. komórkę w ośrodku wodnym o niższym ciśnieniu osmotycznym od ciśnienia osmotycznego protoplazmy doprowadzamy do przepływu wody do wnętrza komórki, wzrostu jej objętości i pęknięcia. Tego rodzaju postępowanie stosuje się przede wszystkim w stosunku do czerwonych ciałek krwi (erytrocytów); zjawisko pękania krwinek pod wpływem różnicy ciśnień osmotycznych i wydalania zawartości tych komórek na zewnątrz nosi nazwę hemolizy. Po hemolizie w roztworze pozostają blade otoczki, zw. cieniami. Przyjmuje się, że cień zawiera błonę lub jest po prostu błoną czerwonej krwinki. Do wyodrębnienia błon komórkowych innych rodzajów komórek stosuje się inne metody. A więc komórki bywają np. rozrywane przy pomocy ultradźwięków lub mechanicznych homogenizatorów. Błony lub fragmenty błon oddziela się od pozostałych składników komórki drogą flotacji (czyli wypłukiwania), wirowania lub ekstrahowania w roztworach buforowych.

dowód — obserwacja w mikroskopie elektronowym

Wątpliwości dotyczące istnienia *błony* pochodziły jednak głównie stąd, że nie można było zobaczyć przekroju błony. Grubość błony jest mniejsza od zdolności rozdzielczej mikroskopu optycznego i dopiero zastosowanie mikroskopu elektronowego pozwoliło „zobaczyć" błonę, ale nie w stanie żywym. Technika mikroskopii elektronowej wymaga bowiem wielu drastycznych zabiegów, które prowadzą co najmniej do zniszczenia właściwości życiowych preparowanych komórek. Przygotowanie próbki polega na utrwaleniu komórki lub jej elementów (np. cieni) przy pomocy m.in. czterotlenku osmu, odwodnieniu, wycięciu ultracienkich skrawków (za pomocą mikrotomu) nie grubszych niż kilkadziesiąt nm i zabarwieniu solami ciężkich metali, np. octanem uranylowym. Pod mikroskopem elektronowym przekrój komórki przedstawia się tak, jak ukazano na il. 128 (tabl. 32). Na zdjęciu tym można wyróżnić warstwę powierzchniową o charakterystycznej strukturze. Struktura ta jest przedstawiona w sposób schematyczny na rys. 3. Widzimy tu dwie gęste warstwy przedzielone warstwą słabiej rozpraszającą elektrony. Sumaryczna grubość tych warstw wynosi od 7 nm do 10 nm. Ta trójwarstwowa struktura jest traktowana jako obraz błony komórkowej. Interpretacja mikrofotografii elektronowej nie jest jednak sprawą prostą i do chwili obecnej nie można na jej podstawie określić szczegółów budowy błony.



Rys. 3. Schematyczny obraz przekroju powierzchniowej warstwy (błony) komórki

błony wewnątrzkomórkowe

Błona komórkowa nie jest jedyną błoną spotykaną w układach biologicznych. Różnorodne organelle komórkowe, jak np. jądro, mitochondria, lizosomy (a w komórkach roślinnych chloroplasty i wodniczki), są zaopatrzone we własne błony. W cytoplazmie komórek roślinnych i zwierzęcych występuje ponadto siateczka śródplazmatyczna (retikulum endoplazmatyczne), która jest złożoną siecią błon tworzących wewnątrz komórki obszary o różnych kształtach, np. wąskie kanały, rurki, okrągłe pęcherzyki, cysterny, i która łączy się z błoną jądra i z błoną komórkową (rys. 1).

błony tkankowe

Poza błonami komórkowymi i błonami wewnętrznymi komórek znamy liczne błony tkankowe złożone z żywych komórek. Należą do nich m.in. przepona jelitowa, błona nerki, skóra żaby i wiele innych.

Układy żywe można więc traktować jako układy błon, tworzące zlokalizowane obszary, w których zachodzą różnorodne procesy życiowe. Podstawową i charakterystyczną funkcją błon jest regulacja transportu substancji dzięki zdolności do selektywnego przepuszczania cząstek. Wszystkie komórki nieustannie wymieniają różnorodne substancje z otoczeniem. Wymiana ta jest specyficzna, ponieważ tylko niektóre związki wnikają i tylko niektóre wydostają się na zewnątrz. Mechanizm tej wymiany jest związany, jak się uważa, całkowicie lub niemal całkowicie z właściwościami błon. Zdolność błon do kierowania procesami przenikania prowadzi do wielu konsekwencji, np. transport jonów przez błony jest źródłem powstawania napięć elektrycznych między wnętrzem komórki i otoczeniem, co umożliwia powstawanie i rozchodzenie się impulsów nerwowych w komórkach nerwowych. Właściwości transportowe błon mitochondriów umożliwiają uzyskiwanie energii z wiązań chemicznych substancji pokarmowych przez ich utlenianie w procesie oddychania. Błony wewnętrzne chloroplastów odgrywają ważną rolę w uzyskiwaniu energii w procesie fotosyntezy a retikulum endoplazmatyczne — w syntezie białek, kwasów tłuszczowych i fosfolipidów.

regulacja transportu przez błony

inne funkcje błon biologicznych

Oprócz wymienionych tu funkcji błon, znane są jeszcze inne formy ich aktywności. Ogólnie, błony biologiczne odgrywają bezpośrednio lub pośrednio zasadniczą rolę w życiowych procesach organizmów żywych. Na ogół uważa się, że błony komórkowe stanowią równie niezbędny składnik żywych komórek jak kwasy nukleinowe. Jeśli np. usuniemy przy pomocy mikromanipulatora jądro komórki ameby, uszkadzając nieznacznie błonę, tak że zrośnie się ona szybko i cytoplazma nie wydostanie się na zewnątrz, a po kilku dniach wprowadzimy do niej jądro pochodzące z innej ameby, odzyskuje ona normalną zdolność do podziałów i ruchów. Natomiast wszelkie trwałe uszkodzenia błony komórkowej, tzn. takie, które się nie regenerują, prowadzą do śmierci komórki.

Błona komórkowa nie jest układem statycznym, lecz dynamicznym. Zmienia ona stale swoje właściwości pod wpływem różnorodnych czynników, a w szczególności pod wpływem zmian temperatury, ciśnienia i składu chemicznego ośrodka, w którym się znajduje.

badania biofizyczne i biochemiczne

Powszechne występowanie błon oraz ich znaczenie w procesach życiowych jest powodem szerokiego zainteresowania strukturą i funkcjami błon we współczesnych badaniach biologicznych, biochemicznych i biofizycznych. Szczególnie istotna jest rola badań biofizycznych, ponieważ podstawowymi procesami badanymi w membranologii są procesy transportu substancji, energii i informacji poprzez błony biologiczne, czyli typowe procesy fizyczne, które mogą być badane jedynie metodami fizycznymi i opisywane przy pomocy praw fizycznych. Podobnie inne badania z tego zakresu, jak np. badania struktury przestrzennej błon, ich właściwości elektrycznych itd. mają typowy charakter fizyczny i wobec tego należą do kręgu problemów biofizycznych. Badania biochemiczne błon są równie rozległe i dotyczą budowy chemicznej błon, reakcji chemicznych zachodzących w błonach, aktywnego transportu, zjawisk enzymatycznych i innych. Badania biofizyczne i biochemiczne w dużym stopniu uzupełniają się, a na poziomie molekularnym (np. w zagadnieniu oddziaływań między cząsteczkami błony) wzajemnie się przenikają. Biofizyka i biochemia umożliwiają nowoczesny rozwój biologii w ogóle, a biologii błon komórkowych w szczególności.

Lipidy i białka membranowe

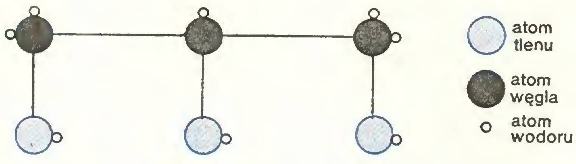
Badania chemiczne błon izolowanych, zarówno błon komórkowych jak i błon organelli wewnątrzkomórkowych, wskazują na dominujący udział w składzie chemicznym błon dwóch typów substancji: lipidów (tłuszczowców) i białek. Z tego względu rozpatrzymy budowę i właściwości cząsteczek lipidowych, cząsteczek białkowych oraz właściwości układów lipid–woda i lipid–białko.

lipidy

Lipidy są to różnorodne substancje, które charakte-

ryzuje mała rozpuszczalność w wodzie. W ściśle chemicznym sensie są to związki kwasów tłuszczowych, przeważnie estry tych kwasów i jedno- lub wielowodorotlenowych alkoholi. Lipidy dzielą się na lipidy proste i złożone. Pierwsze są wyłącznie estrami, drugie zaś zawierają ponadto jeszcze inne związki organiczne. W błonach występują przede wszystkim różne rodzaje lipidów złożonych. Budowę lipidów

fosfolipidy złożonych omówimy na przykładzie jednej z grup tych lipidów, a mianowicie na przykładzie fosfolipi-



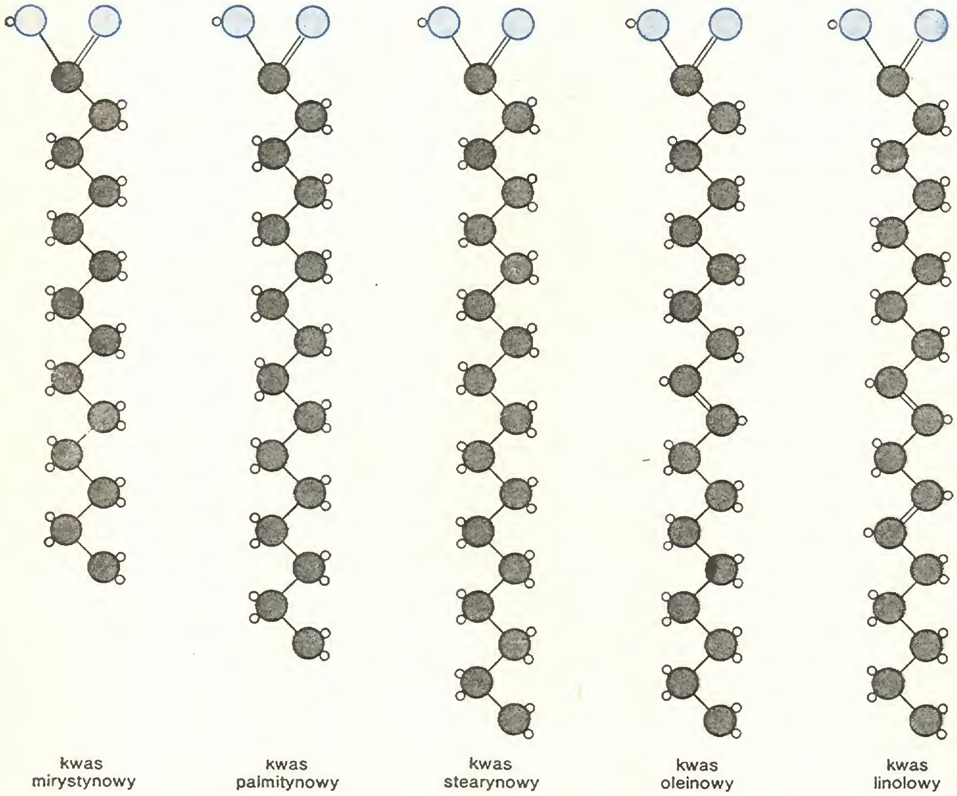
Rys. 4. Cząsteczka glicerolu

rolu są zastąpione długołańcuchowymi resztami kwasów tłuszczowych, np.:
kwasu mirystynowego $—CH_3(CH_2)_{12}COOH$,
palmitynowego $—CH_3(CH_2)_{14}COOH$,
stearynowego $—CH_3(CH_2)_{16}COOH$,
oleinowego $—CH_3(CH_2)_7=CH(CH_2)_7COOH$,
linolowego

$—CH_3(CH_2)_4CH=CHCH_2CH=CH(CH_2)_7COOH$

(rys. 5) i innych. W rozważanych lipidach najczęściej spotykamy łańcuchy węglowodorowe o parzystej liczbie atomów węgla, zawierające na ogół 12 do 18 atomów węgla. Nienasycone (tzn. zawierające podwójne wiązania) łańcuchy węglowodorowe występują powszechnie, lecz szczególnie często podwójne wiązania spotykamy w kwasach tłuszczowych o 18 atomach węgla (z jednym, dwoma lub trzema wiązaniami podwójnymi).

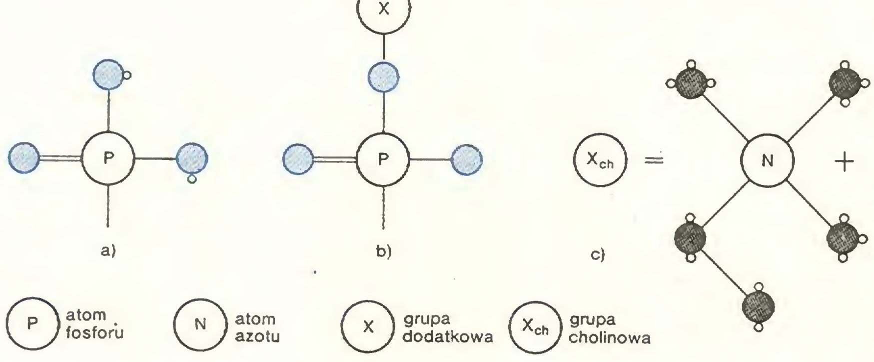
Trzecia grupa hydroksylowa cząsteczki glicerolu jest zwykle (w omawianym przypadku) zastępowana



Rys. 5. Przykłady kwasów tłuszczowych; oznaczenia atomów są takie same jak na rys. 4

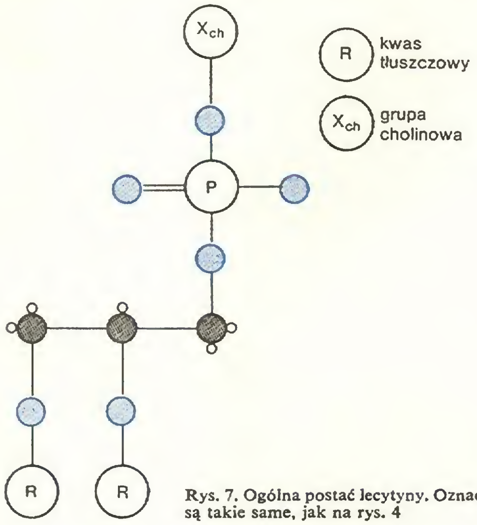
dów, które w poważnym stopniu wpływają na właściwości błon biologicznych. Lipidy te są pochodnymi cząsteczki glicerolu (rys. 4). W fosfolipidach zwykle dwie grupy hydroksylowe (—OH) cząsteczki glice-

przez grupę polarną (na rys. 6a jest nią reszta fosforanowa) lub kombinacją reszty fosforanowej i pewnej grupy dodatkowej (tak jak to jest przedstawione na rys. 6b).



Rys. 6. Grupy zastępujące grupę hydroksylową cząsteczki glicerolu: a) grupa fosforanowa, b) kombinacja grupy fosforanowej i pewnej grupy dodatkowej, c) grupa cholinowa, + dodatni ładunek elektryczny

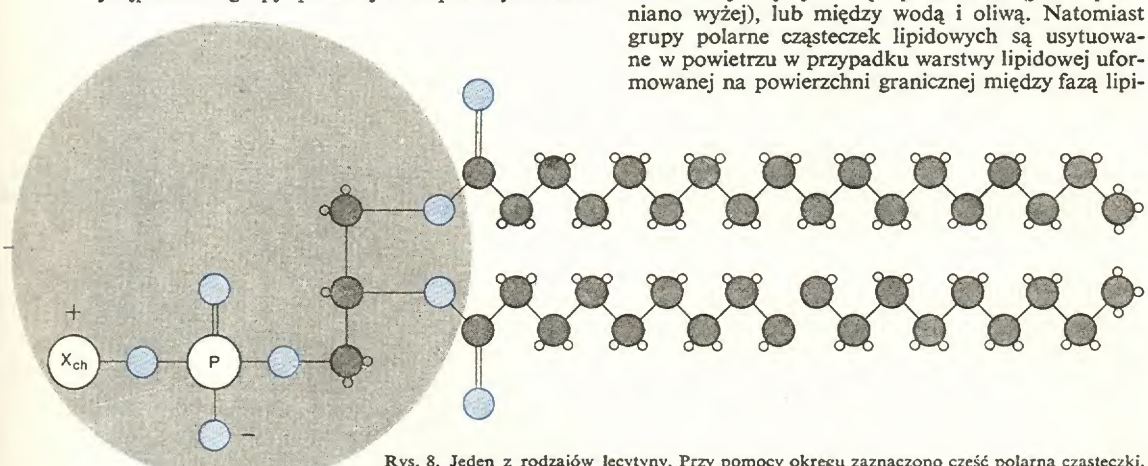
Fosfolipidy dzielą się na kilka klas. Każda klasa zawiera część złożoną z łańcuchów węglowodorowych (które mogą różnić się długością i mogą mieć różne ilości i różne położenia podwójnych wiązań) i część polarną (o charakterze anionu, kationu lub jonu obojnaczego). W jednej z klas fosfolipidów, w lecytynie, grupa fosforanowa związana jest z grupą choli-



Rys. 7. Ogólna postać lecytyny. Oznaczenia są takie same, jak na rys. 4

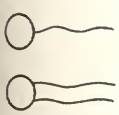
nową (rys. 6c). Wobec tego cząsteczkę lecytyny można ogólnie przedstawić tak, jak na rys. 7. Cząsteczki lecytyny różnych rodzajów zawierają identyczne grupy polarne (w tym przypadku obojnacze), lecz różnią się charakterem łańcuchów węglowodorowych. Jeden z rodzajów lecytyny jest przedstawiony na rys. 9.

Jak widać, ogólną właściwością fosfolipidów jest występowanie grupy polarnej i niepolarnych łańcu-



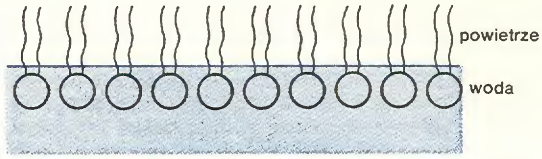
Rys. 8. Jeden z rodzajów lecytyny. Przy pomocy okręgu zaznaczono część polarną cząsteczki. Pozostałe oznaczenia są takie same, jak na rys. 4

chów węglowodorowych. W fosfolipidach z reguły spotykamy dwa łańcuchy kwasów tłuszczowych, a w innych rodzajach lipidów możemy się spotkać z jednym lub trzema łańcuchami węglowodorowymi. Schematycznie przedstawiamy cząsteczki lipidów na ogół tak, jak na rys. 8.



Rys. 9. Schematyczny sposób przedstawienia cząsteczek lipidowych

Fosfolipidy (i inne substancje o podobnej strukturze) nazywają się substancjami amfifilnymi ze względu na dualistyczny charakter oddziaływania cząsteczek lipidowych z wodą. Grupa polarna cząsteczki lipidu ma charakter hydrofilowy, a łańcuchy węglowodorowe mają właściwości hydrofobowe. Oddziaływanie elektrostatyczne między grupą hydrofilową cząsteczki fosfolipidu i cząsteczkami wody prowadzi do wniknięcia tej części lipidu do wody. Łańcuchy węglowodorowe jako hydrofobowe wystają na zewnątrz powierzchni wody. Kropla fosfolipidu rozlana na powierzchni wody rozpływa się na niej, ponieważ grupa polarna każdej cząsteczki lipidowej jest przyciągana przez wodę. W efekcie na powierzchni wody powstaje w granicznym przypadku — jeśli tylko po-



Rys. 10. Warstwa monomolekularna cząsteczek lipidowych na granicy faz woda–powietrze

lipidowa warstwa monomolekularna

wierzchnia wody jest dostatecznie rozległa — lipidowa warstwa monomolekularna. Cząsteczki fosfolipidów w tej warstwie są zorientowane prostopadle do jej powierzchni (rys. 10). Warto wspomnieć, że monomolekularne warstwy były po raz pierwszy badane przez B. Franklina, który w 1765 r. obliczył grubość warstwy oliwy rozlanej na powierzchni wody w stawie i określił ją na 10^{-7} cala, czyli 2,5 nm. J. W. Rayleigh w 1890 r. pierwszy zasugerował, że warstwy tego rodzaju mają charakter monomolekularny, a J. Langmuir w 1917 r. zwrócił uwagę na specyficzną orientację cząsteczek lipidowych w warstwie powierzchniowej. Tak więc pomiędzy fazą wodną i powietrzem fosfolipidy mogą tworzyć monomolekularne błony o specyficznych właściwościach. Ogólniej mówiąc, między dwiema fazami o różnych przenikalnościach elektrycznych powstaje warstwa lipidowa, przy czym grupy polarne tych cząsteczek są zanurzone w ośrodku o wyższej przenikalności elektrycznej, a łańcuchy hydrofobowe — w ośrodku o przenikalności niższej. Łańcuchy węglowodorowe wystają zatem ponad powierzchnię wody w przypadku warstwy lipidowej utworzonej między wodą i powietrzem (jak wspomniano wyżej), lub między wodą i oliwą. Natomiast grupy polarne cząsteczek lipidowych są usytuowane w powietrzu w przypadku warstwy lipidowej uformowanej na powierzchni granicznej między fazą lipidową i powietrzem (jak to jest np. w bańkach mydlanych). Należy zauważyć, że właśnie cząsteczki lipidów zawierających grupę polarną i dwa łańcuchy węglowodorowe (jak np. lecytyna) posiadają szczególną zdolność tworzenia warstw monomolekularnych na granicy faz woda–powietrze. Cząsteczki lipidów zawierające grupę polarną i jeden łańcuch węglowodorowy mają znaczną zdolność rozpuszczania się w wodzie i tworzenia różnorodnych układów przestrzennych w wodzie, ponieważ wpływ hydrofilowej grupy polarnej przeważa nad wpływem hydrofobowych łańcuchów węglowodorowych. Natomiast cząsteczki lipidowe o trzech łańcuchach węglowodorowych w ogóle nie rozpuszczają się w wodzie.

zawartość skład białek w błonach

W błonach biologicznych stosunek zawartości lipidów do zawartości białek zależy przede wszystkim od rodzaju błony. Np. w błonie komórkowej czerwonych ciałek krwi ludzkiej stosunek ten niewiele się różni od jedności, a w tzw. błonie mielinowej komórek nerwowych jest wyjątkowo duży i wynosi dziewięć do jednego. W wielu błonach innego rodzaju przeważa z kolei zawartość białek nad zawartością lipidów.

Ogólne właściwości cząsteczek białka oraz struktura tych makrodrobin zostały omówione w artykule → Białka. W błonach występują białka różnych rodzajów, jednak pełnej i ścisłej ich identyfikacji nie udało się do chwili obecnej przeprowadzić. Wiadomo, że zakres molekularnych rozmiarów białek w błonach, ich skład i rodzaj konformacji nie wykazują różnic w stosunku do białek pochodzenia niemembranowego. Natomiast każda błona ma inny, charakterystyczny dla siebie skład białkowy. Większość białek pochodzenia membranowego należy do białek nierozpuszczalnych w wodzie. Białka i lipidy mogą więc tworzyć trwałe układy w środowisku wodnym. Wiązaniami występującymi w tych układach są wiązania jonowe i hydrofobowe, ale szczegóły oddziaływań nie są jeszcze dokładnie poznane. Trudności w identyfikacji białek membranowych pochodzą stąd, że zabiegi chemiczne mające na celu wyodrębnienie białek z błon naturalnych prowadzą najczęściej do rozpadu tych białek na aminokwasy. Białka występujące w błonach grupuje się na ogół w dwóch klasach: w klasie białek strukturalnych (odgrywających rolę w budowie błony biologicznej) i białek funkcjonalnych (które biorą udział w wielu procesach metabolicznych, m.in., jako enzymy).

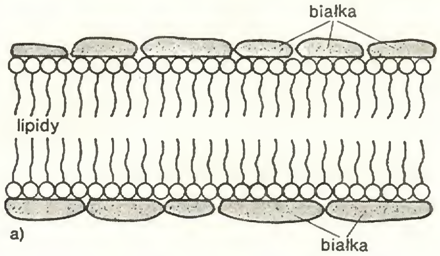
Modele błon komórkowych

Doświadczenia Gortera i Grendela (1925 r.) polegające na wyekstrahowaniu lipidów z czerwonych ciałek krwi i utworzeniu z nich warstwy monomolekularnej na powierzchni wody wskazały, że warstwa ta ma powierzchnię w przybliżeniu dwukrotnie większą od powierzchni krwinek. Wyciągnięto stąd wniosek, że błona komórkowa czerwonych ciałek krwi (i innych komórek, jak również błona organelli komórkowych) jest zbudowana z dwumolekularnej warstwy lipidowej. Ten wniosek leży u podstaw naszych poglądów na budowę błon biologicznych. Pierwszym naukowo uzasadnionym modelem błony komórkowej był model Danielliego–Davsona (1935). Został on zbudowany na podstawie doświadczeń Gortera–Grendela i na podstawie wyników pomiarów napięcia powierzchniowego żywych błon. Pomiary napięcia powierzchniowego żywych komórek wykazały, że nie przekracza ono wartości równej $2 \cdot 10^{-3}$ Nm^{-1}. Ponieważ napięcie powierzchniowe czystych lipidów wynosi około $40 \cdot 10^{-3}$ Nm^{-1}, Danielli i Davson przyjęli, że obniżenie napięcia powierzchniowego żywych komórek w stosunku do napięcia powierzchniowego fazy lipidowej pochodzi od białek zaadsorbowanych na powierzchni lipidowej części błon biologicznych. Stąd powstał model przedstawiony na rys. 11. Błona w tym ujęciu składa się z dwumolekularnej warstwy lipidowej, w której grupy polarne skierowane są na zewnątrz, a łańcuchy węglowodorowe stanowią wnętrze warstwy; na obydwu powierzchniach tej warstwy zaadsorbowane są cząsteczki białka.

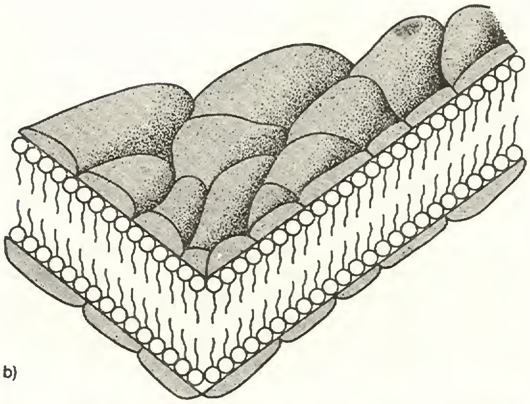
model Danielliego–Davsona

Model Danielliego–Davsona, mimo licznych zalet, nie wyjaśnia wszystkich zjawisk membranowych, w szczególności nie tłumaczy pewnych zjawisk transportowych. Dlatego z upływem lat model ten uległ stopniowym modyfikacjom. W ostatnich latach z największym uznaniem spotkał się model mozaikowy (Singer i Nicholson, Vanderkooi i Green Zahler oraz Wallach). W modelu tym zakłada się, że cząsteczki białek występują nie tylko na powierzchniach błony (jak w modelu Danielliego–Davsona), lecz również wnikają w głąb warstwy lipidowej, a nawet przenikają ją na wskroś (rys. 12). Autorzy modelu mozaikowego sugerują, że białka przenikające całą błonę są odpowiedzialne m.in. za transport pewnych substancji nierozpuszczalnych w fazie lipidowej błony.

model mozaikowy

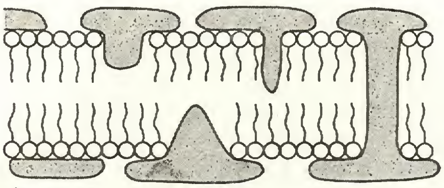


a)

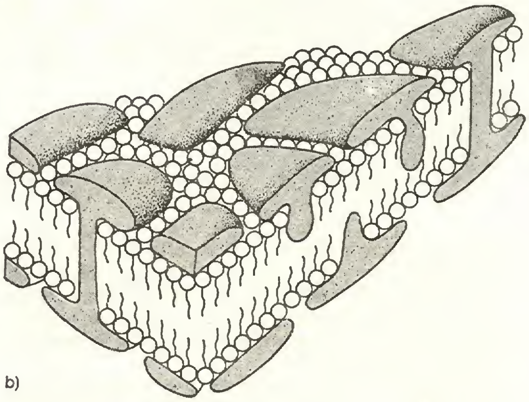


b)

Rys. 11. Model błony Danielliego–Davsona: a) płaski, b) przestrzenny



a)

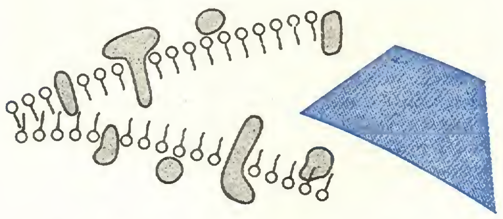


b)

Rys. 12. Model mozaikowy błony: a) płaski, b) przestrzenny

Koncepcja modelu mozaikowego jest poparta przede wszystkim doświadczeniami wykonanymi metodą tzw. wytrawiania zamrożeniowego lub sublimacyjnego (ang. *freeze etching*). W metodzie tej błona

wytrawianie zamrożeniowe błon zostaje nagle zamrożona w temperaturze ciekłego azotu, a następnie „rozłupana" tak, jak to schematycznie pokazano na rys. 13. Błonę dzieli się na dwie warstwy wzdłuż powierzchni przechodzącej przez zakończenia łańcuchów węglowodorowych w dwumolekularnej warstwie lipidowej, tam gdzie wiązania są najsłabsze. Odpowiednio przygotowane preparaty



Rys. 13. Poglądowy rysunek wskazujący, jak sobie wyobrażamy „rozłupanie" błony zawierającej dwumolekularną warstwę lipidów wraz z wbudowanymi białkami

ogląda się pod mikroskopem elektronowym. Na il. 129 (tabl. 32) widzimy otrzymaną w ten sposób mikrofotografię „rozłupanej" błony (a więc zdjęcie jednej z dwóch warstw błony) erytrocytu. Dla porównania na il. 130 (tabl. 32) możemy zobaczyć analogicznie uzyskaną mikrofotografię pojedynczej warstwy lipidowej otrzymanej z dwuwarstwowej błony lipidowej nie zawierającej białek. Na il. 129 (tabl. 32) widać różnorodne wypustki, których nie ma na il. 130 (tabl. 32). Z tego względu przyjmuje się, że wypustki na il. 129 (tabl. 32) przedstawiają ślady cząsteczek białek wnikających w głąb błony.

Obecny pogląd na budowę błony komórkowej jest wynikiem różnorodnych eksperymentów i spekulacji. Błona komórkowa jest na ogół wyobrażana zgodnie z modelem mozaikowym, ale nie jako twór statyczny, lecz dynamiczny. Zachodzą w niej ciągłe zmiany polegające m.in. na przemieszczeniach cząsteczek w błonie, na zmianach ich konformacji, na zmianach orientacji elementów składowych cząsteczek i innych. Z tego względu i ze względów innych, o których częściowo była mowa wcześniej, bezpośrednie badania żywych błon są trudne, lub niekiedy nawet, na dzisiejszym poziomie technik eksperymentalnych, niemożliwe. To jest przyczyną, że wiele badań doświadczalnych prowadzi się nie bezpośrednio na błonach biologicznych, lecz na błonach modelowych. Dzięki równoległym badaniom błon żywych i błon modelowych coraz lepiej poznajemy rzeczywistą strukturę i funkcje błon biologicznych. W badaniach obydwu rodzajów błon wykorzystuje się wiele fizycznych metod eksperymentalnych. Należą do nich metody spektroskopii w podczerwieni i ramanowskiej, rentgenografii; używa się sondy fluorescencyjne, mikroskopy elektronowe; wykorzystuje się również zjawiska rezonansu jądrowego i dichroizmu kołowego.

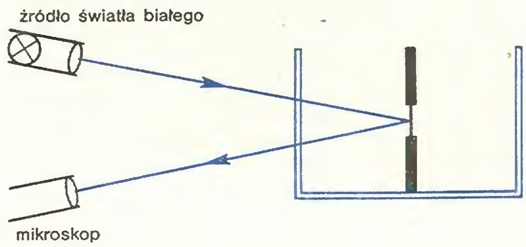
błony modelowe

Do błon modelowych, którymi najczęściej posługują się badacze należą monomolekularne warstwy lipidowe, dwumolekularne warstwy lipidowe oraz obydwa wymienione rodzaje błon modyfikowane przy pomocy białek innych substancji.

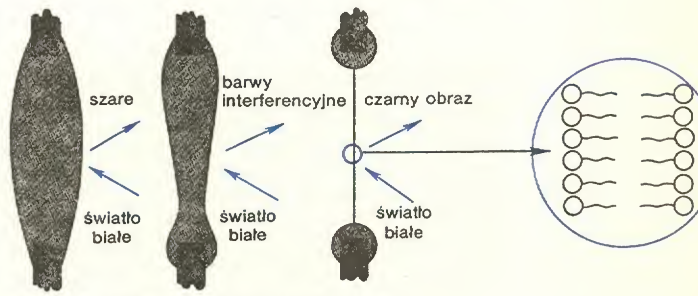
Monomolekularne warstwy lipidowe są wygodnym modelem błony, ale odbiegają znacznie swymi właściwościami od błon komórkowych, ponieważ te pierwsze stykają się z ośrodkiem wodnym tylko z jednej strony, natomiast błony komórkowe mieszczą się między dwoma ośrodkami wodnymi. Ponadto, jak wiadomo z doświadczeń Gortera i Grendela, błony komórkowe składają się z dwumolekularnej warstwy lipidowej, a nie warstwy jednomolekularnej. Tym tłumaczy się konieczność poszukiwań metody otrzymywania dwumolekularnych warstw lipidowych w warunkach sztucznych. Cel ten udało się zrealizować po długotrwałych badaniach w latach 1962–64 poprzez formowanie tzw. błon czarnych i liposomów.

błony czarne

Błony czarne formuje się w urządzeniach typu pokazanego na rys. 14. Składa się ono z dwóch koncentrycznych komór: wewnętrznej (na ogół teflonowej) zaopatrzonej w otworek o średnicy około 1 mm oraz zewnętrznej, wykonanej z materiału przezroczystego. Komory napełnia się wodnym roztworem elektrolitycznym, a następnie nakłada się na otwór (np. przy pomocy pipety) warstwę substancji lipidowej. Warstwa ta w zastosowanym roztworze zmniejsza stopniowo swoją grubość (rys. 15). Grubość tej warstwy ocenia się przy pomocy mikroskopu, do którego dochodzi odbite od warstwy lipidowej światło. Odbita wiązka światła białego przyjmuje początkowo barwę szarą, następnie, wraz ze zmniejszaniem się grubości warstwy, pojawiają się barwne obrazy interferencyjne, aż wreszcie obraz przyjmuje barwę czarną (stąd pochodzi nazwa „błona czarna"). Dzieje się tak wtedy, kiedy grubość błony osiąga wartość mniej więcej 5–6 nm; w tej sytuacji promienie odbite od obu powierzchni błony znajdują się prawie dokładnie w przeciwnych fazach. Po uzyskaniu barwy czarnej proces formowania błony zostaje zakończony. Ponieważ długość rozciągniętej cząsteczki lecytyny wynosi około 8 nm, otrzymane wartości potwierdzają przypuszczenie, że uzyskana błona czarna jest błoną dwumolekularną.



Rys. 14. Układ stosowany do formowania czarnych błon



Rys. 15. Kolejne fazy tworzenia się czarnej błony

Stopniowe zmniejszanie się grubości warstwy lipidowej w procesie formowania się błony czarnej zachodzi spontanicznie wraz ze zmniejszaniem się energii swobodnej układu. Szczegółowe rozważania teoretyczne wskazują, że proces ten odbywa się przy udziale elektrycznej warstwy podwójnej, sił Londona–van der Waalsa i sił grawitacyjnych.

błony w postaci liposomów

Błony dwumolekularne w postaci błon liposomów można wytworzyć dzięki temu, że większość naturalnych fosfolipidów dysperguje (rozdrabnia się) w wodzie i w wodnych roztworach soli. Powstają przy tym cząstki sferyczne i cylindryczne o rozmiarach zawierających się w granicach od ułamków milimetra do kilkudziesięciu nanometrów, zwane liposomami. Każdy liposom składa się z pewnej liczby dwumolekularnych warstw lipidowych ułożonych koncentrycznie i oddzielonych od siebie warstwami wody. Najczęściej liposomy produkuje się przez napromienianie lipidów ultradźwiękami lub metodą wytrząsania lipidów w odpowiednim roztworze wodnym.

właściwośo błon lipidowych

Właściwości błon lipidowych uzależnione są od natury kwasów tłuszczowych występujących w lipidach tworzących daną błonę i od oddziaływania lipidów

między sobą. Im dłuższe są łańcuchy węglowodorowe, tym bardziej błony są zwarte; im bardziej są nienasycone, tym bardziej ich konstrukcja jest luźna.

zachowanie się łańcuchów węglowodorowych

W zależności od temperatury i wielu innych czynników (np. wzrostu kwasowości ośrodka) stwierdza się różne zachowanie łańcuchów węglowodorowych tworzących wnętrze dwumolekularnych warstw lipidowych. W temperaturach odpowiednio niskich łańcuchy te znajdują się w stanie uporządkowanym, są sztywne w pewnym stopniu i usytuowane równolegle względem siebie. W temperaturach wyższych (oraz pod wpływem innych czynników) uporządkowanie to zanika i błona przechodzi w stan zbliżony do stanu płynnego. Płynność błon pozwala wytłumaczyć obecnie wiele zjawisk membranowych, a w szczególności efekty związane z przenikaniem substancji przez błony i efekty związane z działaniem membranowych układów enzymatycznych.

Błony modelowe w postaci czystych błon lipidowych dostarczają istotnych informacji dotyczących związków między strukturą i funkcją błon modelowych, a te z kolei można odnosić w pewnym zakresie do wyjaśnienia właściwości błon komórkowych. Parametry charakteryzujące dwumolekularne błony lipidowe różnią się jednak często od analogicznych parametrów żywych błon komórkowych. Znacznie lepsze przybliżenie otrzymuje się poprzez modyfikowanie czystych błon lipidowych za pomocą białek. Takie białka np. jak walinomycyna i gramicydyna wprowadzone do błon lipidowych zmieniają właściwości transportowe tych błon. W szczególności transport pewnych jonów odbywa się w tak zmodyfikowanych błonach w sposób bardzo zbliżony do transportu tych samych jonów w pewnych błonach komórkowych.

Termodynamiczny opis zjawisk transportu

Transport substancji rozpuszczonej w elemencie objętościowym roztworu może być zrealizowany trojako. Pierwsza możliwość polega na ruchu całego elementu objętościowego — rozpuszczalnika wraz z substancją rozpuszczoną; taki przepływ nazywa się przepływem objętościowym. Druga możliwość — to ruch cząstek (na ogół cząsteczek lub jonów) zachodzący w wyniku ich ruchu cieplnego, zwany dyfuzją. Trzecia możliwość — to niedyfuzyjny ruch rozpuszczonej substancji pod wpływem sił zewnętrznych, zwany migracją.

strumień substancji

Wielkością charakteryzującą proces transportu substancji jest strumień. Przez strumień I substancji rozumiemy stosunek ilości substancji n przechodzącej w pewnym czasie przez daną powierzchnię A do tej powierzchni i do tego czasu, co można przedstawić za pomocą wyrażenia:

$$I = \frac{1}{A}\frac{dn}{dt},$$

lub równoważnie:

$$I = cv,$$

gdzie v oznacza prędkość transportowanych cząstek, c — stężenie molarne.

Zauważmy, że podane równania zostały zapisane w sposób skalarny, ale ogólnie mają one charakter wektorowy, gdyż prędkość, a wobec tego i strumień, są wektorami.

bodźce termodynamiczne

Strumienie powstają pod wpływem różnych czynników, zwanych bodźcami. Np. strumień dyfuzyjny powstaje w zasadzie pod wpływem gradientu potencjału chemicznego (albo, w roztworach nieskończenie rozcieńczonych, gradientu stężenia); tzn., że gradient potencjału chemicznego jest typowym bodźcem dla procesu dyfuzji (albo, mówiąc inaczej, jest bodźcem sprzężonym ze strumieniem dyfuzyjnym). Strumień objętościowy jest sprzężony z gradientem ciśnienia hydrostatycznego, strumień ładunków elektrycznych — z gradientem potencjału elektrycznego, strumień reakcji chemicznej (strumień produktów lub substratów) — z powinowactwem chemicznym. Większość bodźców wyraża się za pomocą gradientów, czyli ma charakter wektorowy. Ważnymi wyjątkami są termodynamiczny strumień reakcji chemicznej i jego termodynamiczny bodziec, które są skalarami. Termodynamicznym strumieniem reakcji chemicznej jest szybkość reakcji chemicznej, a odpowiednim bodźcem — powinowactwo chemiczne.

Uważamy, że dany proces transportu jest znany, jeśli znamy konkretną zależność pomiędzy strumieniem I i bodźcem X, którą ogólnie zapiszemy symbolicznie jako funkcję f:

$$I = f(X).$$

Często sformułowanie konkretnych równań tego typu oraz ich rozwiązanie jest sprawą trudną lub na dzisiejszym poziomie wiedzy — niemożliwą do zrealizowania. Wiele jednak problemów można rozwiązać wykorzystując liniową termodynamikę procesów nierównowagowych (nieodwracalnych). Opiera się ona na postulacie głoszącym, że dla dostatecznie powolnych procesów (i dla procesów stacjonarnych) strumienie są liniowymi funkcjami wszystkich bodźców w układzie:

równania fenomenologiczne

$$I_i = \sum_k L_{ik} X_k, \tag{1}$$

gdzie I_i oznacza strumień i-tej substancji, X_k — k-ty bodziec, a L_{ik} są współczynnikami liniowymi, zwanymi współczynnikami fenomenologicznymi. Współczynniki te są niezależne od strumieni i bodźców. Nazywają się one współczynnikami fenomenologicznymi prostymi, jeśli $i = k$, a współczynnikami fenomenologicznymi krzyżowymi, jeśli $i \neq k$. Wszystkie współczynniki fenomenologiczne muszą być w zasadzie wyznaczane doświadczalnie. Strumienie i bodźce o tych samych wskaźnikach są nazywane wielkościami sprzężonymi lub skoniugowanymi. Równania (1) nazywają się równaniami fenomenologicznymi.

Równania fenomenologiczne można stosować w konkretnych zagadnieniach uwzględniając trzy podstawowe zasady:

II zasada termodynamiki

— Drugą zasadę termodynamiki, a właściwie pewną jej konsekwencję, która posiada następującą postać (dowód pomijamy):

$$\frac{dS}{dt} = \sum I_i X_i > 0,$$

gdzie S jest entropią, t — czasem, dS/dt — szybkością zmiany entropii wywołaną nieodwracalnością procesów.

zasada Onsagera

— Zasadę Onsagera, która mówi, że współczynniki krzyżowe o przestawionych wskaźnikach są sobie równe:

$$L_{ik} = L_{ki} \quad (i \neq k).$$

zasada Curie

— Zasadę Curie, która jest ogólnym prawem przyrody, a w zastosowaniu do naszych rozważań może być sformułowana następująco: w układach izotropowych przyczyna skalarna nie może wywołać skutku o charakterze wektorowym. A więc np. wektorowy strumień dyfuzyjny nie może być w ośrodku jednorodnym wywołany skalarnym bodźcem chemicznym w postaci powinowactwa chemicznego.

Równania fenomenologiczne (1) dla dwóch strumieni i dla dwóch bodźców przyjmują postać:

$$\begin{aligned} I_1 &= L_{11}X_1 + L_{12}X_2 \\ I_2 &= L_{21}X_1 + L_{22}X_2. \end{aligned} \tag{2}$$

Jak widać każdy ze strumieni jest uzależniony nie tylko od bodźca z nim skoniugowanego, lecz również

od bodźca nieskoniugowanego. Ponieważ zgodnie z zasadą Onsagera $L_{12} = L_{21}$, przeto obydwa strumienie są wzajemnie sprzężone i oddziałują na siebie. Jeśli mianowicie z drugiego równania określimy X_2 i otrzymany wynik podstawimy pod X_2 w równaniu pierwszym, otrzymamy:

$$I_1 = \left(L_{11} - \frac{L_{12}^2}{L_{22}}\right) X_1 + \frac{L_{12}}{L_{22}} I_2,$$

co wskazuje na bezpośrednie sprzężenie między obydwoma strumieniami.

zjawisko termodyfuzji

Typowym przykładem fizycznym takiego sprzężenia jest zjawisko termodyfuzji. W zjawisku tym sprzężone są ze sobą strumienie dyfuzyjny i cieplny. W rezultacie, pod wpływem gradientu temperatury, dochodzi do powstania procesu dyfuzji.

Kiedy współczynniki krzyżowe są równe zeru, to sprzężenia między strumieniami nie występują. Np. jeśli $L_{12} = L_{21} = 0$, to z równań (2) pozostają dwie niezależne relacje: $I_1 = L_{11}X_1$ oraz $I_2 = L_{22}X_2$.

Formalizm liniowej termodynamiki procesów nierównowagowych pozwala opisać pewne zjawiska transportu bliskie stanom równowagi. Procesy zachodzące w stanach odległych od stanu równowagi próbuje się opisać za pomocą metod nieliniowej termodynamiki procesów nierównowagowych. Przedstawienie nawet elementów tej ostatniej wykracza poza zakres niniejszego opracowania.

Przedstawione wyżej rozważania o charakterze ogólnym zostaną w dalszym wywodzie wykorzystane do opisu pewnych zjawisk przenikania substancji przez błony komórkowe.

Przenikanie substancji przez błony biologiczne

Błona komórkowa stanowi barierę pomiędzy ośrodkiem zewnętrznym a protoplazmą, a w organizmach wielokomórkowych stanowi również barierę między sąsiednimi komórkami. Przez błonę tę przenikają różnorodne substancje zarówno z ośrodka zewnętrznego do wnętrza komórki, jak i w kierunku przeciwnym. Substancje przenikają również przez błony biologiczne innych rodzajów.

Wszystkie błony biologiczne charakteryzuje zdolność do wybiórczej przepuszczalności substancji. Pewne substancje przenikają z większą szybkością, inne z mniejszą, a jeszcze inne nie są w ogóle przepuszczane przez błonę. Błony biologiczne nie są więc błonami półprzepuszczalnymi, ponieważ umożliwiają transport nie tylko rozpuszczalnika (np. wody), lecz setek innych związków. Charakter przenikania danej substancji uzależniony jest od aktualnego stanu błony i zmienia się wraz ze zmianami błony zachodzącymi pod wpływem czynników zewnętrznych.

przenikanie bierne

Istnieje ogromna różnorodność procesów przenikania; można je podzielić na zasadnicze dwie klasy: procesy przenikania biernego i procesy przenikania aktywnego. Procesy przenikania biernego odbywają się bez nakładu energii zewnętrznej — samorzutnie, natomiast procesy aktywne wymagają nakładu energii (dostarczonej z metabolicznych reakcji chemicznych). Procesy przenikania biernego zachodzą pod wpływem różnych bodźców, takich jak np. gradient potencjału chemicznego, gradient stężenia, gradient potencjału elektrycznego. W związku z tym rozróżniamy takie procesy biernego transportu jak dyfuzja prosta, elektrodyfuzja i przenikanie ułatwione. Gdy przenikanie zachodzi wyłącznie pod wpływem gradientu potencjału chemicznego lub (dla dostatecznie małych stężeń) gradientu stężenia, to jest to tzw. dyfuzja prosta. Jeśli nie ma oddziaływań pomiędzy strumieniami danej substancji i ewentualnymi innymi strumieniami, to zależność między strumieniem i bodźcem (równanie 1) sprowadza się do jednego równania:

dyfuzja prosta

$$I_D = LX_D.$$

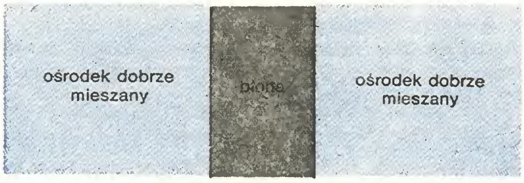
Jeśli bodźcem jest gradient stężenia $\partial c/\partial x$ (dla procesu zachodzącego w jednym tylko kierunku osi x), a $L = -D$, to otrzymamy znane prawo dyfuzji prostej (I prawo Ficka):

I prawo Ficka

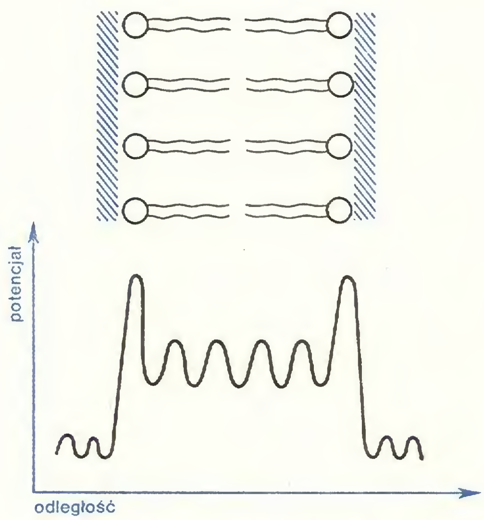
$$I_D = -D\frac{\partial c}{\partial x}, \qquad (3)$$

gdzie D jest współczynnikiem dyfuzji.

Jeżeli mamy dwa ośrodki dobrze mieszane (mieszadłami lub metodą wstrząsania), tzn. takie, że można pominąć w rozważaniach dyfuzję w tych ośrodkach, bo mieszanie zachodzi szybciej niż dyfuzja i jeżeli te ośrodki rozdzielone są błoną (rys. 16), to dyfuzja będzie się odbywała jedynie wewnątrz błony i wobec tego w równaniu (3) $\partial c/\partial x$ oznacza gradient stężenia wewnątrz błony, a D — współczynnik dyfuzji przenikającej substancji przez fazę błony. Jeśli cząsteczka dyfunduje przez ciągłą warstwę lipidową, to ruch jej



Rys. 16. Układ zawierający błonę i dwa dobrze mieszane roztwory



Rys. 17. Rozkład potencjału w obszarze lipidowej fazy błony

ma charakter przeskoków z jednego położenia równowagi do następnego poprzez szereg barier potencjału (rys. 17). Jeśli proces ma charakter stacjonarny i stężenie w jednym ośrodku wynosi c_1, a w drugim c_2 i jeśli grubość błony oznaczymy przez Δx, to równanie (3) przyjmuje postać:

$$I_D = -D\frac{\Delta c}{\Delta x} = -P\Delta c = -P'\Delta\pi,$$

gdzie $\Delta c = c_1 - c_2$, P (zwane stałą przenikania) jest równe $D/\Delta x$, $\Delta\pi$ jest różnicą ciśnień osmotycznych pomiędzy dwoma ośrodkami (ciśnienie osmotyczne π dla roztworów nieskończenie rozcieńczonych wynosi RTc, gdzie R jest stałą gazową, T — temperaturą bezwzględną, a $P' = P/RT$).

Znamy wiele substancji, które przenikają przez błony żywych komórek zgodnie z prawem prostej

dyfuzji. Nie wiadomo jednak, czy substancje te przenikają przez ciągłą fazę lipidową błony, czy też przez hipotetyczne pory w błonie. Wynika to stąd, że proces dyfuzji wyraża się w obydwu przypadkach takim samym prawem. Pewną wskazówką jest stwierdzona doświadczalnie zależność między szybkością dyfuzji i składem lipidowym błony. Im więcej nienasyconych lipidów znajduje się w błonie (im mniej zwarta jest jej struktura), tym szybciej zachodzi proces dyfuzyjnego przenikania biernego.

elektrodyfuzja

Transport jonów wynikły z równoczesnego działania pola elektrycznego i gradientu stężenia nosi nazwę elektrodyfuzji. Różnica potencjałów elektrycznych powstaje zawsze w wyniku różnic w ruchliwościach jonów (tzw. potencjał dyfuzyjny). Ponadto w błonach biologicznych (a często i w sztucznych) występuje napięcie elektryczne, którego źródłem są nie przenikające przez błonę jony i ładunki związane z błoną (tzw. potencjał Donnana). Suma napięć dyfuzyjnych, napięć Donnana i napięć wywołanych przenikaniem aktywnym nazywa się potencjałem membranowym. Przenikanie jonów przez błony żywych komórek często może mieć charakter elektrodyfuzyjny, ponieważ w układzie żywa komórka–ośrodek często może występować równocześnie gradient potencjału elektrycznego i gradient stężenia. W tym wypadku równania fenomenologiczne (2) dla procesu stacjonarnego mogą być przedstawione następująco:

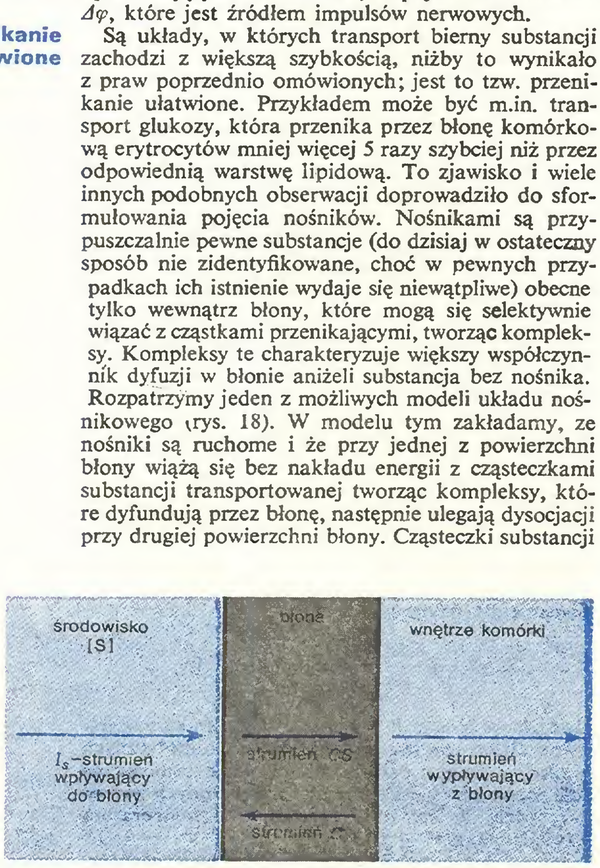
$$I_D = L_{DD}\Delta\pi + L_{D\varphi}\varphi\Delta\varphi$$

$$I_\varphi = L_{\varphi D} + L_{\varphi\varphi}\Delta\varphi,$$

gdzie I_D jest strumieniem dyfuzyjnym, $\Delta\pi$ — różnicą ciśnień osmotycznych, $\Delta\varphi$ — różnicą potencjałów membranowych, I_φ — strumieniem elektrycznym, I_{DD}, $L_{D\varphi}$, $L_{\varphi D}$, $L_{\varphi\varphi}$ — odpowiednimi współczynnikami fenomenologicznymi. Równania elektrodyfuzji pozwalają m.in. obliczyć dla wybranych jonów i przy wielu upraszczających założeniach, napięcie membranowe $\Delta\varphi$, które jest źródłem impulsów nerwowych.

przenikanie ułatwione

Są układy, w których transport bierny substancji zachodzi z większą szybkością, niżby to wynikało z praw poprzednio omówionych; jest to tzw. przenikanie ułatwione. Przykładem może być m.in. transport glukozy, która przenika przez błonę komórkową erytrocytów mniej więcej 5 razy szybciej niż przez odpowiednią warstwę lipidową. To zjawisko i wiele innych podobnych obserwacji doprowadziło do sformułowania pojęcia nośników. Nośnikami są przypuszczalnie pewne substancje (do dzisiaj w ostateczny sposób nie zidentyfikowane, choć w pewnych przypadkach ich istnienie wydaje się niewątpliwe) obecne tylko wewnątrz błony, które mogą się selektywnie wiązać z cząstkami przenikającymi, tworząc kompleksy. Kompleksy te charakteryzuje większy współczynnik dyfuzji w błonie aniżeli substancja bez nośnika. Rozpatrzymy jeden z możliwych modeli układu nośnikowego (rys. 18). W modelu tym zakładamy, że nośniki są ruchome i że przy jednej z powierzchni błony wiążą się bez nakładu energii z cząsteczkami substancji transportowanej tworząc kompleksy, które dyfundują przez błonę, następnie ulegają dysocjacji przy drugiej powierzchni błony. Cząsteczki substancji

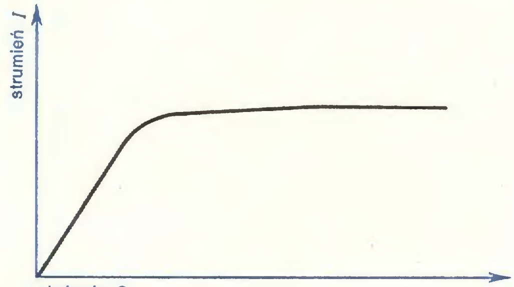


Rys. 18. Przykładowy model układu nośnikowego

transportowej zostają po drugiej stronie błony, a nośniki dyfundują z powrotem i proces się powtarza. Można wykazać, że strumień jednokierunkowy wyraża się tu następująco:

$$I = I_m\frac{[S]}{[S]+K},$$

gdzie I_m i K są pewnymi stałymi, S — stężeniem substancji przenikającej w ośrodku zewnętrznym. Jak widać dla dostatecznie dużych wartości S (tzn. jeśli $S \gg K$) wartość strumienia przyjmuje stałą wartość $I = I_m$. Przebieg zależności strumienia od stężenia zewnętrznego jest przedstawiony na rys. 19. Dla dostatecznie dużych stężeń wszystkie nośniki zostają zaangażowane w transport substancji S i wobec tego dalsze zwiększenie stężenia nie może doprowadzić do wzrostu strumienia — następuje nasycenie.



Rys. 19. Zależność pomiędzy strumieniem i stężeniem przenikającej przez błonę substancji w transporcie nośnikowym

Przenikanie bierne każdego rodzaju może zachodzić zgodnie z gradientem stężenia substancji przenikającej lub też wbrew gradientowi stężenia. Np. w procesie elektrodyfuzji transport jonów może odbywać się wbrew gradientowi stężenia, jeśli tylko w układzie występuje odpowiednie napięcie elektryczne. Możemy również obserwować przepływ substancji pomimo braku różnicy stężeń tej substancji. Np. wyobraźmy sobie komórki, które w swej błonie zawierają nośniki wiążące się zarówno z pewną substancją A jak i z pewną substancją B. Jeśli utrzymujemy komórki w ośrodku o wysokim stężeniu substancji A (w nieobecności B) tak długo, aż strumień wpływający I^A_{wp} stanie się równy strumieniowi substancji wypływającej I^A_{wyp} z komórek, to strumień netto $I^A_{net} = 0$. Jeżeli teraz dodamy do ośrodka zewnętrznego substancję B, to wówczas wystąpi konkurencja substancji A i B o wykorzystanie nośników. Początkowo konkurencja będzie dotyczyła jedynie nośników znajdujących się w pobliżu powierzchni błony od strony ośrodka zewnętrznego. Dlatego strumień wpływający stanie się mniejszy od strumienia wypływającego $I^A_{wp} < I^A_{wyp}$, a więc pojawi się strumień netto różny od zera ($I_{net} \neq 0$) pomimo braku różnicy stężeń substancji A.

przenikanie wbrew gradientowi stężenia

Przeważnie prawom transportu biernego podlegają substancje o mniejszym znaczeniu fizjologicznym. Najważniejsze z biologicznego punktu widzenia substancje (np. jony potasu i sodu, białka, wiele cukrowców) przenikają aktywnie.

przenikanie aktywne

Transport aktywny wymaga dopływu energii swobodnej z jakiegoś niezależnego źródła; przeważnie energii tej dostarczają reakcje metaboliczne. Wobec tego transport aktywny w błonach biologicznych może być określony jako przepływ dyfuzyjny sprzężony z chemicznym procesem metabolicznym.

Jednak według zasady Curie takie sprzężenie między wektorowym strumieniem dyfuzyjnym i skalarnym strumieniem chemicznym w układzie izotropowym jest niemożliwe. Stąd sądzono, że przenikanie aktywne jest specyficznym procesem biologicznym nie podlegającym prawom fizyki. Rozwiązanie tego paradoksu wynika z następującego rozumowania.

Rozważmy układ, w którym występuje bodziec wektorowy w postaci różnicy ciśnień osmotycznych $\Delta\pi$ i bodziec skalarny w postaci powinowactwa chemicznego E oraz strumień dyfuzyjny I i strumień reakcji chemicznej I_r. Równania fenomenologiczne przyjmują wówczas postać:

$$I_D = L_{DD}\Delta\pi + L_{Dr}E, \qquad (4)$$
$$I_r = L_{rD}\Delta\pi + L_{rr}E,$$

gdzie L_{DD}, D_{Dr}, L_{rD}, L_{rr} są odpowiednimi współczynnikami fenomenologicznymi. W układzie izotropowym nie ma sprzężeń, więc współczynniki krzyżowe znikają i równania sprowadzają się do dwóch równań niezależnych: $I_D = L_{DD}\Delta\pi$ oraz $I_r = L_{rr}E$, wskazujących, że w ośrodku jednorodnym mogą występować jedynie nie oddziałujące ze sobą strumienie. Stąd należy wnosić, że gdyby błona komórkowa była tworem izotropowym, to żadne procesy aktywne nie mogłyby w niej powstawać. Błona jest jednak układem anizotropowym, który charakteryzuje przestrzenna i strukturalna asymetria. Z tego względu zasada Curie przestaje obowiązywać (albo też mówiąc inaczej, strumień reakcji chemicznej w odpowiednio asymetrycznym układzie staje się strumieniem ukierunkowanym, a więc wektorowym i zasada Curie pozostaje słuszna) i w błonach biologicznych mogą istnieć sprzężenia między dyfuzją i reakcją chemiczną; w równaniach (4) współczynniki krzyżowe L_{Dr} i L_{rD} nie znikają.

asymetria żywej błony

Tak więc błona żywa ze względu na swoją asymetrię umożliwia sprzężenia między reakcjami chemicznymi i aktywnym transportem substancji i w konsekwencji jest odpowiedzialna za podstawowe procesy życiowe. Dzięki powiązaniom reakcji chemicznej z procesem transportu istnieje możliwość regulacji dopływu substancji do wnętrza komórek przy pomocy subtelnych zmian w kinetyce reakcji chemicznych zachodzących wewnątrz błon komórkowych. W rozważanym zagadnieniu, opierając się na drugiej zasadzie termodynamiki możemy napisać

sprzężenie reakcji chemicznej i transportu aktywnego

$$\frac{dS}{dt} = I_D\Delta\pi + I_r E > 0.$$

Ostatnia nierówność prowadzi do wniosku, że strumień dyfuzyjny sprzężony z reakcją chemiczną może płynąć w kierunku przeciwnym do kierunku działania jego własnego bodźca. Z nierówności tej wynika, że suma obydwu wyrazów musi być większa od zera. Wobec tego jeden z wyrazów może być ujemny, jeśli tylko ten drugi jest dodatni na tyle, że zapewnia dodatnią wartość sumy. Jeśli więc wartość wyrazu metabolicznego I_rE jest dostatecznie duża, to iloczyn $I_D\Delta\pi$ może być ujemny (co oznacza, że kierunki I_D i $\Delta\pi$ są różne) i tym samym może zmniejszać entropię. Przenikanie aktywne może również odbywać się zgodnie z gradientem, np. stężenia. Zachodzi ono jednak w takich razach z większą szybkością aniżeli miałoby to miejsce w procesie dyfuzji. Procesy tego typu spotyka się w żywych komórkach. Np. mechanizmy aktywnego transportu mogą być uruchamiane w błonie komórkowej wówczas, kiedy konieczne jest jak najszybsze wydalanie trujących produktów rozpuszczonych wewnątrz komórki.

Przenikanie aktywne zachodzi na ogół dzięki energii wyzwalanej w procesie hydrolizy kwasu adenozynotrójfosforanowego (ATP). Jest to związek wysokoenergetyczny. Oderwanie się jednej grupy fosforowej od ATP prowadzi do wydzielenia się energii potrzebnej do podtrzymywania procesu aktywnego. Powiązanie tej energii z procesem aktywnego przenikania wynika ze stwierdzeń doświadczalnych: zahamowanie procesu rozpadu ATP powoduje wstrzymanie procesu aktywnego transportu. Również bilans energetyczny wskazuje na ścisłe powiązanie pomiędzy energią pochodzącą z ATP i procesem transportu aktywnego. Nie wiadomo jednak, jak dochodzi do przekształcenia energii chemicznej w energię mechaniczną transportu. Istnieją jedynie liczne, mniej lub bardziej uzasadnione hipotezy i spekulacje. Najczęściej przyjmuje się, że przenikanie aktywne jest realizowane przy pomocy nośników, które (inaczej niż to ma miejsce w biernym transporcie nośnikowym) wiążą się z transportowaną substancją dzięki dostarczonej energii metabolicznej.

ATP — źródło energii dla przenikania aktywnego

Od czego zależy dalszy rozwój biofizyki błon

Błony biologiczne są niezbędnym składnikiem wszystkich żywych komórek i organizmów wielokomórkowych. Liczne funkcje błon związane z podstawowymi funkcjami życiowymi — odżywianiem się, fotosyntezą, powstawaniem i rozchodzeniem się impulsów nerwowych, procesami metabolicznymi — sprowadzają się przede wszystkim do zdolności regulowania procesów transportu. Matematyczny opis procesów transportu jest obecnie utrudniony z kilku co najmniej ważnych powodów, a przede wszystkim z powodu braku dobrej znajomości budowy i struktury błon; zadowalającej teorii budowy cieczy i pełnej teorii termodynamiki procesów nierównowagowych (jeśli w ogóle termodynamiczny opis może być stosowany do układów o rozmiarach 10 nm, odpowiadających grubości błony). Dalszy rozwój biofizyki błon jest więc uzależniony od rozwoju fizycznych metod pomiarowych, głównie spektroskopowych, rozwoju teorii cieczy, w szczególności roztworów i rozwoju termodynamiki małych układów, względnie innych ujęć matematyczno-fizycznych zagadnień.

trudności opisu matematycznego

M. Cereijido, C. A. Rotunno *Introduction to the Study of Biological Membranes*, New York 1970; R. M. Dowben (red.) *Błony biologiczne*, Warszawa 1973; A. Kotyk, K. Janacek *Cell Membrane Trasport*, New York 1975; W. Korochoda *Znaczenie błon komórkowych w procesach kontrolujących ruch i metabolizm komórek*, Kraków 1970; W. Leyko (red.) *Wykłady z biofizyki*, t. 1, Łódź 1975; A. P. M. Lockwood *The Membrane of Animal Cells*, London 1971; S. Przestalski *Opis zjawiska transportu przez błony biologiczne*, Postępy Biologii Komórki 2, 165 (1975); E. D. P. de Roberts i in. *Biologia komórki*, Warszawa 1974.

Białka

Kazimierz Zakrzewski

Chociaż nie potrafimy zdefiniować, czym jest życie, intuicja na ogół bezbłędnie pozwala nam odróżnić twory żywe od martwych. Skomplikowana, lecz harmonijna budowa, celowe reagowanie na bodźce zewnętrzne, a przede wszystkim zdolność do rozmnażania się, do odtwarzania w organizmie potomnym organizmów rodzicielskich to nieodłączny i nierozłączny zespół cech charakteryzujących zjawisko życia. Egzystencja zaś organizmu jest bezwzględnie uzależniona od jego metabolizmu — systemu reakcji chemicznych umożliwiających przyswajanie, przetwarzanie i wykorzystywanie substancji z otoczenia jako materiału budulcowego i źródła energii.

metabolizm organizmów

Wyjaśnienie zasad reakcji metabolicznych stworzyło podstawy obecnego burzliwego rozwoju nauk biologicznych. Udowodnienie zaś, że białka są zdefiniowanymi i zróżnicowanymi związkami chemicznymi, a nie bezpostaciową i pasywną masą „biokoloidu”,

znaczenie białek dla życia

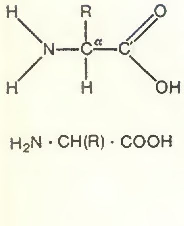
jak niegdyś sądzono, oraz że enzymy są białkami, miało znaczenie przełomowe. Enzymy, białka wykazujące czynność katalityczną, umożliwiają komórkom żywym prowadzenie złożonych reakcji chemicznych w wąskich, dopuszczalnych dla życia granicach temperatury. Bez istnienia enzymów nie byłby możliwy metabolizm, niemożliwa byłaby także replikacja (odtwarzanie) aparatu genetycznego, warunek rozmnażania się organizmów i istnienia życia na ziemi. Nie tylko jednak enzymy okazały się białkami, ale również wszystkie inne, dotąd poznane narzędzia molekularne komórki: specjalne białka odpowiedzialne są za transport substancji i ich wymianę z otoczeniem, inne — za odbieranie i przekazywanie sygnałów chemicznych i fizycznych o stanie środowiska wewnętrznego i zewnętrznego oraz za koordynację funkcji biologicznych. Jeżeli się do tego doda, że z białek powstaje też zrąb strukturalny, układ składników zapewniających trwałą a elastyczną organizację wewnętrzną tworów żywych, to stanie się oczywiste, że mają one uniwersalne znaczenie dla życia.

Budowa białek

Wszechstronnym funkcjom białek towarzyszy ogromna różnorodność ich własności fizycznych i chemicznych — wielkości, kształtu i ładunku elektrycznego cząsteczki, barwy, rozpuszczalności itd. Wspólną jednak ich cechą jest to, że zbudowane są z tych samych dwudziestu różniących się między sobą aminokwasów, połączonych w długi i nierozgałęziony łańcuch. I to właśnie leży u podstaw różnorodności struktury i funkcji białek. W łańcuchu bowiem każdy z dwudziestu aminokwasów może występować w dowolnej pozycji i wielokrotnie. Proste obliczenie wskazuje, że w małym białku, zbudowanym np. ze stu aminokwasów, dwadzieścia różnych aminokwasów może być ułożonych na 10^{130} sposobów, może więc dać tyleż różniących się rodzajów cząsteczek. Jest to więcej niż wynosi liczba atomów we Wszechświecie, szacowana na 10^{80}, ale w przyrodzie występuje tylko niewielka ich część, być może — ok. 10^{12} różnych białek. Znacznie więcej było ich bez wątpienia wypróbowywanych w toku ewolucji i odrzuconych jako nie przyczyniających się do przeżycia gatunku. Białka są na ogół bardzo dobrze dostosowane do warunków bytowania organizmów, chociaż są i takie — produkty świeżych lub starszych mutacji — które mają własności niekorzystne, oraz białka różniące się nieco od optymalnych, nie pogarszające szansy przeżycia i jak gdyby oczekujące na ujawnienie swoich walorów wówczas, gdy się zmienią warunki bytowania organizmu.

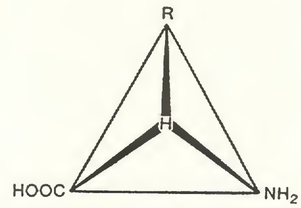
aminokwasy — cegiełki budowy

W przyrodzie występuje wiele dziesiątków różnych aminokwasów (a więcej ich można otrzymać syntetycznie), lecz białka budowane są tylko z tych dwudziestu, które mają w zapisie informacji genetycznej odpowiednie kodony (→ Kwasy nukleinowe). Są to α-aminokwasy (rys. 1) o charakterystycznej i jednakowej konfiguracji atomowej przy węglu C^{α}, lecz różniące się swymi łańcuchami bocznymi (tabela). Ponieważ cztery wartościowości węgla C^{α} wysycone są różnymi podstawnikami, α-aminokwasy wykazują



Rys. 1. Wzór strukturalny α-aminokwasu (u dołu w postaci skróconej). Aminokwasy, podstawowe cegiełki budowy białka, są związkami organicznymi, zawierającymi w swojej cząsteczce grupę aminową ($-NH_2$) o własnościach zasadowych i grupę karboksylową (—COOH) o własnościach kwasowych. Białka budowane są z jednego tylko rodzaju aminokwasów — takich, w których grupa aminowa i grupa karboksylowa znajdują się przy tym samym atomie węgla, oznaczanym jako C^{α}; zwane są one α-aminokwasami. W pozycji oznaczonej R znajduje się łańcuch boczny aminokwasu (zob. tabela). Atomy węgla w łańcuchu bocznym oznacza się kolejnymi literami greckimi β, γ itd. zaczynając od węgla C^{α}; atom węgla grupy kwasowej oznacza się jako C′

czynność optyczną (rys. 2), a z dwu możliwych stereoizomeronów w białkach występują tylko L-α-aminokwasy (wyjątkiem jest tu pozbawiona czynności optycznej glicyna, która ma dwie wartościowości C^{α} podstawione atomami wodoru: tabela). Przyczyny, dla których przyroda wybrała tę właśnie możliwość, są nie znane. Mogły to spowodować warunki panujące na naszej planecie w okresie powstawania życia na ziemi, np. charakter pola magnetycznego, polaryzacja światła słonecznego, siły związane z obrotem kuli ziemskiej wokół osi. Również możliwe jest, że konfiguracja L ma jakąś wyższość nad D albo że o wyborze konfiguracji zdecydował przypadek: pierwszy zdolny do długotrwałej samoreprodukcji twór mógł być zbudowany z L-α-aminokwasów i uniemożliwił wykształcenie się innych form życia.



Rys. 2. Struktura przestrzenna α-aminokwasów. Tworzy on tetraedr, w którego środku znajduje się atom węgla C (niewidoczny na rysunku), natomiast grupy aminowa, karboksylowa i łańcuch boczny R rozmieszczone są w wierzchołkach tetraedru zgodnie z ruchem wskazówek zegara w L-α-aminokwasach i przeciwnie — w D-α-aminokwasach

peptydy — szkielet budowy

Białka syntetyzowane są w procesie translacji (→ Kwasy nukleinowe) przez wyłączenie cząsteczki wody z grupy —COOH jednego aminokwasu i grupy NH_2 następnego (rys. 3). Powstaje między nimi wiązanie peptydowe (rys. 4), ważny czynnik strukturotwórczy w białkach, i formuje się łańcuch polipeptydowy. W roztworze wodnym, jedynym środowisku, w jakim na naszej planecie mogą zachodzić zjawiska życia, przybiera on ukształtowanie przestrzenne uzależnione od właściwości wiązania peptydowego oraz od kolejności występowania w polipeptydzie rozmaitych łańcuchów bocznych aminokwasów (rys. 3). Kolejność ta, zwana sekwencją aminokwasową lub strukturą pierwszorzędową białka, jest bezpośrednim odzwierciedleniem zapisu informacji genetycznej w genie i determinuje przestrzenne ukształtowanie (konformację, strukturę wyższego rzędu) całej cząsteczki. Struktura wyższego rzędu sprawia, że w cząsteczce białka przybliżają się do siebie aminokwasy, na ogół znajdujące się w odległych miejscach łańcucha polipeptydowego, i powstają charakterystyczne dla danego białka ich zgrupowania, będące centrami czynnymi (np. o funkcji katalitycznej). Dopiero ostatecznie ukształtowana cząsteczka białka jest zdolna do spełniania właściwych funkcji biologicznych, a nie pierwotnie syntetyzowany łańcuch polipeptydowy. Sens więc biologiczny zawarty w zapisie informacji genetycznej ujawnia się nie w sekwencji aminokwasowej białka, ale w jego przestrzennym ukształtowaniu.

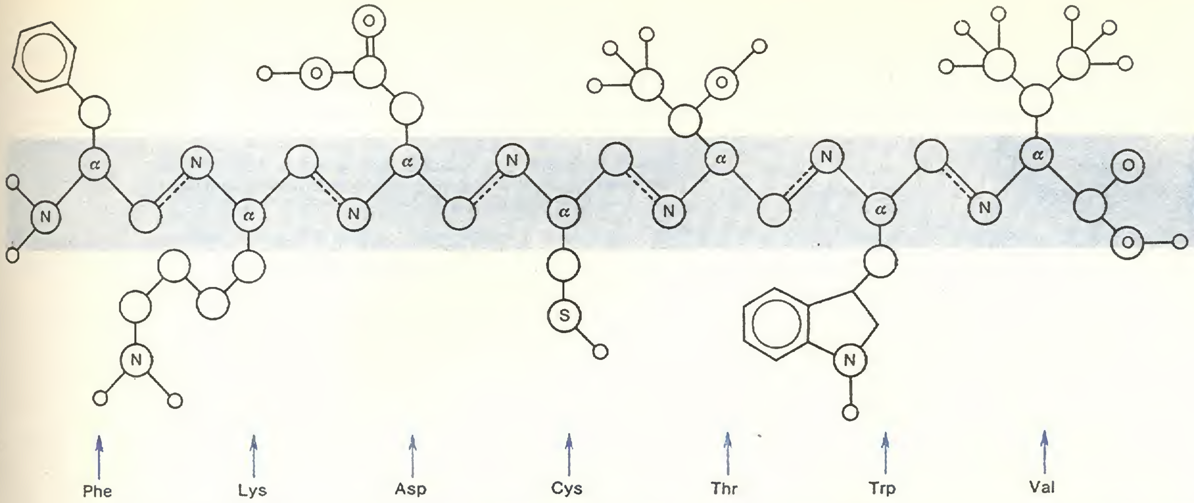
struktura pierwszorzędowa białek

W strukturze przestrzennej białka występują dwa elementy. Jeden z nich zależy od ciągłego powtarzania się wzdłuż łańcucha głównego układu atomów $—C^{\alpha}—C'—N—$ (rys. 3), identycznego we wszystkich białkach i określającego możliwości konformacyjne łańcucha głównego. Elementem drugim jest genetycznie ustalona kolejność różnych aminokwasów, ograniczająca te możliwości i nadająca danemu białku właściwą mu strukturę i funkcję.

Łańcuch główny ma znaczną swobodę w przybieraniu różnych postaci przestrzennych w roztworze wodnym, gdyż wiązania idące do węgla C^{α} (rys. 3) są nasycone i pozwalają na swobodny obrót atomów wokół ich osi — w odróżnieniu od sztywnego i płaskiego wiązania peptydowego (rys. 4). Zmienne kąty

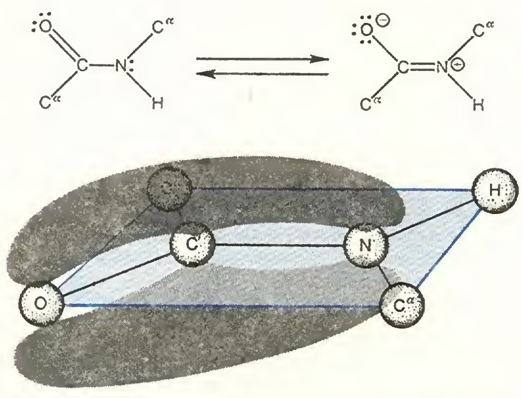
podstawowa budowa białka

$$H_2N \cdot CH(R_1) \cdot COOH + H_2N \cdot CH(R_2) \cdot COOH \xrightarrow{-H_2O} H_2N \cdot CH(R_1) \cdot CO \cdot NH \cdot CH(R_2) \cdot COOH$$



Rys. 3. Podstawowa budowa białka. Aminokwasy łączą się ze sobą w łańcuch w reakcji polegającej na wydzieleniu 1 cząsteczki wody z grupy karboksylowej jednego aminokwasu i grupy aminowej następnego (wzór w formie skróconej na górze). W powstającym łańcuchu znikają grupy aminowe i karboksylowe przy C^{α} z wyjątkiem grupy aminowej pierwszego aminokwasu (tzw. koniec aminowy lub koniec *N* łańcucha) i grupy karboksylowej ostatniego (koniec karboksylowy czyli koniec *C*). Rysunek ukazuje schematycznie budowę związku utworzonego przez 7 reszt aminokwasowych. W schemacie tym, nie odzwierciedlającym budowy przestrzennej łańcucha (z wyjątkiem faktu, że łańcuchy boczne odchodzą naprzemiennie w przeciwne strony od łańcucha głównego), zaznaczono: symbolem α atomy C^{α}, atomy azotu (N), siarki (S), tlenu (O); puste duże kółka oznaczają atomy węgla (ale w pierścieniach aromatycznych np. tyrozyny, atomy węgla pominięto), małe kółka — atomy wodoru (zaznaczone tylko na grupach skrajnych). Nazwy trójliterowe aminokwasów są wpisane dokładnie pod C^{α} danego aminokwasu. Związki zbudowane z dwu reszt aminokwasowych nazywane są dwupeptydami, z trzech — trójpeptydami itd. Związki zbudowane z kilku do kilkunastu aminokwasów określa się ogólnie jako oligopeptydy, a większe — jako polipeptydy. Pojęcie „białko" jest bardziej ogólne i mniej precyzyjne: obejmuje ono łańcuchy polipeptydowe (na ogół dłuższe niż 100-aminokwasowe) a także cząsteczki zbudowane z kilku łańcuchów polipeptydowych oraz polipeptydy z przyłączonymi związkami nie mającymi charakteru aminokwasów. Glikoproteidy np. są łańcuchem polipeptydowym z przyłączonymi resztami cukrowca, w metaloproteidach zaś występują atomy metalu jako integralny składnik danego białka

torysyjne przy C^{α} (rys. 5) nie mogą przybierać takich wartości, które by przybliżyły do siebie jakiekolwiek atomy, nie związane chemicznie, na odległości mniej-

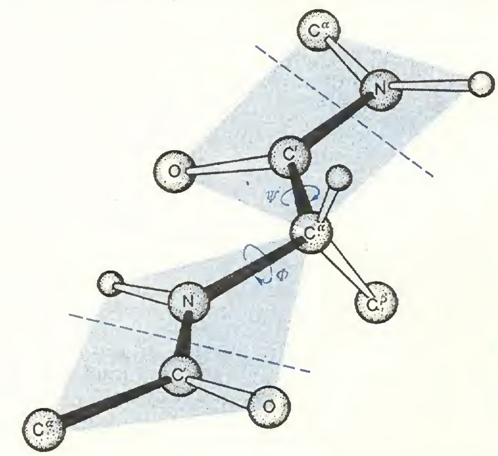


Rys. 4. Wiązanie peptydowe. Wiązanie peptydowe —CO—NH—, powstające przez kondensację grupy karboksylowej jednego aminokwasu z grupą aminową innego (zob. rys. 3) ma zdelokalizowane elektrony, tworzące orbital molekularny rozciągający się od atomu tlenu poprzez atom węgla C′ do atomu azotu (rysunek na dole). Jego skrajne postacie pokazuje rysunek na górze. Rezonansowa energia stabilizacji wiązania peptydowego wynosi ok. 88 $\cdot 10^3$ J/mol (21 kcal/mol) i stanowi przeszkodę dla obrotu atomów wokół osi —C′—N—. W rezultacie, wszystkie atomy spięte wiązaniem peptydowym znajdują się (w przybliżeniu) na jednej płaszczyźnie. Mogą one być w układzie *trans* jak na rysunku dolnym (atomy C^{α} skierowane w przeciwne strony) lub *cis* (atomy C^{α} skierowane w tę samą stronę). W białkach wiązanie peptydowe ma konfigurację *trans*

sze niż suma ich promieni van der Waalsa (→ Chemia kwantowa). Cząsteczka zaś jako całość osiągnie możliwie najniższy stan energetyczny, a więc przybierze trwałą konformację (→ Przedmiot i problemy biofizyki molekularnej), gdy pomiędzy możliwie najwięk-

rola wiązań wodorowych

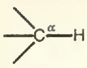
szą liczbą grup —CO i —NH (jedynych reaktywnych grup w łańcuchu głównym) powstaną wiązania wodorowe (rys. 6). Muszą one, ze względu na sztywność wiązania peptydowego, tworzyć się pomiędzy resztami aminokwasów bardziej od siebie odległymi, a nie tymi, które są ze sobą połączone wiązaniem peptydowym. Maksymalna liczba wiązań wodorowych może powstać wtedy, gdy łańcuch polipeptydowy ma sy-



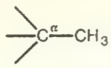
łańcuch polipeptydowy

Rys. 5. Łańcuch polipeptydowy — widok perspektywiczny. Rysunek ukazuje jedną resztę aminokwasową (między niebieskimi kreskami) i początki dwu reszt sąsiadujących. Płaszczyzny wiązań peptydowych zaznaczono kolorem niebieskim. Czarna gruba linia wytycza łańcuch główny, łańcuchy boczne pominięto, zostawiając tylko ich pierwsze węgle (C^{β}). Atomy wodoru (małe kółka) pozostawiono tylko przy atomach C^{α} i N. Łańcuch polipeptydowy na rysunku znajduje się w postaci maksymalnie rozciągniętej. Zgodnie z obowiązującą konwencją, kąt torsyjny wiązania peptydowego przyjęto za 180°, a w maksymalnie rozciągniętym łańcuchu kąty torsyjne Φ i Ψ przyjęto jako również mające wartość 180°. Obrót płaszczyzn w kierunku strzałek opisuje się jako wzrost wartości kątów torsyjnych Φ i Ψ

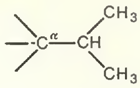
Aminokwasy hydrofobowe, apolarne



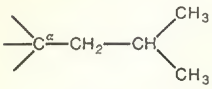
glicyna
(Gly)



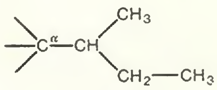
alanina
(Ala)



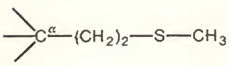
walina
(Val)



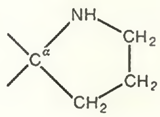
leucyna
(Leu)



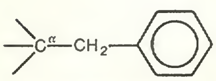
izoleucyna
(Ile)



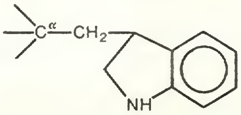
metionina
(Met)



prolina
(Pro)

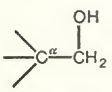


fenyloalanina
(Phe)

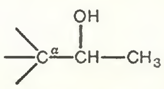


tryptofan
(Trp)

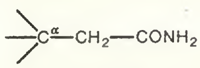
Aminokwasy hydrofilne, polarne,* niezjonizowane
w warunkach fizjologicznych



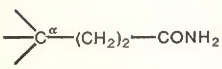
seryna
(Ser)



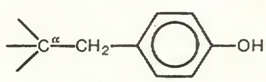
treonina
(Thr)



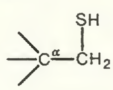
asparagina
(Asn)



glutamina
(Glu)

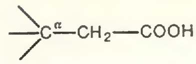


tyrozyna
(Tyr)

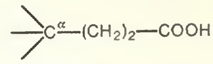


cysteina
(Cys)

Aminokwasy hydrofilne, ujemnie zjonizowane
w warunkach fizjologicznych

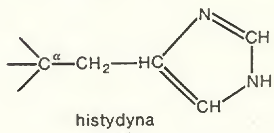


kwas asparaginowy
(Asp)

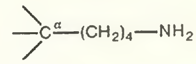


kwas glutaminowy
(Glu)

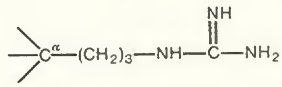
Aminokwasy hydrofilne, dodatnio zjonizowane
w warunkach fizjologicznych



histydyna
(His)

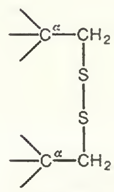


lizyna
(Lys)

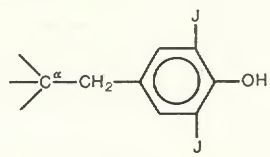


arginina
(Arg)

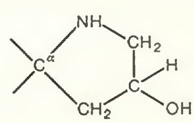
Niektóre ważniejsze produkty potranslacyjnej** modyfikacji aminokwasów



cystyna



dwujodotyrozyna



hydroksyprolina

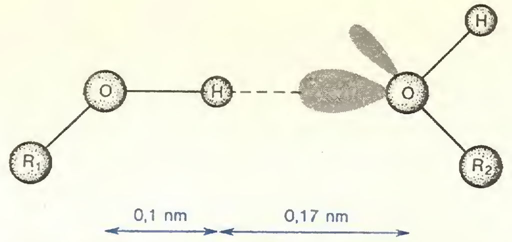
* Stan jonizacji polarnych aminokwasów podany jest tylko w przybliżeniu, gdyż w zależności od różnych czynników jonizacja może być częściowo zahamowana lub może się pojawić.

** Potranslacyjna modyfikacja aminokwasów jest wynikiem działania enzymów na aminokwas już wbudowany do łańcucha polipeptydowego: dwie cysteiny po utlenieniu ich grup-SH dają cystynę, z proliny powstaje hydroksyprolina itd. Modyfikacji takich znanych jest kilkadziesiąt.

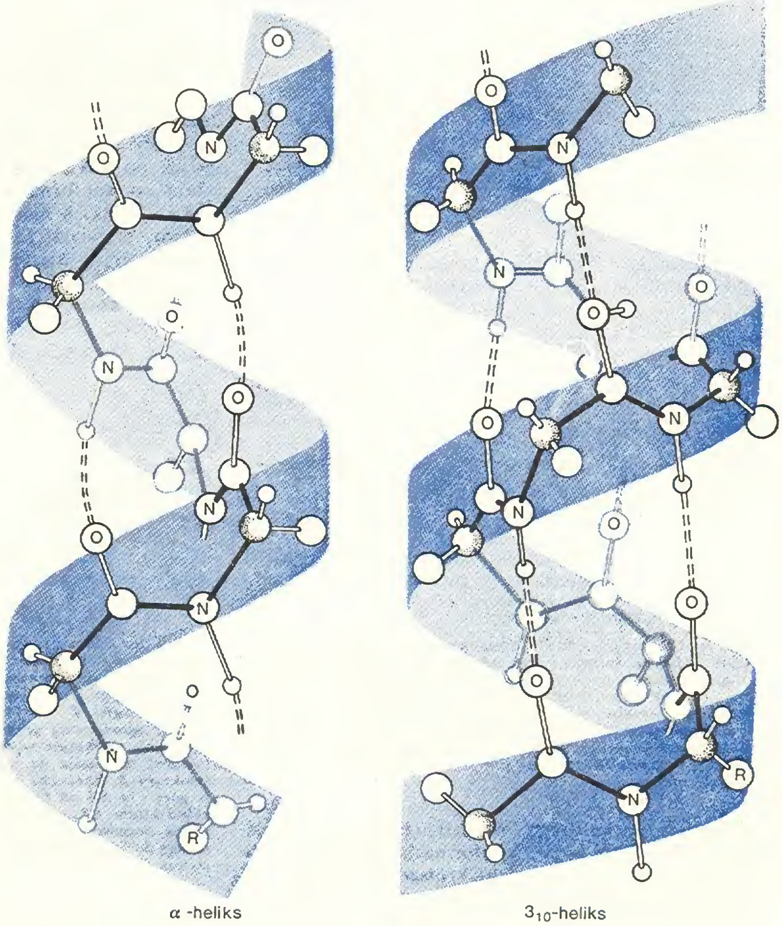
metrię śrubową (→ Budowa kryształów), a to dzięki temu, że obrót i translacja kolejnych reszt aminokwasowych stwarzają warunki do odpowiedniego zbliżenia odległych od siebie wiązań peptydowych. Drugą możliwością osiągnięcia maksymalnego wysycenia wartościowości potencjalnych partnerów wiązania wodorowego jest bliskie ulokowanie równolegle ułożonych łańcuchów polipeptydowych w postaci rozprostowanej (stosunki przestrzenne między sąsiadującymi w łańcuchu polipeptydowym resztami aminokwasów nazywane bywają strukturą drugorzędową białka).

Istnieje kilka teoretycznie możliwych struktur łańcucha polipeptydowego o symetrii śrubowej, tj. dających w efekcie cząsteczkę o kształcie określanym jako heliks. W białkach wykryto tylko heliksy prawoskrętne, należące do rodziny zwanej strukturami α (rys. 7). Ponieważ duże wymiary atomu tlenu w grupie —CO bardzo ograniczają możliwości uzwojenia łańcucha głównego, struktury α mogą mieć w jednym całkowitym zwoju tylko od ok. 2 do prawie 5 reszt aminokwasowych. Wiązania wodorowe w nich mogą łączyć odpowiednio grupę —CO każdego wiązania peptydowego z grupą —NH trzeciego, czwartego, piątego lub szóstego aminokwasu. Struktury α różnią się między sobą nie tylko liczbą reszt aminokwasowych w jednym zwoju, ale także pochyleniem uzwojenia, a to sprawia, że w niektórych heliksach wiązania wodorowe są dosyć naprężone i mało trwałe. Najbardziej korzystne warunki do utworzenia wiązań wodorowych pojawiają się w α-heliksie (rys. 7b) i on też najczęściej bywa obserwowany w białkach (a przynajmniej tych, które zbadano krystalograficznie); α-heliks w białkach często zakończony jest krótkim odcinkiem 3_{10}-heliksu (rys. 7a). Lewoskrętnych struktur α w białkach nie wykryto. Występowanie ich jest mało prawdopodobne, gdyż asymetrycznie ulokowany w L-α-aminokwasach łańcuch boczny stanowi przeszkodę do swobodnego uzwajania się łańcucha w lewo. Lustrzanym odbiciem

prawoskrętne α-heliksy



Rys. 6. Wiązanie wodorowe. Wiązanie wodorowe tworzy atom wodoru, znajdujący się między dwoma silnie elektroujemnymi atomami. W białkach wiąże ono najczęściej atomy tlenu, azotu i siarki w rozmaitych kombinacjach. Na rysunku przedstawiono wiązanie wodorowe między dwoma atomami tlenu, z których znajdujący się po lewej stronie nazywany jest donorem, a po prawej — akceptorem wodoru. Przy akceptorze zaznaczono gęstości elektronowe na niebiesko, a skala odległości podana jest pod rysunkiem. Znaczenie wiązania wodorowego dla białek polega na tym, że działa ono na stosunkowo dużą odległość (0,25–0,30 nm) i jest dosyć silne ($1-3 \cdot 10^4$ J/mol). Istnienie wiązania wodorowego uzależnione jest od odpowiedniego zbliżenia donora i akceptora, ale także od stanu środowiska. Przy zmianie jego przenikalności elektrycznej, kwasoty i in., wiązanie wodorowe może pęknąć lub powstać, co wpływa na strukturę białka i uzależnia ją w ten sposób od zjawisk chemicznych zachodzących w środowisku



Rys. 7. Struktury helikalne w białkach — modele atomowe struktury 3_{10}-heliksu i α-heliksu. Taśmą niebieską zaznaczono wyidealizowany przebieg heliksu, a grubą kreską — aktualny układ jego łańcucha głównego. Atomy C^{α} są zakropkowane, symbolami N i O oznaczono atomy azotu i tlenu wiązań peptydowych, między którymi powstają wiązania wodorowe (kreski przerywane). Obie struktury należą do struktur α, które opisuje wzór:

$$\overset{O\text{-------------------------------}}{\text{—C—}}[NH \cdot CH(R) \cdot CO]_n\overset{H}{\text{—N—}};$$

n oznacza liczbę reszt aminokwasowych zamkniętych w pętlę wiązaniem wodorowym, w 1 pętli zatem znajduje się $3n+4$ atomów. Strukturę α wystarczająco opisują dwie liczby, określające liczbę reszt w jednym zwoju i liczbę atomów w jednej pętli. Np. 3_{10}-heliks ma 3 reszty aminokwasowe w zwoju i 10 atomów w pętli; heliks tradycyjnie nazywany α-heliksem ma formalną nazwę $3{,}6_{13}$-heliks. α-heliks ma bardzo mało zdeformowane wiązania wodorowe i on to głównie występuje w białkach, czasem będąc zakończony krótkim 3_{10}-heliksem

prawoskrętnych heliksów zbudowanych z L-α-aminokwasów są lewoskrętne heliksy zbudowane z D-α-aminokwasów, a te nie są w przyrodzie ziemskiej do tworzenia białek wykorzystywane.

Drugą regularną, termodynamicznie stabilną konformacją łańcucha polipeptydowego są struktury β (rys. 8). Łańcuchy polipeptydowe są w nich prawie maksymalnie rozprostowane, ułożone w stosunku do siebie równolegle lub przeciwrównolegle i połączone wiązaniami wodorowymi. Mają one charakter płaskich, lekko pofałdowanych arkuszy (stąd ich nazwa „plisowane arkusze"; rys. 9) o bardzo regularnym rozmieszczeniu łańcuchów bocznych. W białkach mogą one mieć nawet bardzo duże rozmiary (rys. 10) i wtedy cały arkusz wykazuje lekkie skręcenie płaszczyzny w prawo, a całość struktury układa się w kształt zbliżony do cylindrycznego (tzw. β-baryłka). Mniejsze lub większe struktury β występują we wszystkich białkach i odgrywają w nich szczególnie ważną rolę.

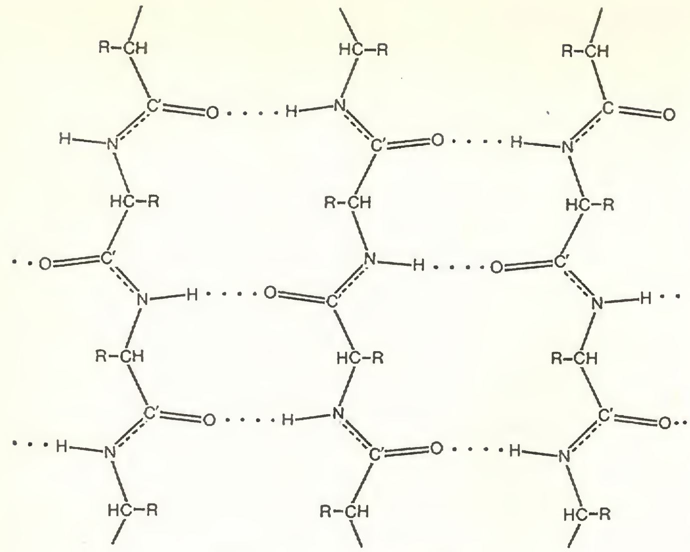
struktury β

Są one bowiem znacznie bardziej sztywne niż struktury helikalne, nadają więc białkom dobre właściwości mechaniczne, a jednocześnie mogą zapewnić aminokwasom o decydującym znaczeniu, np. w centrach katalitycznych, precyzyjnie ustaloną pozycję. Struktury helikalne — w odróżnieniu od struktur β — mają charakter bardziej sprężysty, ponieważ stabilizowane są wiązaniami wodorowymi, dosyć odpornymi na rozciąganie i tolerującymi pewne zmiany w położeniu tworzących je partnerów. Dlatego to struktury helikalne odgrywają dużą rolę w formowaniu przestrzennego kształtu cząsteczki białka po zakończeniu syntezy łańcucha polipeptydowego, a ostatecznej cząsteczce zapewniają ruchliwość konformacyjną, dynamiczne zmiany stosunków przestrzennych między różnymi odcinkami łańcucha polipeptydowego (→ Przedmiot i problemy biofizyki molekularnej).

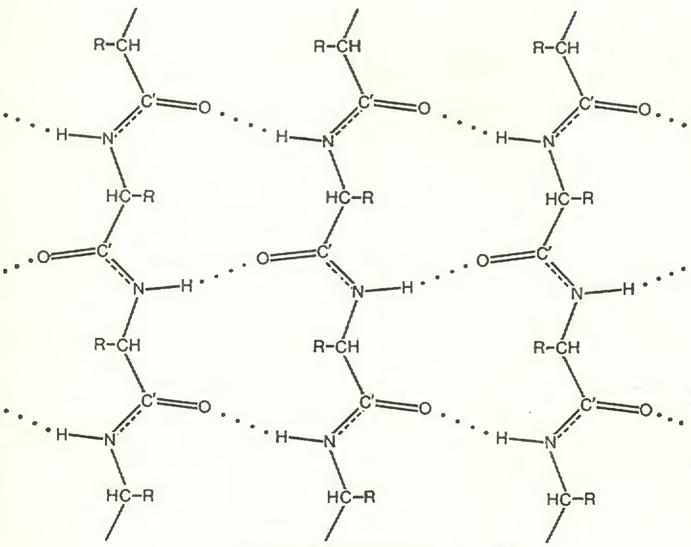
struktury helikalne

białka fibrylarne

W niektórych białkach, zwanych fibrylarnymi, czyli włókienkowymi, dominuje regularna konformacja (α lub β) większej części czy nawet całego ich łańcucha polipeptydowego. Cząsteczka białka fibrylarnego ma wówczas kształt bardzo wydłużony, jak np. w białkach podporowych, stanowiących osnowę zrębu strukturalnego organizmów wyższych (np. kolagen). Większość jednak białek ma cząsteczkę o kształcie



struktura β przeciwrównoległa



struktura β równoległa

Rys. 8. Struktury β. Prawie rozprostowane łańcuchy polipeptydowe (odrębne łańcuchy lub odcinki tego samego łańcucha) mogą wytworzyć między sobą sieć wiązań wodorowych, utrzymujących ich układ przestrzenny. Na rysunku pokazano wzory chemiczne struktur β o równoległym lub przeciwrównoległym przebiegu łańcucha (kierunek określa pozycja jego końca *N*). W układzie przeciwrównoległym, który w białkach występuje częściej, wiązania wodorowe są mniej zdeformowane niż w układzie równoległym

białka globularne

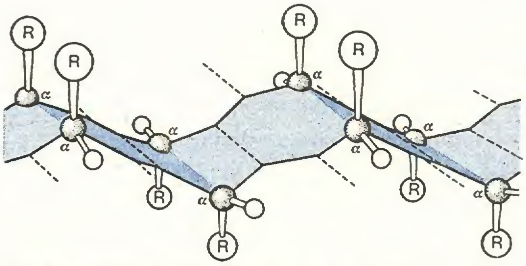
zbliżonym do kulistego. W białkach takich, określanych jako globularne, konformacja łańcucha polipeptydowego jest mieszana: krótsze lub dłuższe odcinki o konformacji α lub β połączone są odcinkami o konformacji nieregularnej (stosunki przestrzenne między odcinkami łańcucha polipeptydowego nazywane bywają strukturą trzeciorzędową białka). Mechanizmy, które kierują części łańcucha polipeptydowego ku określonej konformacji nie zostały jeszcze dobrze poznane (→ Przedmiot i problemy biofizyki molekularnej), dwa jednak czynniki mają szczególnie duże znaczenie: zróżnicowane powinowactwo aminokwasów do wody oraz rozmaitość własności chemicznych łańcuchów bocznych.

Grupy polarne, znajdujące się w łańcuchach bocznych aminokwasów hydrofilowych (np. —OH, —COOH; tabela), wiążą cząsteczki wody i stąd ich termodynamicznie faworyzowaną pozycją w cząsteczce białka jest ta, która im zapewnia dobry kontakt z wodą. Odwrotnie się zachowują węglowodorowe łańcuchy boczne aminokwasów hydrofobowych: w środowisku wodnym otaczają się one parakrystaliczną siatką cząsteczek wody, co powoduje spadek entropii wody i wzrost swobodnej energii układu. Układ więc dąży do maksymalnego zmniejszenia kontaktów między wodą a apolarnymi łańcuchami bocznymi, co osiąga przez wydzielenie ich w postaci bezwodnych skupień. Przeciętny łańcuch apolarny aminokwasu hydrofobowego w skupieniach tych zyskuje ok. 17 kJ/mol (4 kcal/mol) swobodnej energii stabilizacji, głównie kosztem entropii wody, której cząsteczki odzyskują swobodę przyjęcia mniej uporządkowanego rozmieszczenia. W rezultacie — cząsteczka przybiera kształt „kropli oliwy w wodzie": aminokwasy polarne lokują się na powierzchni w kontakcie z wodą, aminokwasy apolarne zaś tworzą bezwodne, silnie skupione wnętrze, mające bardziej charakter ciała stałego niż roztworu, gdyż upakowanie atomów może w nim wynosić do 75% dostępnej przestrzeni.

aminokwasy hydrofilowe i hydrofobowe

wpływ proliny na strukturę

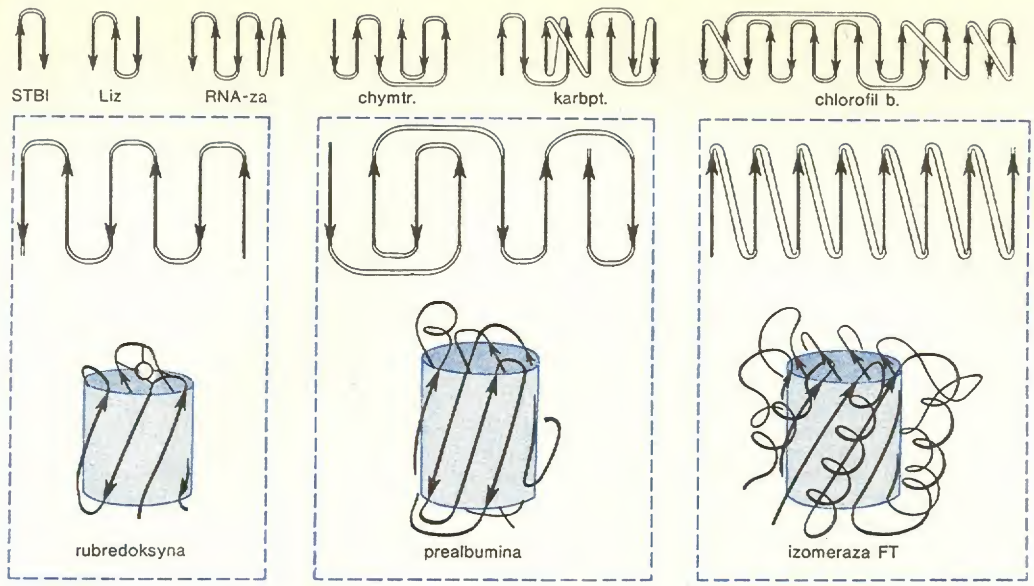
Na ostateczne ukształtowanie przestrzenne cząsteczki białka szczególnie duży wpływ mają niektóre aminokwasy. Prolina np. nie może być ulokowana w prawoskrętnej strukturze α (ale może — w lewoskrętnym heliksie, jak np. w kolagenie, o czym niżej) ze względu na odmienną niż w innych aminokwasach konfigurację atomów przy węglu C^{α}. Tworzący się prawoskrętny heliks musi ulec przerwaniu w miejscu, w którym występuje prolina, i może ponownie powstać po kilku aminokwasach, które na ogół pozostają w konformacji nieregularnej. Występowanie takich krótkich nieregularnych odcinków za proliną często sprawiają, że następny odcinek helikalny ma zupełnie inny kierunek niż poprzedni. Aminokwasy o polarnych, lecz nie zjonizowanych grupach w łańcuchach bocznych, np. seryna, asparagina (tabela)



Rys. 9. Struktura β układająca się tak jak pofałdowany (plisowany) arkusz. Łańcuchy polipetydowe w strukturach β nie są całkowicie rozprostowane, tj. ich kąty torsyjne Φ i Ψ (zob. rys. 5) odbiegają nieco od wartości 180° (np. w strukturze przeciwrównoległej $\Phi = -140°$, a $\Psi = 135°$). W rezultacie płaszczyzna ograniczona przez łańcuch główny wykazuje regularne załamania. Wiązania wodorowe między łańcuchami (linie przerywane) pozwalają na wytworzenie bardzo rozległych arkuszy struktury β. Łańcuchy boczne R skierowane są naprzemiennie w górę i w dół od płaszczyzny arkusza. Na rysunku pokazano tylko atomy C^{α}, atom wodoru przy C oraz pierwszy atom łańcucha bocznego R. Płaszczyzny arkusza struktury wykazują tendencję do lekkiego skrętu w prawo (nie zaznaczoną na rysunku), co czasem prowadzi do powstania tzw. β-baryłki (zob. rys. 10)

utrudniają tworzenie regularnych konformacji, gdyż między ich grupami polarnymi a grupami —CO czy —NH łańcucha głównego mogą powstać wiązania wodorowe, utrudniające wytworzenie wiązań wodorowych potrzebnych do stabilizacji struktury α czy β. Jednocześnie jednak takie wiązanie wodorowe między łańcuchem bocznym aminokwasu a łańcuchem głównym utrwala zwrot łańcucha głównego o ok. 180° (zwany zwrotem β lub zwrotem typu szpilki do włosów) i umożliwia zapoczątkowanie struktury β. Całość struktury białka stabilizowana jest rozmaitymi wiązaniami, wśród których szczególnie dużą rolę odgrywają silne, kowalencyjne wiązania

wiązania stabilizujące strukturę

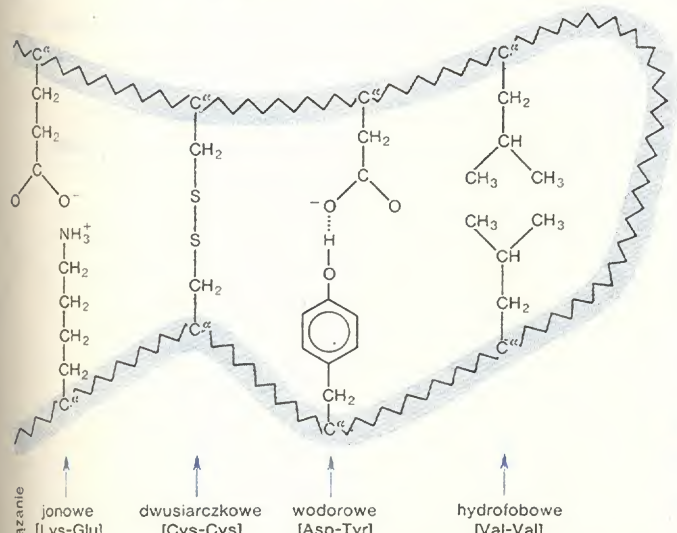


typologia struktur β

Rys. 10. Topologia struktur β — struktury β kilku spośród wielu zbadanych krystalograficznie białek. Grube proste strzałki obrazują przebieg łańcucha głównego polipeptydu, znajdującego się w płaszczyźnie struktury (płaszczyzna papieru), a połączenia między odcinkami łańcucha głównego przedstawiają linie cienkie pojedyncze — jeżeli przebiegają pod płaszczyzną struktury β i kreską podwójną — jeżeli przebiegają nad płaszczyzną. W strukturach β występują dwa rodzaje połączeń: łańcuch polipeptydowy może opuścić płaszczyznę struktury β i powrócić na nią z tej samej strony (zwrot typu „szpilki do włosów") lub też powraca ze strony przeciwnej („przekrzyżowanie"). W rysunkach górnych nie uwzględniono ani długości ani ich konformacji. Najmniejsze struktury β zbudowane są z łańcuchów o przebiegu przeciwrównoległym (STBI — inhibitor enzymów proteolitycznych oraz Liz — enzym lizozym), a układy równolegle pojawiają się dopiero w strukturach większych (enzymy: RNA-za — rybonukleaza, chymtr. — chymotrypsyna, karbpt. — karboksypeptydaza, oraz chlorofil bakteryjny). Większe struktury β, a zwłaszcza te, które mają bardzo regularny przebieg i brak przekrzyżowań, układają się cylindrycznie, w tzw. β — baryłkę. Połączenia między łańcuchami w strukturze β często mają charakter helikalny (rubredoksyna — metaloproteid bakteryjny, prealbumina — białko osocza krwi, izomeraza FT — enzym przemiany cukrowcowej)

—S—S— (tzw. mostki dwusiarczkowe) między dwiema resztami cysteiny (rys. 11).

Nie wszystkie polarne reszty aminokwasów znajdują się na powierzchni i nie wszystkie apolarne wewnątrz cząsteczki białka. Tam, gdzie pozycja danego aminokwasu nie odpowiada jego termodynamicznie faworyzowanej lokalizacji, w cząsteczce białka powstają lokalne naprężenia, mimo że jako całość może się ona znajdować w minimum energetycznym. Ma to bardzo duże znaczenie dla funkcji biologicznej białka. Skryte w hydrofobowym wnętrzu grupy



Rys. 11. Niektóre wiązania międzyłańcuchowe w białkach. Dwa odcinki tego samego łańcucha polipeptydowego (linia zygzakowata na niebieskim tle) mogą się ze sobą połączyć wiązaniem jonowym, dwusiarczkowym, wodorowym lub hydrofobowym. Są to przykłady najczęściej w białkach występujących wiązań, służących do stabilizacji ich struktury trzeciorzędowej. Wzory reszt aminokwasów, tworzących dane wiązanie, podano pod schematem w skrótach trójliterowych

polarne silniej oddziałują elektrostatycznie, niż gdyby się znajdowały w środowisku wodnym, mającym dużą przenikalność elektryczną. W rezultacie odgrywają one dużą rolę w utrzymywaniu struktury przestrzennej białka (np. wtedy, gdy wiązanie wodorowe jest chronione przez bezwodne środowisko), a także w jego właściwościach katalitycznych (gdy w bezwodnym obszarze wiązania substratu znajduje się nie zjonizowana grupa —COOH, mogąca katalizować przekształcenie kowalencyjne substratu, o czym niżej). Apolarne, lecz powierzchniowo ulokowane łańcuchy boczne dążą do uzyskania bardziej bezwodnego otoczenia przez tworzenie kompleksów z innymi białkami czy z substancjami, znajdującymi się w otaczającym roztworze, a mającymi cechy hydrofobowe. Siła takiego wiązania (stała stabilności kompleksu) uzależniona jest od stopnia dostosowania (komplementarności) powierzchni hydrofobowej białka do substancji wiązanej (zwanej ogólnie ligandem). Jest to podłożem molekularnym najważniejszej biologicznie funkcji białka — jego zdolności do rozpoznawania swoistych dla niego ligandów w otoczeniu. Właściwość ta jest podłożem wybiórczości w działaniu białek (np. receptorów, enzymów, transporterów), a także możliwości samoorganizacji białek do wysoko spolimeryzowanych superstruktur molekularnych.

zdolność do rozpoznawania swoistych ligandów

Rodzaje białek

Białka — elastyczny szkielet organizmów

Zrąb organizmów żywych tworzy subtelne utkanie, w skład którego wchodzą białka, wielocukry, czasem również lipidy oraz krystaliczne składniki mineralne (np. kości). Białka zwane podporowymi lub strukturalnymi tworzą osnowę zrębu: utrzymują jego organizację, nadają mu spoistość i sprężystość, w ich sieci ulokowane są komórki produkujące poszczegól-

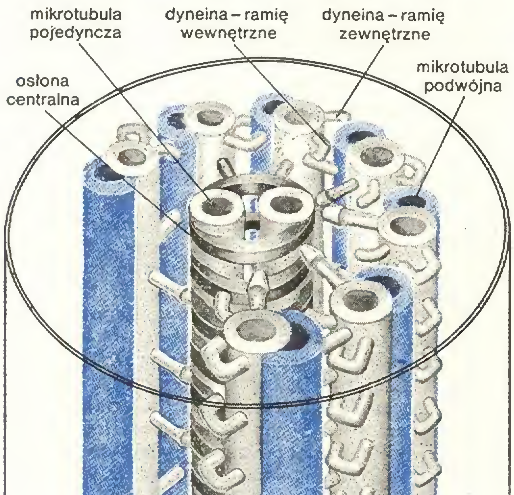
białka strukturalne

ne składniki zrębu. Jednakże rola białek strukturalnych nie jest tylko pasywna, podtrzymująca. Organizm jest plastyczny, a w jego ruchach białka strukturalne często są składnikiem czynnym. Wreszcie — oddziaływanie białek strukturalnych z różnymi składnikami ustroju, lokalnymi czy dopływającymi z krwią, sprawia, że stanowią one istotny element w integracji organizmu.

Białka strukturalne z reguły są białkami polimerycznymi. Zbudowane są z dużej liczby podjednostek, połączonych ze sobą głównie wiązaniami niekowalencyjnymi (często polimer jest dodatkowo wzmocniony wiązaniami kowalencyjnymi, np. mostkami dwusiarczkowymi). Utworzenie polimeru jest procesem spontanicznym: podjednostki (monomery lub gdy najmniejsza zdolna do samoorganizacji podjednostka nie jest pojedynczym łańcuchem polipeptydowym — protomery) łączą się ze sobą (w sposób odwracalny) tak, że ostatecznie utworzona superstruktura molekularna jako całość znajduje się w minimum energetycznym i jest trwała. Układ stabilizujących ją wiązań określa stosunki przestrzenne między podjednostkami (tzw. strukturę czwartorzędową białka), a ponieważ białka polimeryczne budowane są z identycznych lub prawie identycznych podjednostek, wykazują wysoką regularność swojej organizacji wewnętrznej. Budowa podjednostek determinuje zarówno budowę białka polimerycznego, jak jego funkcję.

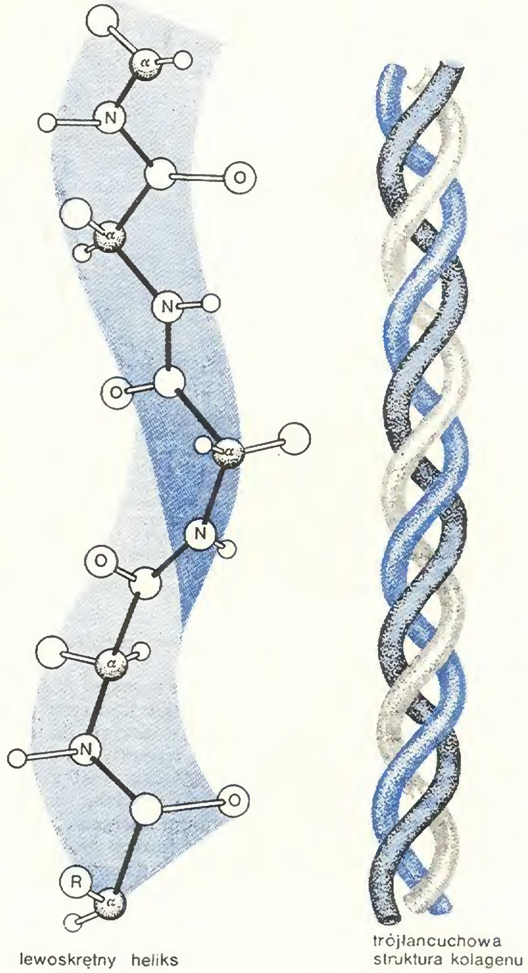
we wnętrzu komórek — tubulina

Wewnętrzny szkielet komórek tworzy siateczka śródplazmatyczna (→ Błony komórkowe) oraz — w komórkach zawierających jądro, roślinnych i zwierzęcych — system delikatnych rureczek, zwanych mikrotubulami (il. 114, tabl. 28). Mają one charakter tworów pojedynczych lub podwójnych, długości ułamków mikrometra, choć czasem znacznie dłuższych. W aparacie ruchowym komórki, np. w witkach plemników, mikrotubule tworzą precyzyjny (i spotykany w innych białkach strukturalnych) splot „9+2" (rys. 12), sprzęgnięty z cząsteczkami enzymu, dostarczającego energii koniecznej do wykonywania pracy mechanicznej. Mikrotubula zbudowana jest z podjednostek, małego globularnego białka zwanego tubuliną (il. 116, tabl. 28), ułożonego helikalnie w ścianie mikrotubuli. Helikalny układ podjednostek sprawia, że dwa końce mikrotubuli różnią się od



Rys. 12. Aparat ruchowy komórki. Na powierzchni wielu komórek znajdują się długie cienkie wypustki, będące w stałym ruchu. Ich głównym elementem strukturalnym i ruchowym są mikrotubule. Wypustki te stanowią aparat ruchowy komórki, np. witki plemników, lub przesuwają wzdłuż powierzchni komórki płyn, zawiesinę czy ciała obce, np. rzęski (il. 116, tabl. 28). Rzęski i witki mają podobną budowę: dwie centralne mikrotubule pojedyncze otoczone są przez 9 podwójnych mikrotubuli, tworząc tzw. splot „9+2". Liczne inne białka zapewniają trwałość całego narządu, a wśród nich znajduje się również dyneina, enzym udostępniający energię, zawartą w ATP (→ Organizacja procesów życiowych komórki) dla wykonania pracy mechanicznej. Ruch rzęski czy witki podobny jest do ruchu szybko uderzającego bata. Wywołany on jest przez kolejne przesuwanie się jednej połowy podwójnej mikrotubuli wzdłuż jej drugiej połowy, która pozostaje nieruchoma (il. 114, tabl. 28)

struktura kolagenu

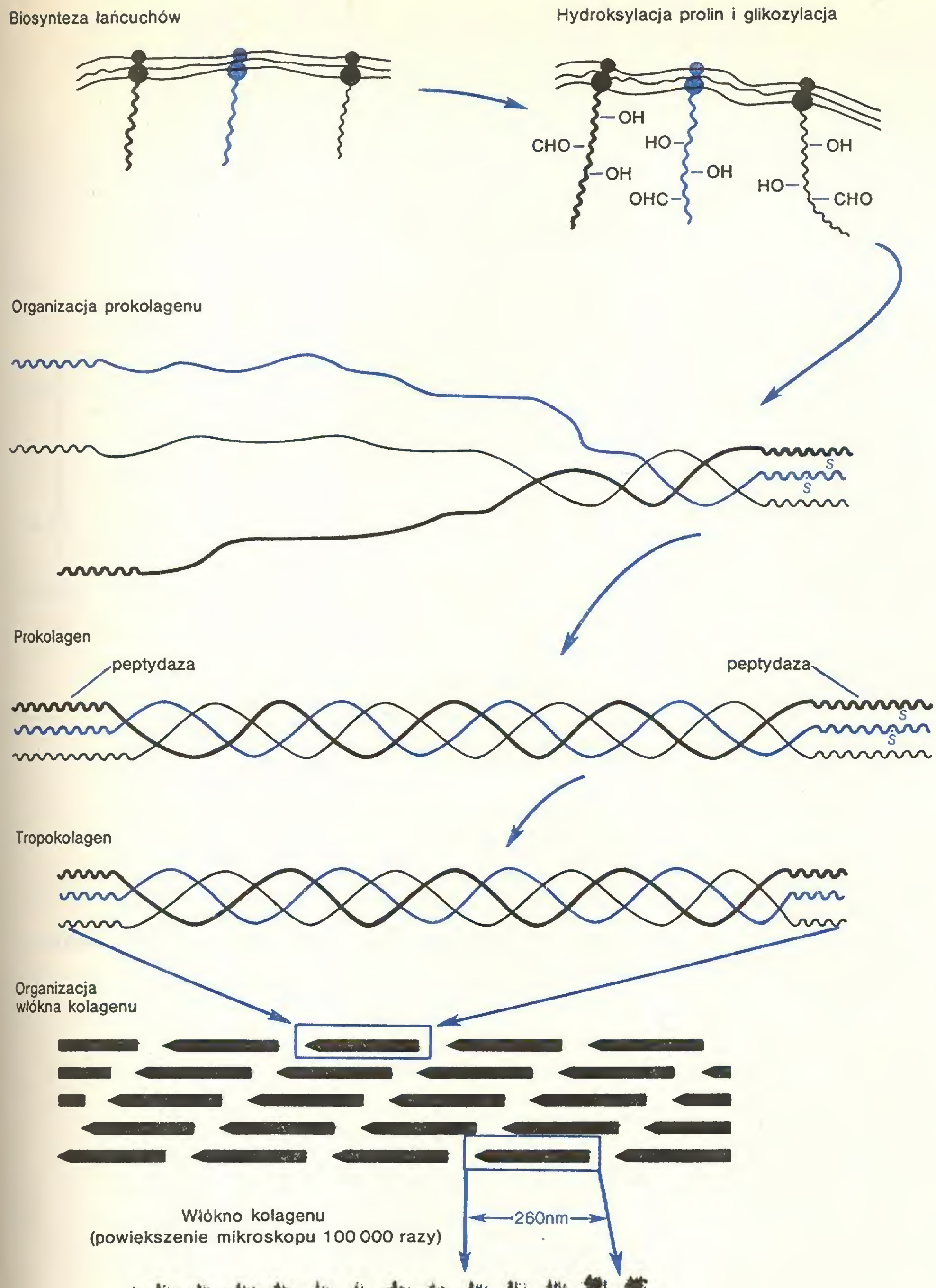


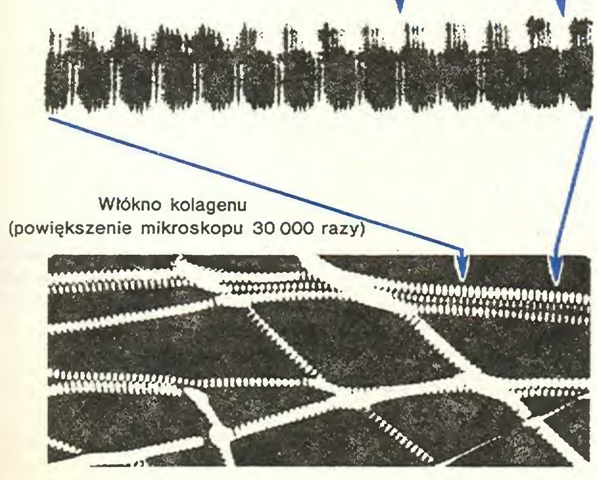
Rys. 13. Struktura kolagenu. Lewoskrętny heliks kolagenu ma 3 reszty aminokwasowe w jednym zwoju. Pochylenie zwojów jest znacznie większe niż w analogicznej strukturze α (rys. 7), w wyniku czego łańcuchy boczne nie stanowią przeszkody dla uzwojenia w lewo, ale grupy —CO i —NH skierowane są na zewnątrz heliksu i nie mogą utworzyć wewnątrzcząsteczkowych wiązań wodorowych charakterystycznych dla struktur α. Lewoskrętny heliks kolagenu może być ustabilizowany tylko wtedy, gdy trzy podobnie ukształtowane łańcuchy polipeptydowe zbliżą się do siebie na tyle, by powstały między nimi wiązania wodorowe. Z prawej strony ukazana jest trójłańcuchowa struktura kolagenu (rysunek nie uwidacznia tego, że struktura trójłańcuchowa jest jako całość lekko skręcona prawoskrętnie)

siebie, a to jest skojarzone z jej funkcją biologiczną: mikrotubula jest kierunkowskazem i organizatorem procesów nadających komórce jej anizotropię — różnokierunkowość jej budowy i czynności. Mikrotubule formowane są w cytoplazmie, z obfitego zasobu podjednostek zawsze tam, gdzie ulega polaryzacji struktura i funkcja komórki: przy wytwarzaniu wrzeciona mitotycznego, narządów ruchu przy wydłużaniu ciała komórki, przy sekrecji enzymów hormonów.

Funkcja mikrotubuli jest bezpośrednio związana z podziałem komórki i jej różnicowaniem się; procesy te ulegają zahamowaniu przez różne czynniki (m.in. farmakologiczne), które uniemożliwiają samoorganizację wolnej tubuliny w mikrotubule (znalazło

powstawanie włókna kolagenu





Rys. 14. Powstawanie włókna kolagenu. Łańcuchy polipeptydowe kolagenu powstają na rybosomach, następnie niektóre reszty proliny są w nich przekształcone w reszty hydroksyproliny (tabela) i do łańcucha podstawiane są reszty cukrowe. Tak zmodyfikowane łańcuchy łączą się ze sobą wiązaniami dwusiarczkowymi (—S—) na jednym końcu w obszarze tzw. peptydów ustalających (faliste końce). Struktura trójłańcuchowa kolagenu ulega w ten sposób stabilizacji na długości ok. 1000 aminokwasów. Peptydy ustalające mające inny skład aminokwasowy niż reszta łańcucha (w której co trzecim aminokwasem jest glicyna) ulegają skróceniu przez swoisty enzym, peptydazę kolagenu, i cała trójłańcuchowa struktura staje się zdolna do spontanicznej polimeryzacji. Włókno kolagenu organizuje się w ten sposób, że wszystkie cząsteczki tropokolagenu układają się równolegle i w tym samym kierunku, a każda z nich jest w stosunku do sąsiednich przesunięta o $^1/_4$ swojej długości. Ta regularność budowy znajduje swój wyraz w obrazach uzyskanych w mikroskopie elektronowym (powiększenie 100 tys. razy — preparat barwiony metalem, 30 tys. razy — preparat napylany metalem)

to zastosowanie m.in. w praktyce przeszczepiania tkanek). Mikrotubule występują wyłącznie u organizmów posiadających jądra komórkowe (brak ich u bakterii); u wszystkich tych organizmów skład aminokwasowy tubuliny oraz forma i funkcja mikrotubuli są jednakowe. Wskazuje to, że tubulina powstała jako białko o decydującym znaczeniu dla przekształcenia się organizmów prostych, bezjądrzastych, w organizmy wyższe, jądrzaste i wielokomórkowe. Geny kontrolujące syntezę tubuliny są zapewne produktem szczególnie starannej selekcji, gdyż jak się wydaje, przetrwały bez istotnych zmian ponad miliard lat.

między komórkami — kolagen

Międzykomórkowy zrąb strukturalny organizmów zwierzęcych tworzy kolagen, białko stanowiące do 30% masy wszystkich białek ich ustroju. Długie, silne i dosyć sztywne włókna kolagenowe, związane z substancją podstawową (z białkami, wielocukrami, materiałem mineralnym), są osnową wszystkich rodzajów tkanki łącznej — kości, chrząstek, skóry itd. Uniwersalne znaczenie kolagenu dla budowy ciała zwierząt, w zestawieniu z jego nie spotykaną w innych białkach konformacją (rys. 13), wskazuje, że jest on wynikiem szczególnie surowej presji selekcyjnej. Z tym też zapewne jest związany niezwykle złożony proces formowania włókna kolagenowego, w którym potranslacyjna modyfikacja łańcucha polipeptydowego (modyfikacja chemiczna aminokwasów, zachodząca po zakończeniu jego biosyntezy) odgrywa istotną rolę.

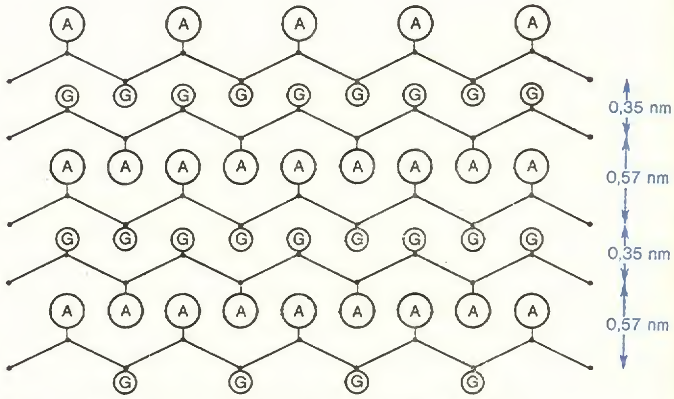
Włókno kolagenowe powstaje spontanicznie z protomerów tropokolagenu (rys. 14), jednakże uformowanie tych zdolnych do samoorganizacji podjednostek wymaga działania szeregu enzymów, mogących rozpoznać wybrane odcinki sekwencji aminokwasowej w pierwotnie syntetyzowanym łańcuchu polipeptydowym. Najpierw niektóre reszty proliny zostają przekształcone w reszty hydroksyproliny, co ułatwia lewoskrętne uzwojenie heliksu (il. 114, tabl. 28). Konformacja ta jest jednak nietrwała i jej powstanie byłoby mało prawdopodobne, gdyby końce trzech łańcuchów nie zostały połączone ze sobą (w obszarze tzw. peptydów ustalających; rys. 14). Wtedy dopiero całemu zespołowi zostaje nadany zwijający go ruch obrotowy (podobny do przędzenia na kołowrotku) o nieznanym mechanizmie, ale zapewne przy jednym końcu ustabilizowanym, i drugim — ruchomym. Wreszcie — peptydy ustalające są odcinane (przez następny swoisty enzym, rozpoznający ich sekwencję) i gotowy tropokolagen wydalany jest z komórki. W przestrzeni pozakomórkowej tropokolagen spontanicznie polimeryzuje (rys. 14) w sposób analogiczny do tego, jakim się posługuje powroźnik przy splataniu liny: ciągłość całej struktury zapewnia niewielkie przesunięcie poszczególnych jej elementów składowych w stosunku do siebie. Samoorganizacja włókna dokonuje się z ogromną precyzją: każda cząsteczka tropokolagenu jest przesunięta dokładnie o 234 ± 3 aminokwasy w stosunku do sąsiedniej. To drugie, pozornie mało prawdopodobne zdarzenie w formowaniu włókna kolagenowego umożliwione jest przez szczególną sekwencję jego łańcucha: cząsteczki tropokolagenu są właśnie o tyle przesunięte, że wytwarza się między nimi maksymalna ilość wiązań niekowalencyjnych, a także powstają wiązania kowalencyjne między wcześniej odpowiednio zmodyfikowanymi resztami lizyny.

osłony zewnętrzne

Osłony zewnętrzne organizmów zwierzęcych spełniają ważne, lecz mało skomplikowane zadania. Mają one też dość prostą budowę, nadającą im wyróżniające się własności mechaniczne i znaczną na ogół odporność na czynniki środowiskowe. Ich łańcuch polipeptydowy w całości lub w znacznej części ma jednolitą konformację — albo strukturę α, albo strukturę β.

jedwab

Przykładem białka osłonowego o strukturze β jest jedwab, z którego stawonogi, a głównie owady i pajęczaki budują oprząd i gniazda oraz tworzą instrument lokomocyjny — pajęczynę. Elementem nośnym włókna jedwabiu jest fibroina, białko zbudowane z fałdowanych arkuszy o przeciwrównoległej konformacji. W łańcuchu polipeptydowym fibroiny na długich odcinkach występują powtarzające się sekwencje (Gly-Ser-Gli-Ala-Gly-Ala)$_n$. Włókno jedwabiu utworzone jest z wielu takich arkuszy nałożonych na siebie (i wtopionych w serycynę — białko o nieregularnej konformacji, zwane klejem jedwabnym). W rezultacie — we włóknie występują regularne zgrupowania łańcuchów bocznych glicyny i alaniny (rys. 15)



Rys. 15. Jedwab — przekrój pionowy przez nakładające się na siebie pofałdowane („plisowane") arkusze fibroiny (struktura β), model otrzymany na podstawie badań krystalograficznych. Regularne rozmieszczenie powtarzających się wzdłuż łańcucha głównego reszt — alaniny (A) i glicyny (G) — prowadzi do utworzenia krystalicznej struktury białkowej

o charakterze krystalicznym; zgrupowania te nadają mu miękkość i giętkość oraz typowy dla jedwabiu piękny połysk.

Tam, gdzie sprężystość ma być główną cechą materiału osłonowego, budowany jest on z elastycznych α-heliksów. Sposób jednak, w jaki się formuje superstruktura molekularna o konformacji α, nadaje jej również i inne pożądane cechy, a zwłaszcza odporność na zerwanie, jak to ilustruje przykład α-keratyny (rys. 16). Na najniższym szczeblu samoorganizacji powstaje protomer (protofibryla, rys. 16a), będący superheliksem trójłańcuchowym o splocie liny. Na szczeblu następnym protofibryle asocjują do splotu „9+2" (rys. 10), konstrukcją bardziej podobnego do kabla niż liny i odpowiednio silniejszego. Dalsza zaś organizacja α-keratyny zachodzi dopiero pod wpływem potranslacyjnej modyfikacji jej łańcucha polipeptydowego: reszty cysteiny utleniane są do cystyn i tworzą mostki dwusiarczkowe, wzmacniające heliksy wzdłuż i łączące je poprzecznie ze sobą i z substancją podstawową, białkiem bogatym w cysteinę, ale nie mającym regularnej konformacji. Włókno keratynowe staje się w ten sposób odporne na zerwanie, pozostaje giętkie i elastyczne, natomiast przy bardzo dużej ilości mostków dwusiarczkowych jest twarde, jak na przykład w rogach, kopytach i paznokciach.

α-keratyna

We wszystkich białkach strukturalnych ogólna zasada ich tworzenia się jest jednakowa: spontaniczna polimeryzacja podjednostek do wielkiej superstruktury molekularnej. Znajduje też ona zastosowanie w konstruowaniu wielu innych białek, np. enzymów oligomerycznych oraz wirusów.

dlaczego samoorganizacja?

Powstaje pytanie, dlaczego przyroda wybrała ten sposób konstruowania białek zamiast syntetyzować od razu jeden bardzo długi łańcuch polipeptydowy, który by przecież mógł spełniać identyczną funkcję. Jedną z przyczyn jest niewątpliwie ogólna tendencja ewolucji do maksymalnie oszczędnego gospodarowania materiałem genetycznym: znacznie mniej DNA potrzeba do zakodowania małej podjednostki i wbudowania w nią wszystkich potrzebnych do samoorga-

włos — struktura

nizacji wskazówek niż do zakodowania olbrzymiego łańcucha polipeptydowego. Maleje także szansa błędu, jeżeli wynosi ona np. 10^{-3} na 1 aminokwas, to błąd pojawi się w łańcuchu składającym się z wielu tysięcy aminokwasów, natomiast krótkie podjednostki będą w znacznej części od błędu wolne. Co więcej — podjednostka z defektem może być łatwo pominięta przy tworzeniu superstruktury, podczas gdy wycięcie błędnego odcinka sekwencji wymaga istnienia aparatu, który może go rozpoznać, i enzymów rozszczepiających wiązania peptydowe. Wreszcie — niekowalencyjny sposób łączenia ze sobą podjednostek sprawia, że superstruktura — przynajmniej zanim zostanie ustabilizowana wiązaniami kowalencyjnymi — jest agregatem odwracalnym i jej powstawanie staje się zależne od aktualnego stężenia podjednostek, co umożliwia lepszą kontrolę procesu samoorganizacji i uzależnia go od różnych czynników ustrojowych.

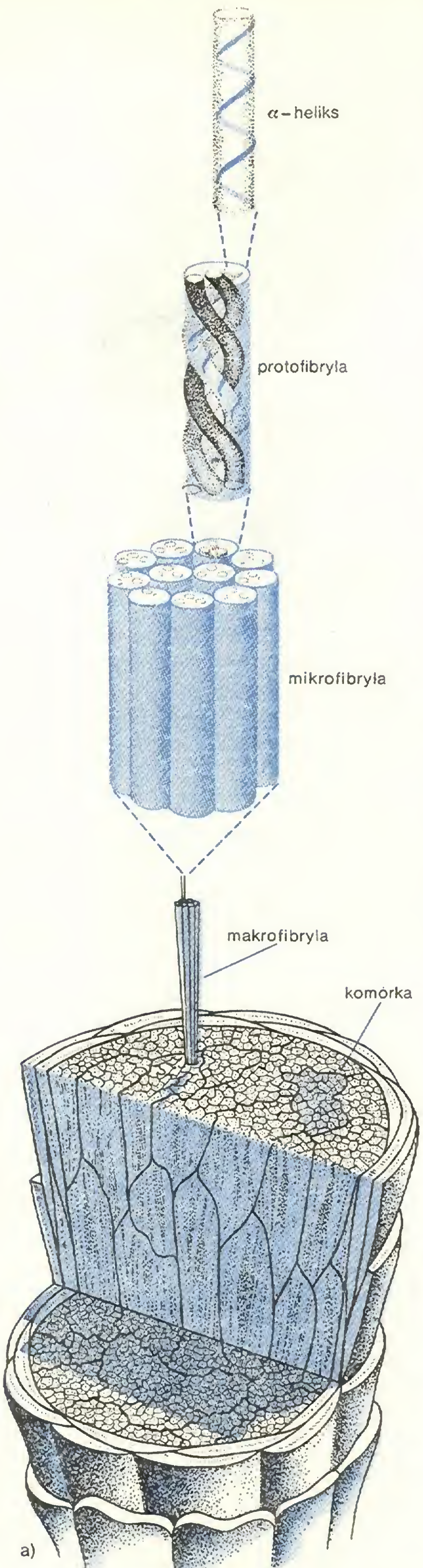
Białka — narzędzia transportu

Złożony układ błon selektywnie przepuszczalnych sprawia, że wymiana substancji z otoczeniem i ich wewnętrzne przemieszczanie w organizmach żywych nie zachodzi przez swobodną dyfuzję, ale prawie wyłącznie przy udziale białek transportowych. Białka te ułatwiają przenoszenie związków nierozpuszczalnych, np. tłuszczowców, przyspieszają zgodną z różnicą stężeń wędrówkę związków rozpuszczalnych, np. tlenu z płuc do tkanek, i umożliwiają ich przenikanie wbrew różnicy stężeń, np. jonów nieorganicznych przez błony komórkowe; często też stanowią one magazyn substancji drobnocząsteczkowych, z którego komórka czerpie w miarę potrzeby (→ Organizacja procesów życiowych komórki; zob. też Błony komórkowe).

białka transportowe

Funkcja biologiczna białka transportowego została wykształcona przez wytworzenie w nim swoistej powierzchniowej konfiguracji aminokwasów, służącej za centrum wiązania ligandu. Centrum to determinuje wybiórczość transportu albo bezpośrednio, wiążąc ligandy o pasującej do niego konfiguracji, albo pośrednio — wiążąc tzw. grupę prostetyczną, związek niebiałkowy, a często atom metalu, stanowiący właściwe miejsce przyłączania ligandu. Ligand jest wiązany niekowalencyjnie, tak że w jego wysokich stężeniach równowaga reakcji przesunięta jest na stronę kompleksu ligand–białko, a w niskich — ligand się odłącza (oddysocjowuje) i białko może ponownie służyć za transporter.

grupa prostetyczna





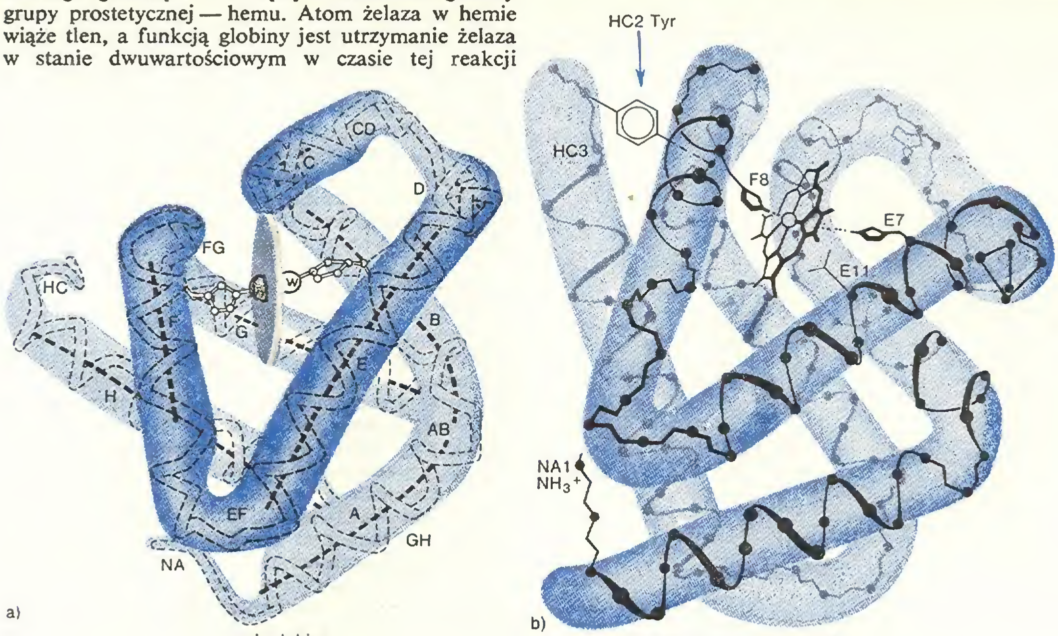
Rys. 16. Włos: a) Przejście od pierwotnie syntetyzowanego łańcucha polipeptydowego poprzez kolejne stadia organizacji cząsteczek do dojrzałego włosa. b) Fotografia włosa ludzkiego (powiększenie ok. 300 razy, mikroskop elektronowy skanningowy) wyłaniającego się z mieszka skórnego. Widoczna jest zewnętrzna powłoka włosa, utworzona przez ułożone jak łuska rybia płaskie komórki (obfitujące w keratynę), oraz złuszczające się, też bogate w keratynę, komórki naskórka (zaznaczone strzałkami)

włos — fotografia

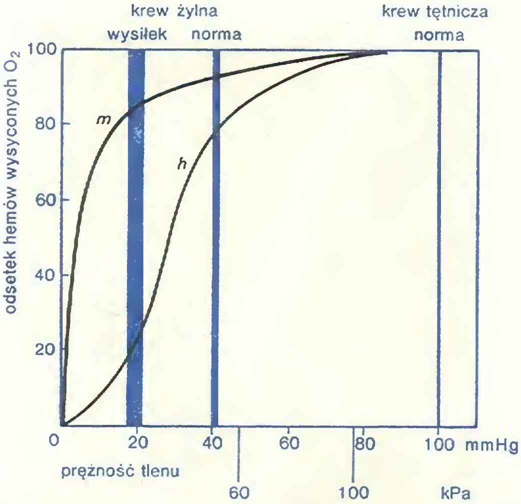
mioglobina

Przykładem prostego białka transportowego jest mioglobina (rys. 17), stanowiąca magazyn tlenu w tkance mięśniowej. Zbudowana jest ona z białka zwanego globiną i z tkwiącej w kieszonce globiny grupy prostetycznej — hemu. Atom żelaza w hemie wiąże tlen, a funkcją globiny jest utrzymanie żelaza w stanie dwuwartościowym w czasie tej reakcji (reakcja odłączenia elektronu od żelaza ma podwyższoną energię aktywacji w wyniku hydrofobowego otoczenia hemu). Na jedną globinę w mioglobinie

globina i hem



Rys. 17. Mioglobina i hemoglobina: a) budowa mioglobiny, b) budowa jednej podjednostki hemoglobiny (jej cała, złożona z 4 podjednostek cząsteczka — zob. rys. 19). Mioglobina jest białkiem przenoszącym tlen z krwi do mięśni. Hemoglobina wykształciła się w toku ewolucji z mioglobiny jako białko transportujące tlen z płuc do tkanek. Mioglobina i podjednostki hemoglobiny zbudowane są z ok. 150 aminokwasów, a ich łańcuch polipeptydowy w ok. 80% jest w konformacji α-heliksu. Zarys przebiegu łańcucha w mioglobinie pokazany jest podwójną przerywaną linią, w hemoglobinie nieco dokładniej, z zaznaczonymi atomami C^{α}. Zarys struktury trzeciorzędowej, pokazany niebieskim kolorem, wskazuje na daleko idące podobieństwo tych dwu białek. Odcinki helikalne oznaczone są literami A, B, ... (gruba przerywana linia na rys. a oznacza oś heliksu), a odcinki nieregularne, łączące heliksy, dwiema literami, np. EF — między heliksem E i F. Koniec *N* cząsteczki znajduje się przed heliksem A (NA), koniec *C* — za heliksem H (HC). W kieszonce uformowanej przez łańcuch polipeptydowy znajduje się hem, związek żelazoorganiczny (żelazoporfiryna z centralnym atomem żelaza — kulką — i czterema sprzężonymi pierścieniami pięcioatomowymi, zob. wzór w kieszonce hemoglobiny; na modelu mioglobiny hem jest zaznaczony schematycznie). Do hemu jest przyłączany tlen molekularny O_2. Atom żelaza hemu jest związany koordynacyjnie z 6 ligandami, dostarczającymi po jednej parze elektronowej do jego orbitali. Cztery z nich pochodzą z pierścieni organicznych (pyrolowych) porfiryny, jedna z histydyny tzw. bliskiej, z heliksu F (F8 na modelu hemoglobiny i wzór wpisany w model mioglobiny). Szósta para elektronowa w mioglobinie pochodzi z cząsteczki wody (W), skoordynowanej z drugiej strony z histydyną z heliksa E (histydyna „daleka"). W hemoglobinie w obecności tlenu szósta wartościowość koordynacyjna pozostaje jak się zdaje niewysycona; w obecności tlenu molekularnego O_2 para elektronowa dostarczana jest przez tlen. W mioglobinie tlen molekularny wchodzi na miejsce wody. Otoczenie hemu jest hydrofobowe, co — wraz ze sposobem wiązania hemu z łańcuchem polipeptydowym — zapewnia to, że wiązanie tlenu przez żelazo nie doprowadza do utlenienia żelaza dwuwartościowego na trójwartościowe, a to jest warunkiem spełniania przez hemoglobinę i mioglobinę ich funkcji transportowych



Rys. 18. Wiązanie tlenu kooperatywne i niekooperatywne. Wykres pokazuje zależność między ciśnieniem cząstkowym tlenu a stopniem wysycenia tlenem cząsteczek hemoglobiny i mioglobiny. Zależność ta jest hiperboliczna (krzywa *m*) w przypadku mioglobiny, która wiąże tlen niekooperatywnie, i sigmoidalna (krzywa *h*) w przypadku hemoglobiny, wiążącej tlen kooperatywnie. Wysokie wysycenie mioglobiny tlenem przy niskiej jego prężności pozwala jej służyć jako magazyn tlenu dla pracującego mięśnia nawet wtedy, gdy hemoglobina prawie całkowicie tlen oddysocjowała

przypada jeden atom żelaza, tak że może ona być albo całkowicie tlenem wysycona, albo wolna od tlenu. Wysycenie więc mioglobiny tlenem przebiega zgodnie z kinetyką reakcji bimolekularnej i jest hiperboliczną funkcją stężenia tlenu w środowisku (rys. 18).

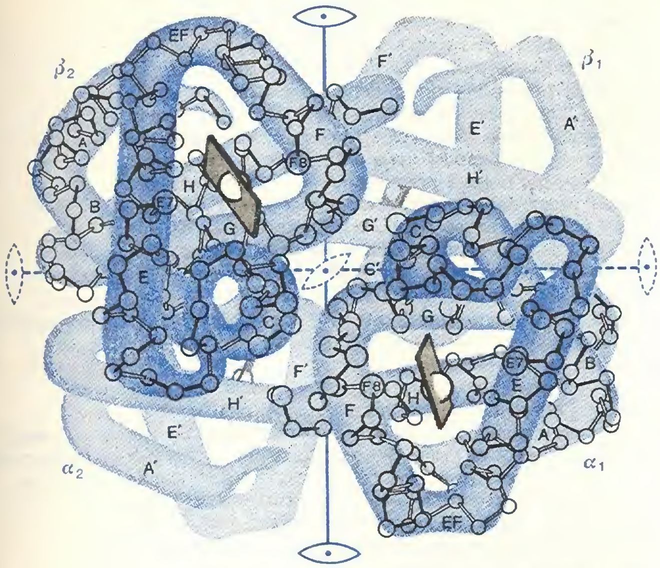
kooperatywność wiązania ligandów

W cząsteczce białka może być więcej niż jedno centrum wiązania ligandu i wówczas z reguły oddziałują one na siebie. Białka takie wykazują kooperatywność wiązania ligandów: wysycenie jednego centrum może ułatwiać (kooperatywność dodatnia) lub utrudniać (kooperatywność ujemna) wysycenie następnego. Krzywa wysycenia ma wówczas kształt esowaty (sigmoidalny), jak to ilustruje reakcja hemoglobiny z tlenem.

hemoglobina

Hemoglobina i mioglobina wykazują daleko idące podobieństwo: mają tę samą grupę funkcyjną, hem, i bardzo zbliżoną strukturę trzeciorzędową (rys. 17). Różnią się jednak tym, że mioglobina jest monomerem, a hemoglobina tetramerem, zbudowanym z 4 podjednostek (rys. 19), z których każda ma zresztą sekwencję aminokwasową, wykazującą bliskie pokrewieństwo z sekwencją mioglobiny. Hemoglobina ma cztery hemy w cząsteczce, wiąże więc cztery cząsteczki tlenu, a kooperatywność tego wiązania zależy od kontaktów między podjednostkami. Łańcuchy polipeptydowe hemoglobiny wykształciły się z łańcucha mioglobiny w wyniku duplikacji jej genu i nagromadzenia się w nim wielu punktowych mutacji (→ Kwasy nukleinowe). Nie zmienione zostały te odcinki se-

kwencji, które są odpowiedzialne za podstawową funkcję biologiczną transportera tlenu: wiązanie tlenu i ochronę dwuwartościowości żelaza. Powstały jednak nowe odcinki sekwencji sprawiające, że łańcuchy mogą się ze sobą wiązać, i to w taki sposób, że mini-



Rys. 19. Cząsteczka hemoglobiny. Hemoglobina jest tetramerem, tj. białkiem zbudowanym z czterech podjednostek, złączonych niekowalencyjnie. W jej skład wchodzą dwie podjednostki α i dwie β, mające budowę bardzo podobną do siebie i do mioglobiny (zob. rys. 17). Cząsteczka hemoglobiny ma obrys prawie kulisty, a podjednostki są w niej rozmieszczone symetrycznie (klasa symetrii 222). Na rysunku pokazano prawdziwą oś symetrii (linia ciągła) i jedną pseudooś (linia przerywana); druga pseudooś jest prostopadła do płaszczyzny papieru

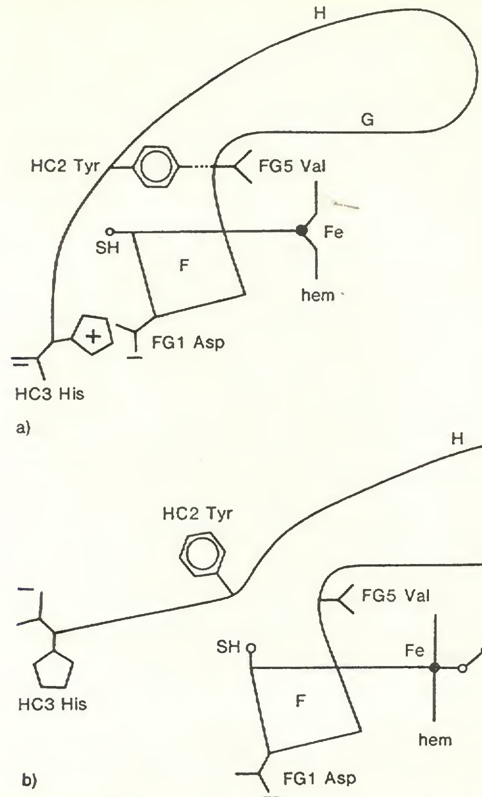
mum energetyczne osiągnięte jest w tetramerze. Reakcja z tlenem w jednej podjednostce wywołuje w niej małą zmianę konformacji (rys. 20), pociągającą jednak za sobą zerwanie niektórych kontaktów między podjednostkami. Sąsiadujące więc podjednostki znajdują sobie nową, korzystną termodynamicznie konformację, a cały tetramer uzyskuje w związku z tym nieco zmienioną strukturę czwartorzędową: ze sprężonej, charakterystycznej dla stanu beztlenowego przechodzi w rozluźnioną, właściwą dla wiązań z tlenem (rys. 21). Najtrudniejsze jest wysycenie tlenem pierwszego hemu, gdyż musi zostać przełamana bariera energii stabilizacji stanu sprężonego tetrameru. Każdy jednak następny atom tlenu jest coraz łatwiej wiązany (czwarty — kilkaset razy szybciej niż pierwszy), bo tetramer przeszedł już ze stanu sprężonego w rozluźniony. Molekularny mechanizm kooperatywności polega na tym, że lokalna, a wywołana przez związanie ligandu zmiana struktury trzeciorzędowej staje się bardziej uogólnioną zmianą konformacji, w białkach podjednostkowych wyrażającą się odmienną strukturą czwartorzędową.

oddziaływanie homotropowe i heterotropowe

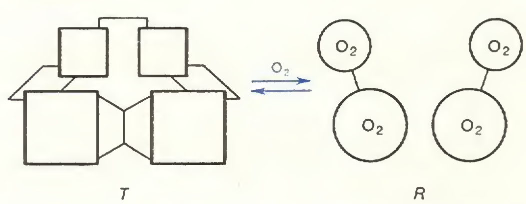
Oddziaływanie takie jak między hemami, tj. między identycznymi centrami wiązania (izosterycznymi), nosi nazwę oddziaływania homotropowego. W białkach istnieją jednak centra o odmiennej budowie (allosteryczne), wiążące inne ligandy, np. w hemoglobinie — dwutlenek węgla, jony wodorowe, kwas fosfoglicerynowy. Wpływają one na wiązanie tlenu (m.in. usztywniając sprężoną strukturę hemoglobiny). Oddziaływanie na siebie centrów o odmiennej budowie nazywa się oddziaływaniem heterotropowym.

Specyficzna geometria centrum wiązania ligandu nadaje białkom zdolność wyszukiwania właściwego ligandu. Homotropowe oddziaływania dokładniej dopasowują reakcje asocjacji (przyłączania) i dysocjacji (odłączania) ligandu do aktualnych potrzeb. Heterotropowe oddziaływania uzależniają reakcje te od innych czynników, np. produktów przemiany materii czy substancji docierających do komórki z zewnątrz. Współistnienie w białku izo- i allosterycznych centrów czyni z niego kluczowy element w regula-

przekształcenie konformacyjne w hemoglobinie



Rys. 20. Pierwszy akt przekształcenia konformacyjnego w hemoglobinie. Schemat pokazuje zmianę stosunków przestrzennych i wiązań niekowalencyjnych w hemoglobinie, zachodzą w wyniku związania przez nią tlenu molekularnego. Oznaczenia heliksów i niektórych aminokwasów, biorących udział w przekształceniu, jak na rys. 17. W nieobecności tlenu (a) atom żelaza (Fe) znajduje się nieco poza płaszczyzną pierścienia porfirynowego. Wiążąc tlen molekularny (O_2) żelazo przechodzi w stan niskospinowy i wsuwa się w płaszczyznę porfiryny (b). Pociąga to za sobą histydynę „bliską", w wyniku czego heliks F odsuwa się od heliksu H i pęka wiązanie wodorowe utrzymujące tyrozynę (HC2) w kieszonce między heliksem H i F. To z kolei osłabia drugie wiązanie (solne, między histydyną HC3 i kwasem asparaginowym FG1), utrzymujące pozycję końca *C* podjednostki. Słabną wówczas kontakty między podjednostkami i ostatecznie cała struktura podjednostkowa tetrameru (jego struktura czwartorzędowa) ulega rozluźnieniu, co ułatwia wiązanie następnych cząsteczek tlenu (zob. rys. 21)



Rys. 21. Przekształcenie konformacyjne tetrametrycznej hemoglobiny. W hemoglobinie w nieobecności tlenu podjednostki mają charakterystyczną strukturę trzeciorzędową (symbolicznie oznaczoną jako kwadraty: małe — podjednostki α, duże — β). Są one połączone ze sobą wieloma wiązaniami (cienkie kreski) a całość struktury czwartorzędowej jest wzmocniona przez efektor allosteryczny, związany między podjednostkami α i β. Związanie tlenu prowadzi do zerwania części wiązań wewnątrzcząsteczkowych (rys. 20), w konsekwencji czego zmienia się struktura trzeciorzędowa podjednostki, która tlen związała (symbolizuje to zmiana kształtu kwadratowego na okrągły). Pociąga to za sobą rozluźnienie kontaktów między podjednostkami i ostatecznie cały tetramer przechodzi z formy „sprężonej", *T* (od ang. *tense*) w formę „rozluźnioną", *R* (od ang. *relaxed*). Zmiana struktury czwartorzędowej zachodzi głównie jako przesunięcie jednego dimeru $\alpha\beta$ w stosunku do drugiego, zmiany stosunków przestrzennych między α i β są nieznaczne

cjach metabolicznych zarówno wtedy, gdy funkcja wiązania występuje w nim w stanie czystym (np. w białkach transportowych), jak i wtedy, gdy sprzężona jest ona z inną funkcją, np. katalityczną (jak w enzymach).

Białka — narzędzia regulowanej katalizy

enzymy

Wczesnym i krytycznym stadium katalizy, tym, które umożliwiło wykształcenie się wszelkich form życia, było utworzenie enzymów, białek o funkcji katalitycznej. Enzymy, jak wszystkie katalizatory, przyspieszają reakcje chemiczne, lecz same nie są w nich zużywane. Działają one jednak sprawniej niż sztuczne katalizatory, a cechą, która je uczyniła uniwersalnym narzędziem katalizy biologicznej, jest wysoka wybiórczość ich działania oraz możliwość uzależnienia aktywności od stanu komórki i od rozmaitych substancji w niej obecnych, m.in. produktów przemiany materii. Gdyby tak nie było, metabolizm stałby się chaotyczny, organizm nie mógłby sprawować kontroli nad zachodzącymi w nim reakcjami chemicznymi i niemożliwa stałaby się celowa odpowiedź na stale zmieniające się warunki otoczenia. Dlatego też wszystkie właściwie reakcje chemiczne w organizmie żywym przebiegają przy udziale enzymów, niezależnie od tego, czy są one w warunkach fizjologicznych szybkie czy powolne. Komórka zawiera od tysięcy (bakterie) do milionów (organizmy wyższe) różnych enzymów, co stanowi przeszło połowę wszystkich jej białek. Każdy enzym jest wyspecjalizowany co do reakcji, jaką może katalizować, oraz co do substratu (czy grupy podobnych substratów), na który może działać. Tabela przedstawia nazwy i podstawowe własności najważniejszych dotąd poznanych klas enzymów.

Enzymy

klasy enzymów

Klasa 1: Oksydoreduktazy (ok. 2000 enzymów) — katalizują odszczepienie atomu wodoru lub elektronu od jednych związków i przeniesienie ich na drugie. Większość utlenień biologicznych rozpoczynają dehydrogenazy odłączające atom wodoru od grup alkoholowych (—CHOH), aldehydowych (—CHO), kwasowych (—COOH), aminowych ($—NH_2$), tiolowych (—SH) i wielu innych, stanowiących donory wodoru. Akceptorami mogą być liczne koenzymy i inne związki. Dehydrogenazy, dla których akceptorem wodoru jest tlen molekularny O_2, nazywane są oksydazami. W transporcie elektronów biorą też udział rozmaite białka, zwłaszcza cytochromy i inne metaloproteiny (np. zawierające atomy miedzi, molibdenu), a także flawoproteiny, białka z wiązaniami —S—S—.

Klasa 2: Transferazy (ok. 800 enzymów) — katalizują przenoszenie grup chemicznych z jednych związków na drugie. Przenoszona może być grupa metylowa ($—CH_3$), aldehydowa, kwasowa, aminowa, reszty kwasu fosforowego, aminokwasów, cukrów, nukleotydów i in.

Klasa 3: Hydrolazy (ok. 900 enzymów) — katalizują rozszczepienie wiązań przy udziale cząsteczki wody, czyli hydrolizę. Działają na wiązania —C—O—, —C—N—, —C—C—, —C—Cl, —C—P—, —P—N—, —S—N—, —C—P—, np. esterazy hydrolizują wiązanie estrowe (—C—O—), peptydazy — peptydowe (—C—N—).

Klasa 4: Liazy (ok. 300 enzymów) — katalizują wyłączanie grupy chemicznej ze związku, pozostawiając w nim wiązanie podwójne. Działają na wiązania —C—C—, -C—O—, —C—N—, —C—S— i inne.

Klasa 5: Izomerazy (ok. 100 enzymów) — katalizują wewnętrzne przekształcenie związku, strukturalne lub geometryczne, np. przekształcają izomer L w D, *cis* w *trans* (rys. 4), przenoszą w obrębie danego związku wiązania podwójne (—C=C—), dwusiarczkowe (—S—S) i in.

Klasa 6: Ligazy, zwane również syntetazami (ok. 70 enzymów) — katalizują połączenie dwu cząsteczek w nowy związek. Tworzą one m.in. wiązania —C—O— (np. łączenie aminokwasu z tRNA), —C—S— (np. w działaniu koenzymu *A*) —C—C—, —P—O— (np. enzymy naprawiające uszkodzenia w kwasach nukleinowych).

Enzymy są białkami o masach cząsteczkowych od tysięcy do kilku milionów daltonów. Część z nich (głównie enzymy wydzielane z komórek na zewnątrz) jest białkami monomerycznymi, zbudowanymi z jednego łańcucha polipeptydowego. Większość jednak są to białka oligomeryczne, złożone z kilku (a czasem więcej) podjednostek. Masy cząsteczkowe enzymów monomerycznych i podjednostek enzymów oligomerycznych mieszczą się w tych samych granicach — od kilkunastu do ok. 80 tysięcy daltonów.

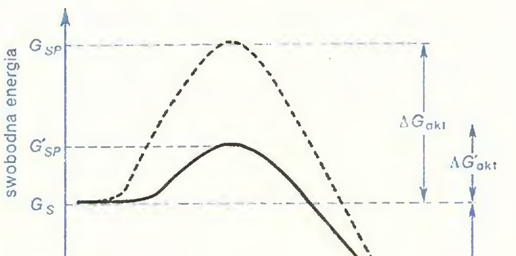
Wszystkie enzymy mają wybiórczo działające centra wiązania różnych ligandów. Funkcja biologiczna enzymu określona jest przez to centrum, z którym sprzężone są grupy katalitycznie czynne i które wraz z nimi tworzy tzw. obszar czynny enzymu. W obszarze tym dokonuje się reakcja chemiczna, przyspieszana przez dany enzym — przetworzenie substratu reakcji w jej produkt. Oprócz centrum wiązania tego liganda, który jest substratem reakcji, enzymy mają zazwyczaj jedno (lub więcej) centrum allosteryczne. Wiązane tam ligandy mają odmienną od substratu budowę i nie ulegają przemianie chemicznej katalizowanej przez dany enzym. Natomiast wysycenie centrum allosterycznego przez taki ligand wywołuje lokalną zmianę konformacyjną w enzymie, przenoszącą się w okolice obszaru czynnego i powodującą — w drodze oddziaływania heterotropowego — zmianę szybkości katalizowanej reakcji. Ligandy allosteryczne są zazwyczaj nazywane efektorami lub modulatorami reakcji enzymatycznych.

efektory reakcji enzymatycznych

Wybiórczość działania enzymów, zwana ich swoistością substratową, jest niezwykle wysoka, nieosiągalna przy pomocy katalizatorów sztucznych. Są enzymy działające na jeden tylko związek chemiczny lub na jeden z jego izomerów przestrzennych czy strukturalnych. Czasem swoistość substratowa enzymu może być szersza, ale zawsze ograniczona jest do jednego rodzaju reakcji i określonego typu wiązania chemicznego (tabela). U jej podłoża leży specyficzna geometria centrum wiązania substratu, a zwiększa się jeszcze wskutek wymagania, by podatne na katalizę wiązanie było dostępne dla grup katalitycznie czynnych, tj. tych grup chemicznych w enzymie, które dokonują zerwania wiązań kowalencyjnych w substracie i utrwalają ostateczną strukturę produktu. Drobne nawet różnice konfiguracji atomowych wokół podatnego na katalizę wiązania mogą spowodować, że związek bardzo podobny do substratu nie ulega katalizie, pomimo że jest wiązany w obszarze czynnym. Staje się on tzw. kompetytywnym inhibitorem: konkurując z substratem o centrum wiązania, uniemożliwia związanie substratu i w rezultacie hamuje reakcję. Wiele takich związków jest wytwarzanych w organizmie i odgrywają one ważną rolę w regulacji procesów

swoistość substratowa enzymów

inhibitory kompetytywne



Rys. 22. Zmiany energetyczne w przebiegu reakcji chemicznej katalizowanej i niekatalizowanej. Wykres przedstawia zmiany swobodnej energii reakcji katalizowanej (linia ciągła) i niekatalizowanej (linia przerywana). Zaznaczono poziomy swobodnej energii dla substratu (*S*), związku przejściowego (*SP*) i produktu (*P*) jako funkcję postępu reakcji chemicznej. Opisana wykresem reakcja jest spontaniczna, gdyż poziom swobodnej energii substratu jest wyższy niż produktu. Aby jednak substrat przekształcił się w produkt, musi powstać związek pośredni, co wymaga doprowadzenia odpowiedniej ilości energii (energia aktywacji G_{akt}). Przy wysokiej barierze energii aktywacji spontaniczna reakcja może być niezauważalna w normalnych warunkach ciśnienia i temperatury. W obecności katalizatora energia aktywacji jest mniejsza i reakcja przebiega ze znacznie większą szybkością

metabolicznych. Syntetycznie otrzymywane inhibitory kompetytywne znajdują szerokie zastosowanie jako leki, pestycydy i inne użyteczne preparaty.

Z wysoką wybiórczością działania enzymów kojarzy się ich druga cecha, odróżniająca je od sztucznych katalizatorów — zdolność do ogromnego przyspieszania reakcji chemicznych. Reakcje katalizowane przez enzymy przebiegają 10^8 i więcej razy szybciej niż w obecności katalizatora sztucznego. Połączenie tych dwu cech sprawia, że organizmy są w stanie przeprowadzać w stosunkowo niskiej temperaturze i przy normalnym ciśnieniu reakcje, dla których technologia współczesna musi stosować bardzo drastyczne i obciążające naturalne środowisko warunki (np. katalityczna redukcja azotu powietrza do amoniaku).

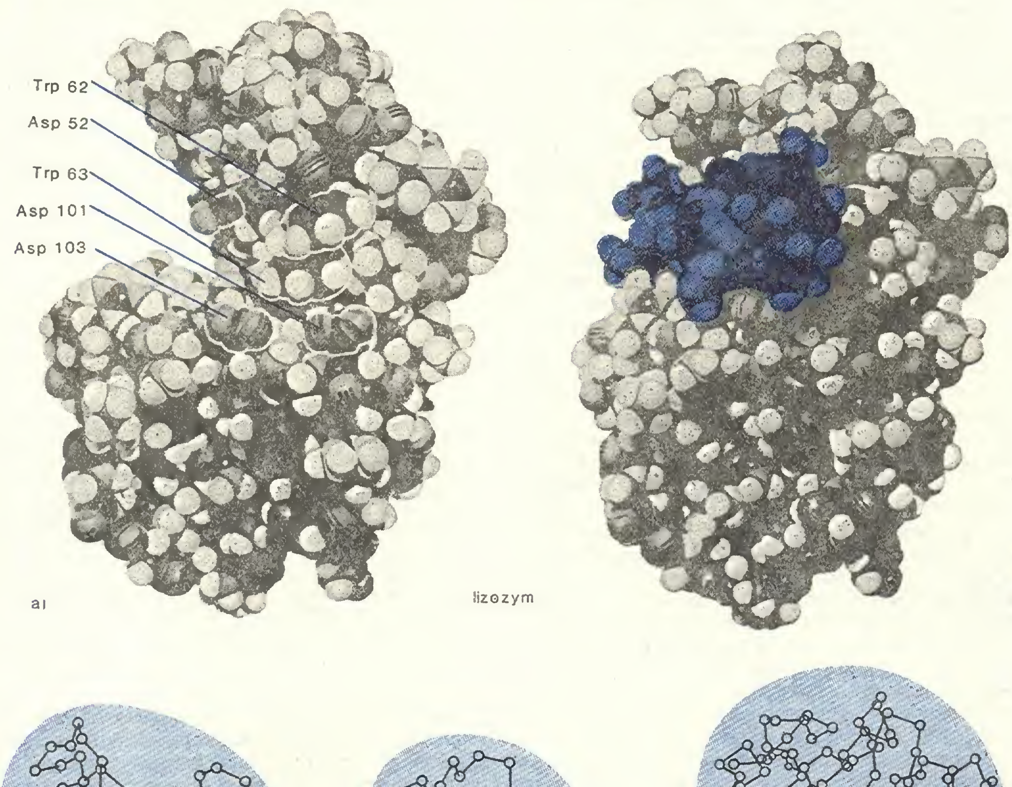
kataliza

W reakcji chemicznej cząsteczka związku ulega przekształceniu kowalencyjnemu, w wyniku którego powstaje produkt reakcji, związek o innej niż substrat budowie, różniący się od niego systemem wiązań kowalencyjnych, a na ogół także liczbą i rodzajem atomów. (W odróżnieniu od tego przekształcenie konformacyjne nie wymaga zajścia reakcji chemicznej, gdyż nie musi się zmienić struktura kowalencyjna danego związku). Aby doszło do reakcji chemicznej, cząsteczka substratu musi się zderzyć niesprężyście np. z inną cząsteczką. Powstaje wówczas tzw. związek przejściowy o budowie pośredniej między substratem a produktem, nie będący zresztą związkiem chemicznym w zwykłym tego słowa znaczeniu, gdyż jest bardzo nietrwały i szybko rozpada się z wytworzeniem produktu czy produktów reakcji. Do zderzeń niesprężystych zdolne są tylko te cząsteczki substratu (zwane aktywowanymi), których energia kinetyczna jest wyższa od średniej energii substratu o wartość określoną jako energia aktywacji (rys. 22). A ponieważ ich udział w całości w warunkach normalnych jest bardzo niewielki (np. 10^{-14}), reakcje chemiczne, nawet te, które są spontaniczne, tj. przebiegają z ubytkiem swobodnej energii, zachodzą bardzo powoli, a czasem są w ogóle niezauważalne.

powstawanie związku przejściowego

Katalizatory przyspieszają reakcje chemiczne dzięki temu, że spośród wielu możliwych dróg od substratu

modele atomowe enzymów



Rys. 23. Modele atomowe enzymów: a) Model cząsteczki lizozymu (zob. rys. 24), enzymu rozszczepiającego niektóre wielocukry, wykonany z modeli atomów uwzględniających promienie van der Waalsa. Z lewej strony — lizozym z jego charakterystyczną szczeliną, mieszczącą obszar czynny; zaznaczono kilka aminokwasów, odgrywających istotną rolę w katalizie (rys. 24). Z prawej — ten sam model lizozymu z substratem (zaznaczonym na niebiesko) wypełniającym szczelinę. Gdy substrat wchodzi w obszar czynny, brzeg szczeliny nieco się na niego nasuwa, a tryptofan (Try 62) przesuwa się o 75 nm (na rysunku niewidoczne). b) Rozmieszczenie atomów łańcucha głównego w trzech enzymach monometrycznych; kolorem niebieskim zaznaczony jest ogólny obrys cząsteczki

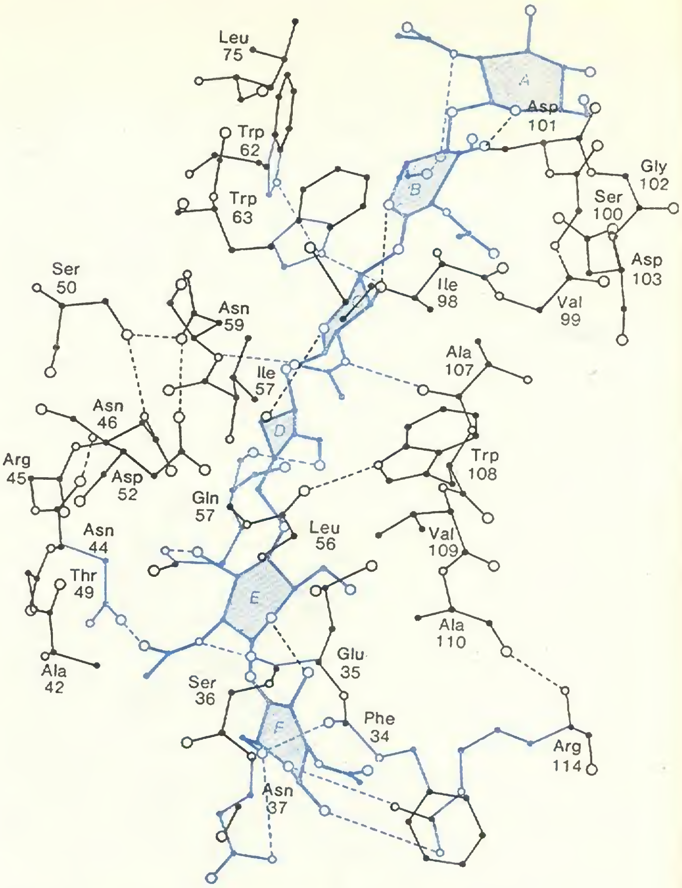
działanie katalizatora

do produktu wybierają tę, która jest najbardziej korzystna energetycznie. Ich działanie polega na obniżeniu energii aktywacji, charakterystycznej dla danej reakcji (rys. 22). Katalizator wiąże aktywowaną cząsteczkę substratu na swojej powierzchni w ten sposób, że oddziaływanie między grupami czynnymi katalizatora i substratu staje się ułatwione. W rezultacie dochodzi między nimi do termodynamicznie korzystnej reakcji chemicznej, stanowiącej właściwy akt katalityczny: zerwaniu ulegają wiązania stare (substratu) i powstają nowe, właściwe dla produktu. Produkt się oddziela, a katalizator ponownie może wejść w reakcję z substratem. Udostępniona w akcie katalitycznym swobodna energia przyczynia się w znacznym stopniu do obniżenia energii aktywacji katalizowanej reakcji.

Akt katalityczny w enzymach, przynajmniej tam, gdzie to było dokładnie zbadane, wydaje się przebiegać tak samo jak w katalizatorach sztucznych. Najczęściej jest to reakcja nukleofilna, czasem elektrofilna, niejednokrotnie prowadząca do powstania kowalencyjnego, nietrwałego związku między substratem a katalizatorem. Niemniej efekt działania katalizatora sztucznego, mierzony wielkością energii aktywacji, jest znacznie słabszy niż enzymu. W warunkach normalnych np. nadtlenek wodoru („perhydrol") rozpada się spontanicznie na wodę i tlen z energią aktywacji ok. 72 kJ/mol (18 kcal/mol); w obecności katalizatora sztucznego, czerni platynowej, energia aktywacji wynosi ok. 50 kJ/mol, a w obecności katalazy, enzymu swoistego dla reakcji rozkładu nadtlenku wodoru, energia aktywacji wynosi zaledwie 8 kJ/mol (szybkość reakcji rozkładu w obecności katalazy jest ponad 10^8 razy wyższa niż przy czerni platynowej). Wobec podobieństwa mechanizmu aktu katalitycznego w enzymach i w katalizatorach sztucznych sam akt nie może być wystarczającym tłumaczeniem znacznie większej sprawności enzymów.

Katalizatory sztuczne wymagają znacznego dopływu energii, by można było osiągnąć należytą szybkość katalizowanej reakcji, gdyż zachodzi konieczność aktywowania bardzo dużej liczby cząsteczek substratu, niezbędnych do zapoczątkowania reakcji z katalizatorem. Enzymy nie mogą funkcjonować w tak drastycznych warunkach, a wręcz przeciwnie — musiały być tak wykształcone, by miały dużą sprawność w środowisku, w jakim się mogła toczyć ewolucja materii żywej — w niskiej temperaturze, normalnym ciśnieniu, w rozcieńczonym roztworze substratów. Elementem struktury białka, który sprawił, że pierwotnie (w ewolucji prebiotycznej, zanim się pojawiły organizmy) istniejące katalizatory chemiczne, np. metale i związki metaloorganiczne, przekształciły się w enzymy, było prawdopodobnie utworzenie w białkach, w bezpośredniej bliskości grupy katalitycznie czynnej, centrum wiązania substratu. Centrum to, w dzisiejszych enzymach komplementarne raczej w stosunku do stanu przejściowego substratu niż podstawowego, wiąże substrat silnie i może go wychwytywać nawet przy bardzo jego niskim stężeniu w środowisku. Prawdopodobieństwo reakcji np. między dwiema cząsteczkami jest w takim roztworze bardzo małe. Gdy jednak zostaną one upakowane w centrum wiązania, to tracą swobodę ruchu (rotacyjnego i translacyjnego) i znajdują się w bezpośredniej bliskości ze sobą i z enzymem, szybkość więc reakcji bardzo wzrasta. Wskutek posiadania swoistego centrum wiązania substratu enzym staje się niejako maxwellowskim „demonem" entropii: wprowadza chwilowy stan wysokiego uporządkowania materii bez wydatkowania (netto) energii. Jest to ważny czynnik w działaniu enzymów, być może krytyczny we wczesnym okresie ewolucji. Sam przez się nie wystarcza jednak do wyjaśnienia całego ogromnego przyspieszenia reakcji przez enzym. Większe znaczenie ma tutaj zdolność enzymu do wykorzystania energii wiązania substratu dla umożliwienia przekształcenia konformacyjnego i substratu, i enzymu. Zjawisko to zostało ujawnione w ostatnim dziesięcioleciu dzięki badaniom rentgenowskim nad krystalicznymi enzymami i ich kompleksami z substratem (a ściślej — z inhibitorami kompetytywnymi, analogami substratu). Badania te w połączeniu z badaniami biochemicznymi i rozważaniami ewolucyjnymi dały po raz pierwszy obraz — może jeszcze niedoskonały, ale wewnętrznie spoisty — mechanizmów molekularnych w działaniu enzymów.

enzymy w procesie ewolucji



Rys. 24. Substrat w obszarze czynnym enzymu. Schemat pokazuje lokalizację substratu (zbudowanego z 6 reszt cukrowcowych) w obszarze czynnym lizozymu (rys. 23). Sześć pierścieni cukrowcowych (nieregularne sześciokąty niebieskie) jest złączonych z wieloma atomami reszt aminokwasowych lizozymu wiązaniami wodorowymi (linie kreskowane). Podobszary wiązania dwu skrajnych reszt z każdego końca wielocukru utrzymują całą jego cząsteczkę w rozprostowanej formie. Dla trzeciego (od góry) i czwartego pierścienia cukrowca nie ma dosyć miejsca w obszarze czynnym, są one skręcone (ustawione są niemal prostopadle do płaszczyzny papieru) co wprowadza do ich wiązań naprężenia, zwłaszcza do pierścienia czwartego od góry. Grupy polarne silnie elektroujemnych aminokwasów, Asp 52 i Glu 35 znajdujące się w pobliżu wiązania między czwartym a piątym pierścieniem, zrywają to wiązanie i wielocukier rozpada się na dwie części.

obszar czynny enzymu

Obszar czynny enzymu jest ulokowany w zagłębieniu (szczelinie, kieszonce) powierzchni jego cząsteczki (widoczne na rys. 23). W obszarze tym istnieje zazwyczaj kilka podobszarów (rys. 24), komplementarnych w stosunku do poszczególnych części substratu lub przyjmujących różne substraty, gdy enzym działa więcej niż na jeden substrat jednocześnie (większość reakcji enzymatycznych jest dwusubstratowa, choć często dla uproszczenia pomija się, że np. woda jest też substratem w wielu reakcjach). Specyficzna geometria obszaru czynnego zmusza substrat do przyjęcia konformacji odmiennej od tej, którą miał w stanie wolnym w roztworze. Wprowadzone zostają w ten sposób naprężenia do jego cząsteczki: ulegają zmianie kąty wartościowości, odległości atomowe, maleje energia stabilizacji rezonansowej niektórych wiązań. Cząsteczka się zniekształca, na ogół geometrycznie, ale

często również elektrostatycznie, i przybliża do formy związku przejściowego, w której jest silniej związana z enzymem. Energia wiązania substratu jest też wykorzystywana w lokalnych zmianach konformacji enzymu: katalitycznie czynne grupy (w enzymach, w odróżnieniu od katalizatorów sztucznych — często jednocześnie elektro- i nukleofilne) zbliżają się do strategicznie ważnych miejsc substratu, przede wszystkim tych, w których w stanie przejściowym pojawią się formalne lub częściowe ładunki. Łatwo już wtedy zajdzie właściwy akt katalityczny: ostateczna polaryzacja elektronów w nietrwałym związku przejściowym, prowadząca do wytworzenia trwałych wiązań kowalencyjnych produktu reakcji.

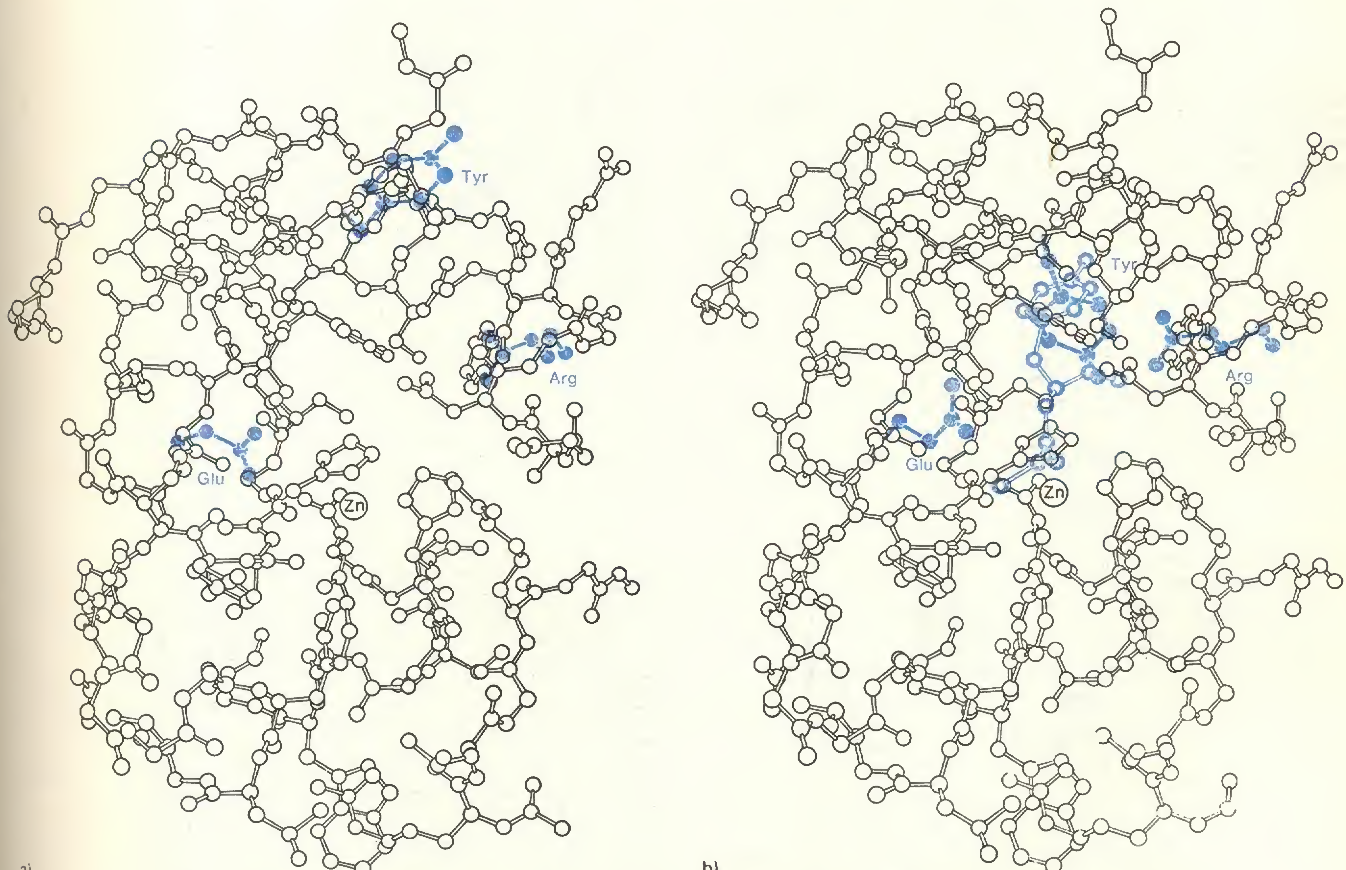
Nasilenie zmian konformacyjnych w substracie i w enzymie bywa różne w rozmaitych reakcjach enzymatycznych. Gdy substrat ma dużą swobodę zmiany swojej konformacji, jak np. wielocukier (rys. 24), to enzym może być dosyć sztywny i może ulegać tylko bardzo niewielkim przekształceniom. Wyraźne natomiast zmiany konformacji enzymu zachodzą, gdy rozrywane wiązanie jest sztywne, jak np. wiązanie peptydowe (rys. 25). Czasem zaś zmiany konformacyjne substratu są bardzo małe, natomiast duże jest przemieszczenie w nim ładunków elektrostatycznych (rys. 27).

enzymy a katalizatory techniczne

Mechanizm katalizy enzymatycznej odzwierciedla warunki, jakie biosfera stawiała, by ewolucja organizmów żywych zachodziła pomyślnie. Bezwzględne wymagania selekcji biologicznej doprowadziły do utworzenia katalizatorów, działających z niezwykłą sprawnością w bardzo wąskich granicach niskiej temperatury i w niewielkim stężeniu substratów. Człowiek w swoim empirycznym podejściu nieświadomie wykorzystał do opracowania katalizatorów technicznych jedną tylko cechę strukturalną enzymów — udział metalu w obszarze czynnym, cechę bardzo wielu enzymów i praktycznie wszystkich katalizatorów technicznych. Bogactwo jednak rozwiązań ewolucyjnych problemu katalizy jest znacznie większe. Wiele enzymów nie zawiera metalu w swojej cząsteczce, a ich grupy katalitycznie czynne nie wykazują — w odróżnieniu od metalu — aktywności katalitycznej, gdy są odłączone od białka. Nie ulega wątpliwości, że głębsze poznanie subtelnej struktury enzymów i molekularnych mechanizmów w ich działaniu dostarczy przesłanek do utworzenia katalizatorów technicznych „drugiej generacji", opartych na wzorach wypracowanych przez przyrodę, wydajniejszych niż dotychczasowe, nie wymagających tak dużego wkładu energii i mniej obciążających naturalne środowisko człowieka.

regulacja funkcji enzymów

Szybkość katalizy enzymatycznej, podobnie jak szybkość innych reakcji w roztworach wodnych, zależy od stężeń substratu, produktu, jonów wodorowych oraz od temperatury, potencjału oksydoredukcyjnego, lepkości i in. Czynniki te wpływają na przebieg reakcji metabolicznych, są jednak zbyt mało swoiste by zapewnić dostateczną subtelność regulacji funkcji enzymów. Duże natomiast znaczenie mają tu swoiste cechy obszaru czynnego enzymów. Enzym np. może skierować przemianę substratu w określonym kierunku, jeżeli z kilku enzymów działających na



Rys. 25. Zmiany konformacyjne enzymu w czasie katalizy. Rysunek przedstawia część (ok. $^1/_4$) łańcucha polipeptydowego karboksypeptydazy (zob. rys. 10), enzymu rozrywającego ostatnie od końca *C* wiązanie peptydowe w białkach. Na rys. a) widoczna jest szczelina prowadząca do obszaru czynnego, który na rys. b) wypełniony jest substratem (dwupeptyd Gli-Tyr, puste kółka niebieskie). Substrat jest związany swoim końcem *N* (grupą —CO glicyny) z atomem cynku (Zn), stanowiącym integralną część obszaru czynnego; koniec *C* substratu (pierścień tyrozyny) znajduje się w hydrofobowej kieszonce obszaru czynnego. Związanie substratu wywołuje znaczne przemieszczenie trzech reszt aminokwasowych, zaznaczonych pełnymi kółkami niebieskimi. Grupa hydroksylowa Tyr 248 wiąże dwoma wiązaniami wodorowymi grupę aminową substratu i grupę —NH jego łańcucha głównego. Arg 145 wiąże jego grupę karboksylową, a grupa karboksylowa Glu 270 atakuje spolaryzowane przez związanie z cynkiem wiązanie peptydowe substratu

dany substrat jeden przekształca go sprawniej niż inne. A także w wielu reakcjach metabolicznych (rys. 26) wytwarzane są produkty będące kompetytywnymi inhibitorami różnych enzymów, co stwarza możliwość sprzężenia rozmaitych ciągów reakcji enzymatycznych ze sobą. Ważnym czynnikiem regulacji dla niektórych enzymów są koenzymy, drobnocząsteczkowe

koenzymy

związki organiczne, wiązane przez enzym w jednym z podobszarów czynnego obszaru i stanowiące współsubstrat w jego działaniu katalitycznym. Koenzym wiąże się (kowalencyjnie) z produktem reakcji, w tej postaci oddysocjowuje od enzymu i jest transportowany do innych enzymów. Tam produkt zostaje odłączony od koenzymu i przetwarzany dalej, a koenzym regeneruje się i ponownie wchodzi w cykl reakcji. Zasób koenzymów w komórce jest niewielki (większość z nich stanowią witaminy i ich pochodne), tak że dostępność koenzymu dla danej reakcji enzymatycznej staje się czynnikiem kontrolującym szybkość całego ciągu przemian metabolicznych, jak np. NAD — w łańcuchu oddechowym (→ Organizacja procesów życiowych komórki).

Praktycznie nieograniczone możliwości regulacji funkcji enzymu wynikają z oddziaływań heterotropowych, związane z obecnością w nim centrów allosterycznych: uzależniają one bowiem szybkość katalizy, a nawet jej swoistość nie od substratu czy produktu,

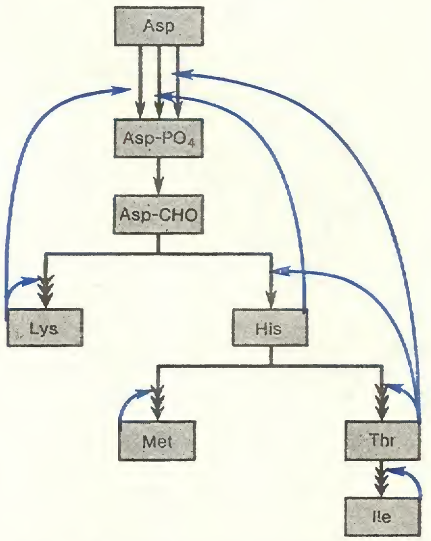
enzymy allosteryczne

ale od czynników o zupełnie innej budowie. Oddziaływania heterotropowe występują zarówno w enzymach monomerycznych, jak oligomerycznych (często nazywanych enzymami allosterycznymi), zwłaszcza jednak silnie są rozwinięte w tych ostatnich. Enzymy bowiem monomeryczne mają na ogół tylko jeden obszar czynny w swojej cząsteczce (wyjątki są nieliczne), podczas gdy w enzymach oligomerycznych może być ich tyle, ile jest podjednostek, a więc efektory reakcji wpływają na cały system oddziaływań homotropowych. Co więcej — w enzymach oligomerycznych podjednostki

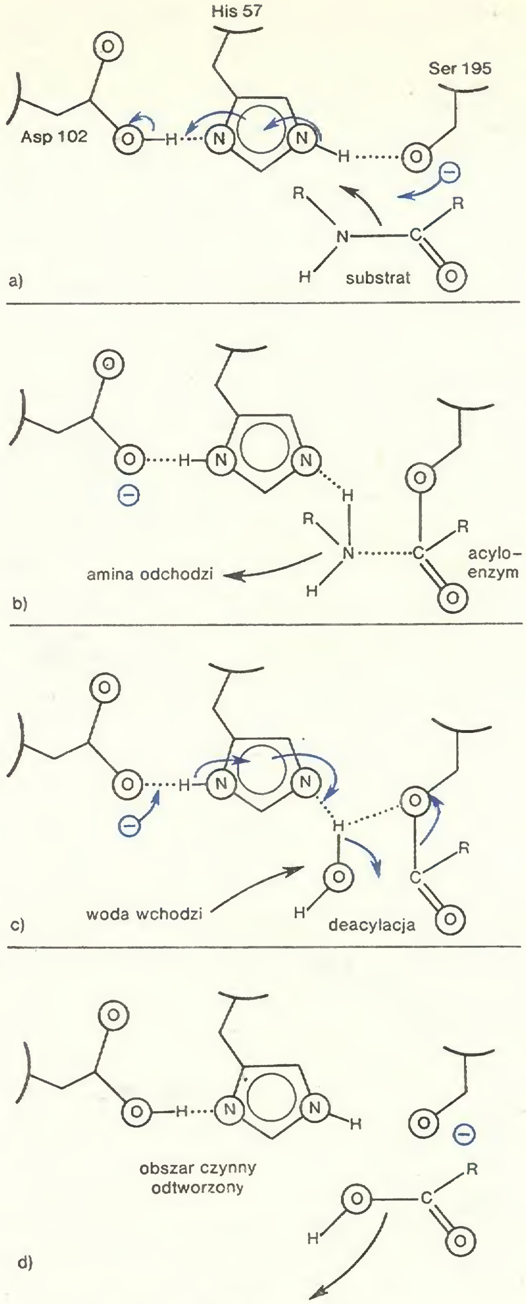
podjednostki katalityczne i podjednostki regulacyjne

nie muszą być identyczne, mogą się różnić swoistością substratową, a część z nich może być wyspecjalizowana jako podjednostki regulacyjne pozbawione funkcji katalitycznej, tylko hamujące (czy aktywujące) funkcję podjednostki katalitycznej w oligomerze.

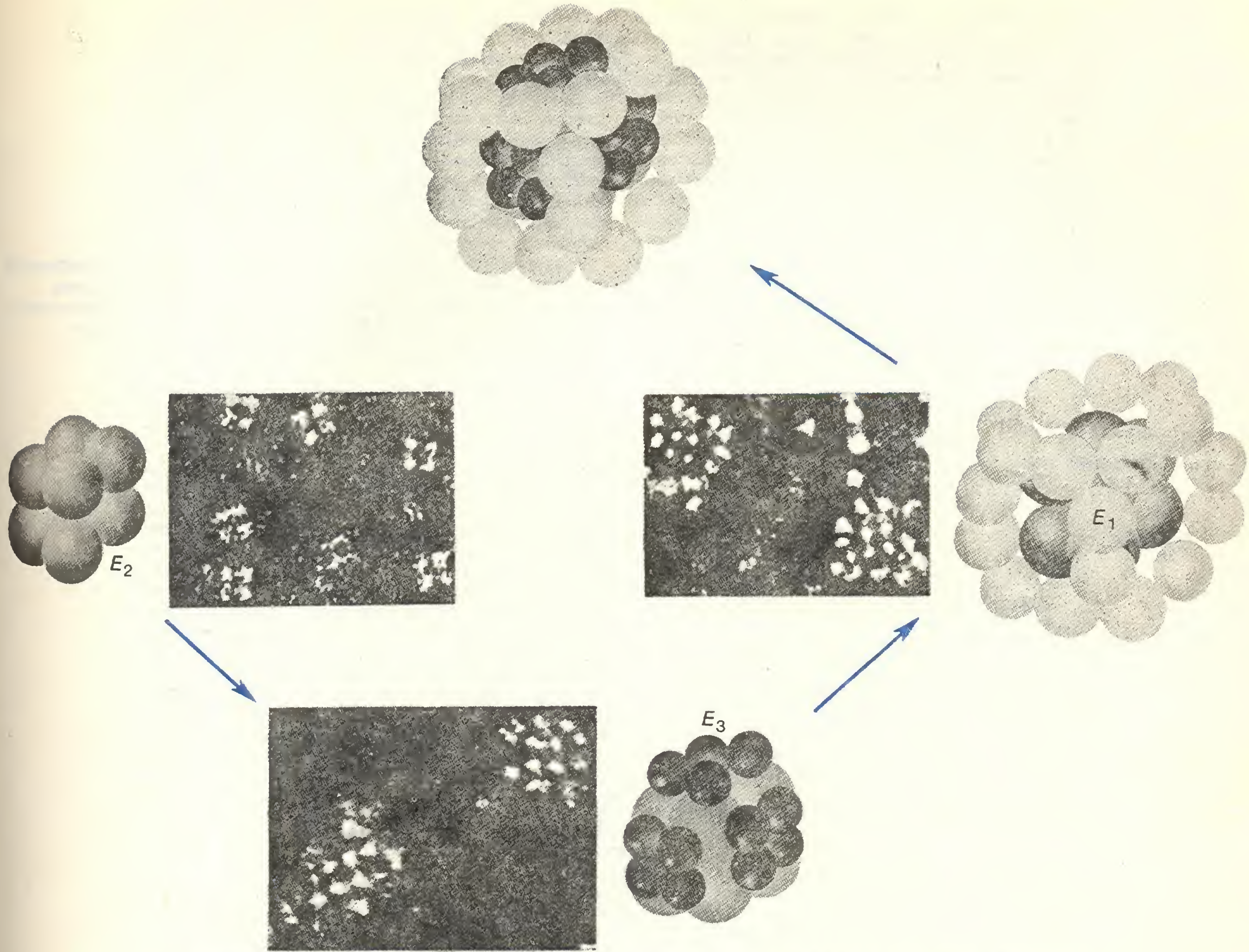
Enzymy oligomeryczne zostały w toku ewolucji tak wyselekcjonowane, że przynajmniej jednym efekto-



Rys. 26. Allosteryczne sprzężenia zwrotne w biosyntezie. Schemat pokazuje powstawanie kilku różnych aminokwasów z kwasu asparaginowego jako substratu (aminokwasy oznaczone skrótami trójliterowymi, tak jak w tabeli; Hse — homoseryna, Asp-PO_4 i Asp-CHO — fosforan asparatylu i semialdehyd asparatylu, odpowiednio). Powstające aminokwasy (droga biosyntezy — strzałki czarne, kilka grotów na strzałce oznacza, że pominięto kilka produktów pośrednich) stają się efektorami allosterycznymi, hamującymi (strzałki niebieskie) niektóre wcześniejsze etapy ich biosyntezy



Rys. 27. Elektronowe przesunięcia w enzymie w czasie katalizy. Chymotrypsyna jest enzymem proteolitycznym (proteazą), tj. rozszczepiającym wiązania peptydowe w białkach. W czasie katalizy nie ulega ona wyraźnym przekształceniom konformacyjnym (w odróżnieniu od np. karboksypeptydazy — rys. 25) ani też nie doprowadza do geometrycznego naprężenia w substracie (w odróżnieniu od np. lizozymu, rys. 24). Katalityczne działanie chymotrypsyny odbywa się głównie, jeżeli nie wyłącznie, w wyniku przesunięć elektronowych w enzymie i w substracie: a) W obszarze czynnym chymotrypsyny (i wielu innych proteaz zwanych „serynowymi") występuje aminokwas seryna w bezpośrednim sąsiedztwie histydyny i kwasu asparaginowego. Aminokwasy te nie sąsiadują ze sobą w sekwencji, ale są zbliżone przez pofałdowania łańcucha polipeptydowego chymotrypsyny. Substrat jest silnie wiązany przez enzym w bezpośrednim sąsiedztwie układu Asp-His-Ser (od góry rysunku), a atomy C' i N wiązania peptydowego znajdują się w kontakcie van der Waalsa z grupami czynnymi enzymu. Powoduje to przesunięcia elektronowe (niebieskie strzałki) nadające tlenowi seryny własności nukleofilne. b) Atak nukleofilny na grupę —CO substratu prowadzi do utworzenia wiązania kowalencyjnego między enzymem a częścią substratu zaczynającą się od —CO wiązania peptydowego, rozrywanego przez enzym. Histydyna staje się donorem wodoru do grupy —NH substratu, a powstająca amina odłącza się od obszaru czynnego. c) Na miejsce aminy wchodzi cząsteczka wody, hydrolizująca wiązanie kowalencyjne enzym–substrat. d) Druga część substratu z utworzoną grupą —COOH odłącza się i enzym jest zregenerowany



Rys. 28. Kompleks wieloenzymowy. Wiele enzymów oligomerycznych, katalizujących kolejne etapy w ciągu reakcji metabolicznych uformowało się w toku ewolucji w struktury polimeryczne zw. kompleksami wieloenzymowymi. W kompleksach tych znacznie sprawniej przekazywane są produkty pośrednie do następnych enzymów. Rysunek przedstawia jeden z takich kompleksów wieloenzymowych, dehydrogenazę ketokwasów, która wprowadza produkt końcowy przemian cukrowych i tłuszczowych, kwas pirogronowy, do cyklu Krebsa (→ Organizacja procesów życiowych komórki). Pokazane są modele strukturalne (kule reprezentują protomery) oraz zdjęcia obrazu otrzymanego za pomocą mikroskopu elektronowego. Kompleks zbudowany jest z trzech enzymów oligomerycznych: enzymu usuwającego grupę —COOH kwasu pirogronowego i redukującą jego grupę —CO (E_1); enzymu utleniającego produkt poprzedniej reakcji do kwasu octowego (E_2) oraz enzymu przenoszącego resztę kwasu octowego na koenzym A (E_3). Ten ostatni związek wchodzi do cyklu Krebsa. Organizatorem kompleksu jest E_2, zbudowany z 8 protomerów w układzie sześciennym. Przyłącza on na swoich ścianach 6 cząsteczek E_3 (każda złożona z 4 protomerów) oraz na krawędziach 12 cząsteczek E_1 (każda złożona z 2 protomerów). Masa cząsteczkowa kompleksu wynosi ok. 4,5 miliona daltonów

rem reakcji jest produkt końcowy tego ciągu reakcji, w którym dany enzym stanowi stadium początkowe (rys. 27). Hamowanie funkcji enzymu może być całkowite lub stopniowe. Na przykład enzym, który syntetyzuje glutaminę i zapoczątkowuje co najmniej 6 różnych ciągów reakcji w przemianie azotowej bakterii, zbudowany z 12 identycznych podjednostek, jest częściowo hamowany przez każdy z 6 końcowych produktów, a całkowicie — przez wszystkie razem. Wiele jednak enzymów allosterycznych ma podjednostki nieidentyczne i zwykle zbudowane z dimerycznych protomerów, w których jeden monomer jest podjednostką regulacyjną, drugi zaś — katalityczną. W dimerze takim podjednostka katalityczna, oddzielona od regulacyjnej, może być zupełnie nieaktywna lub też może wykazywać całkowitą aktywność. Stąd też powstają dwie klasy enzymów regulacyjnych tego typu: jedne na efektor odpowiadają zahamowaniem aktywności, drugie — jej wyzwoleniem.

Nie tylko hamowanie i aktywacja są rezultatem oddziaływań allosterycznych w enzymach. Zmiana struktury trzeciorzędowej może doprowadzić do zmiany geometrii centrum wiązania substratu, pociągającej za sobą zmianę swoistości substratowej enzymu. Przykładem może tu być dehydrogenaza glutaminianowa, enzym sprzęgający biosyntezę aminokwasów z cyklem kwasów trójkarboksylowych Krebsa (→ Organizacja procesów życiowych komórki), który może katalizować, w zależności od efektora, przyłączenie grupy aminowej do α-ketoglutaranu i tworzyć kwas glutaminowy, lub do pirogronianu i tworzyć alaninę. Enzym ten jest zbudowany z 6 identycznych podjednostek i w roztworze występuje jako heksamer mogący mieć dwie różne postaci (oktaedr lub piramida trygonalna, → Budowa kryształów). W obecności efektorów, a także niektórych hormonów, jedna postać heksameru przechodzi w drugą z jednoczesną zmianą swoistości substratowej enzymu.

zmiana struktury i zmiana swoistości substratowej

W enzymie oligomerycznym podjednostki mogą się od siebie różnić swoistością w stosunku do substratu. Enzym nabiera wówczas właściwości układu (kompleksu) wieloenzymowego (staje się tzw. multienzy-

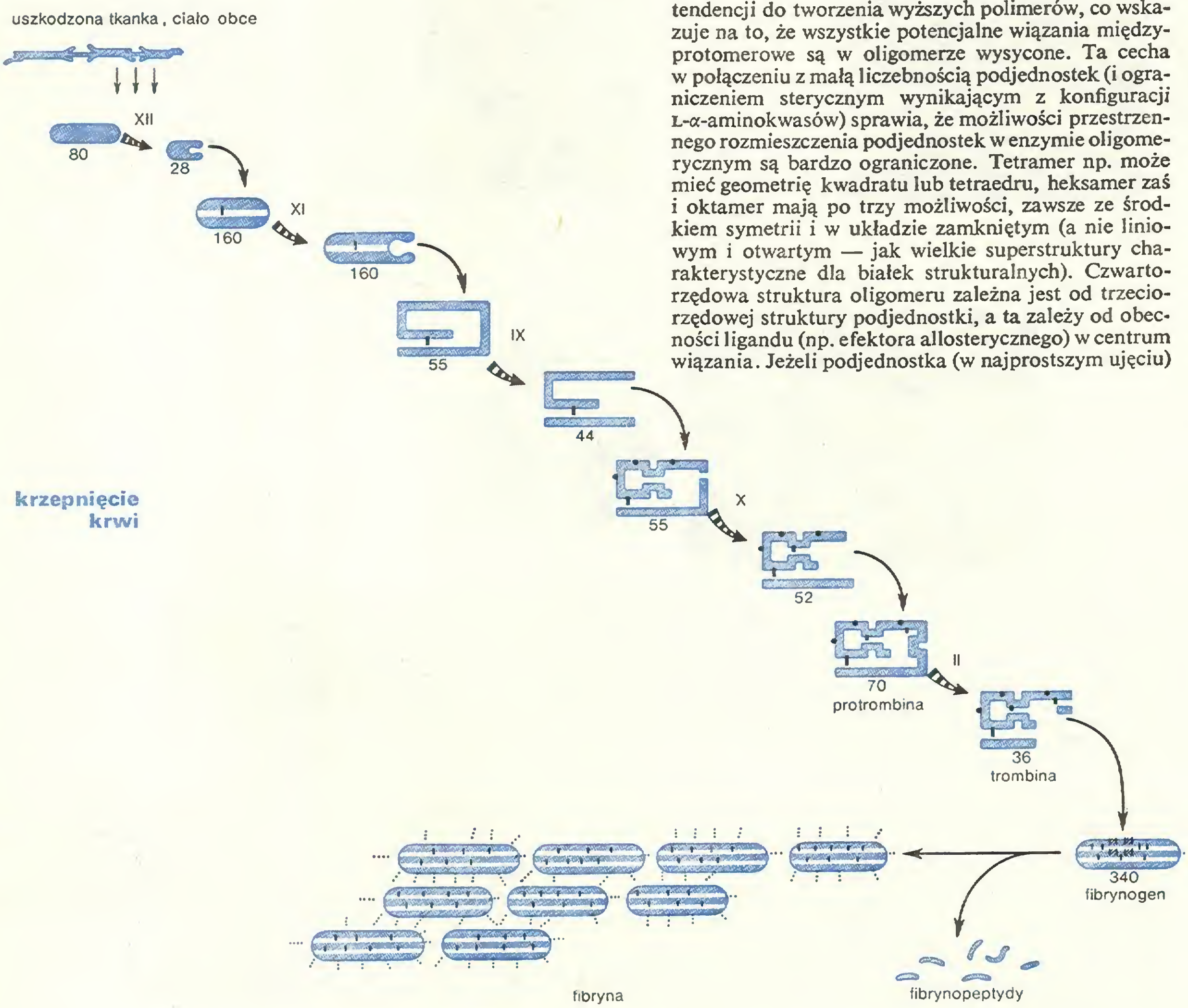
mem). Jednym z najprostszych jest enzym, który syntetyzuje tryptofan, zbudowany z dwu par nieidentycznych podjednostek $\alpha_2\beta_2$. Podjednostka α katalizuje powstanie indolu, a podjednostka β przyłącza do niego serynę, tworząc ostatecznie tryptofan. Aktywności enzymatyczne wolnych podjednostek są raczej niskie, a w kompletnym enzymie rosną prawie stukrotnie. Zespół dwu podjednostek ($\alpha_2\beta_2$) pozwala na efektywne sprzężenie dwu kolejnych reakcji, a podjednostki wzajemnie się w nim aktywują. Te korzystne cechy są w przyrodzie szeroko wykorzystywane i bardzo wiele kluczowo ważnych enzymów uformowanych jest w takie kompleksy wieloenzymowe (rys. 28). Składają się one z kilku enzymów oligomerycznych i mają ściśle określoną budowę przestrzenną, w której jeden enzym tworzy rdzeń i jest organizatorem całości. Przykładem kompleksu może być dehydrogenaza ketokwasów (rys. 28), organizatorem jej jest transacylaza liponianowa (enzym zbudowany z 24 bardzo podobnych, może identycznych podjednostek), która przyłącza do siebie po 12 cząsteczek pozostałych dwu enzymów (w jej nieobecności nie agregują one ze sobą). W toku ewolucji enzymy wielofunkcyjne tak zostały wyselekcjonowane, że ich częściami składowymi są enzymy oligomeryczne, katalizując kolejne etapy w danym ciągu przemian metabolicznych. Stworzyło to możliwość bezpośredniego przekazywania produktów jednych reakcji do enzymów katalizujących etap następny, bez strat spowodowanych dyfuzją i konkurencją innych układów enzymatycznych.

kompleksy wieloenzymowe

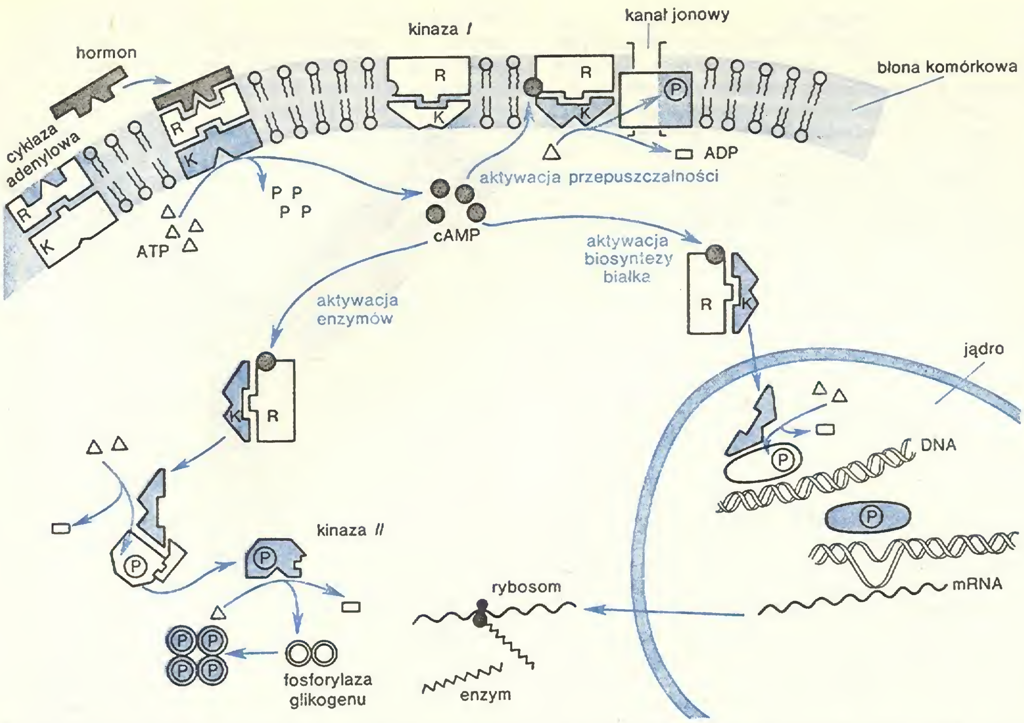
Większość (ok. 95%) zbadanych enzymów oligomerycznych zbudowana jest z 2, 4, 6 lub 8 podjednostek. Protomery ich są prawie zawsze identyczne (choć mogą się składać z dwu lub nawet więcej jednostek — i to niekoniecznie identycznych), a w oligomerze znajdują się w równoważnym (pseudoidentycznym) otoczeniu. Swobodne protomery łączą się we właściwą dla danego enzymu strukturę czwartorzędową, bez tendencji do tworzenia wyższych polimerów, co wskazuje na to, że wszystkie potencjalne wiązania międzyprotomerowe są w oligomerze wysycone. Ta cecha w połączeniu z małą liczebnością podjednostek (i ograniczeniem sterycznym wynikającym z konfiguracji L-α-aminokwasów) sprawia, że możliwości przestrzennego rozmieszczenia podjednostek w enzymie oligomerycznym są bardzo ograniczone. Tetramer np. może mieć geometrię kwadratu lub tetraedru, heksamer zaś i oktamer mają po trzy możliwości, zawsze ze środkiem symetrii i w układzie zamkniętym (a nie liniowym i otwartym — jak wielkie superstruktury charakterystyczne dla białek strukturalnych). Czwartorzędowa struktura oligomeru zależna jest od trzeciorzędowej struktury podjednostki, a ta zależy od obecności ligandu (np. efektora allosterycznego) w centrum wiązania. Jeżeli podjednostka (w najprostszym ujęciu)

podjednostki enzymu oligomerycznego

krzepnięcie krwi



Rys. 29. Krzepnięcie krwi jest wynikiem złożonej sekwencji reakcji enzymatycznych (kaskady), wywołanych m.in. uszkodzeniem tkanki, kontaktem z ciałem obcym (na lewo na górze), a doprowadzającej do powstania fibryny (na dole), wysokospolimeryzowanego, nierozpuszczalnego białka, zorganizowanego w rozległą sieć, w której utkwione są komórki krwi (skrzep krwi). Rysunek przedstawia kolejną aktywację 5 enzymów protolitycznych, zw. czynnikami krzepnięcia i oznaczonych cyframi rzymskimi, cyfry arabskie — oznaczają masy cząsteczkowe w tys. daltonów. Każdy z nich występuje w dwu postaciach: nieczynnego prekursora i czynnego enzymu (strzałka prążkowana pokazuje przekształcenie prekursora w enzym). Utworzony enzym modyfikuje (strzałka pojedyncza) kowalencyjną straukturę prekursora następnego enzymu przez nacięcie w nim kilku wiązań peptydowych (grot strzałki wskazuje na obszar ulegający wycięciu). Po usunięciu odcinka łańcucha polipeptydowego ujawnia się obszar czynny enzymu. Przedstawione czynniki krzepnięcia są proteazami serynowymi (rys. 27), a lokalizacja katalitycznie czynnych reszt asparaginy, histydyny i seryny pokazana jest kropkami na schemacie budowy czynnika X i II; proste czarne kreski na schematach wskazują pozycje wiązań dwusiarczkowych, utrzymujących integralność cząsteczki po wycięciu części łańcucha. Ostatnim enzymem w kaskadzie jest czynnik II (prekursor — protrombina i enzym — trombina). Trombina — ale nie pozostałe czynniki — atakuje niektóre wiązania peptydowe w fibrynogenie (zacieniowane na modelu, na prawo na dole) i doprowadza do odszczepienia od niego fibrynopeptydów (na dole). Tak zmodyfikowany fibrynogen polimeryzuje spontanicznie do fibryny. W fibrynie monomery są początkowo połączone wiązaniami niekowalencyjnymi (kreski przerywane), potem jednak zostaje ona usztywniona wiązaniami kowalencyjnymi (niewidoczne na rysunku)



Rys. 30. Pobudzenie metabolizmu komórkowego przez hormon. Wiele hormonów jest wiązanych na powierzchni komórki, co stanowi pierwszy etap oddziaływania hormonu na komórkę. Wyzwala to w komórce wiele procesów, ale hormon już nie bierze w nich bezpośredniego udziału. W procesach tych zasadniczą rolę odgrywa cykliczny nukleotyd, którego synteza jest pobudzana przez związanie podjednostki regulacyjnej enzymu, cyklazy nukleotydowej (np. cyklazy przekształcającej ATP w cykliczny nukleotyd adenylowy z uwolnieniem nieorganicznego fosforanu — na lewo u góry). Aktywacja cyklazy przez hormon nie jest dobrze poznana, np. nie wiadomo czy podjednostka regulacyjna cyklazy (R) jest receptorem wiążącym hormon (jak na rys.), czy też istnieją jakieś etapy pośrednie. Cykliczny nukleotyd (cAMP), zw. „drugim posłańcem'' w mechanizmie działania hormonów, wprowadza przy udziale enzymu zw. kinazą, resztę fosforanową (P w kółku) do rozmaitych białek i aktywuje je w ten sposób. Rysunek przedstawia przekrój ok. $^1/_4$ komórki żywej i ukazuje 3 główne ciągi reakcji aktywowanych przez hormon za pośrednictwem cAMP: 1) Aktywacja przepuszczalności błony komórkowej jest spowodowana wprowadzeniem reszty fosforanowej do białka blokującego kanały jonowe i wywołaną tym zmianę jego konformacji; 2) aktywacja enzymów, na przykładzie aktywacji przemiany cukrowej, w czym biorą udział trzy enzymy: kinaza I aktywuje kinazę II, a ta aktywuje fosforylazę glikogenu, enzym prowadzący do rozszczepienia glikogenu z wydzieleniem glikozy; 3) aktywacja biosyntezy białka; katalityczna podjednostka kinazy (K) przenika przez błonę jądrową i wprowadza resztę kwasu fosforowego do białka stabilizującego DNA, co w bliżej nie poznany sposób ułatwia transkrypcję genu

ma tylko dwie możliwe struktury trzeciorzędowe (jedną charakterystyczną dla stanu wolnego, drugą — dla stanu związania z ligandem), to enzym oligomeryczny będzie miał dwie możliwe struktury czwartorzędowe, odpowiadające najbardziej korzystnej interakcji identycznych podjednostek w ich dwu wzajemnie wykluczających się formach. Wysycenie jednej podjednostki ligandem doprowadza do przekształcenia jej struktury trzeciorzędowej, zmieniając jednocześnie sposób, w jaki się wiąże ona z pozostałymi podjednostkami. Uniemożliwia to utrzymanie poprzedniej struktury czwartorzędowej i zmusza cząsteczki do zmiany ich konformacji na alternatywną. Powstanie nowa struktura czwartorzędowa, alternatywna do poprzedniej. Chwilowo podjednostki w niej, z wyjątkiem pierwszej, nie są wysycone ligandem, ale mają zwiększoną łatwość takiego wiązania, gdyż nowa ich struktura trzeciorzędowa posiada taką geometrię centrum wiązania jak podjednostka, która jest już z ligandem skompleksowana. Wiele enzymów (oraz białek transportowych, np. hemoglobina, str. 734) tak się zachowuje i efekt allosteryczny definiowano pierwotnie właśnie jako zgrane i symetryczne przekształcenie wszystkich podjednostek oligomeru. Dobrze to tłumaczyło kinetykę wielu reakcji enzymatycznych i procesów transportu. Dziś się jednak sądzi, że konformacja może się zmieniać stopniowo w kolejnych podjednostkach i że symetryczne przekształcenie struktury czwartorzędowej jest tylko skrajną z wielu możliwości, sięgających aż do dysocjacji enzymu oligomerycznego do wolnych podjednostek.

funkcjonalna architektura enzymu

W każdym jednak razie funkcjonalność architektury enzymu wyraża się ścisłym sprzężeniem między jego strukturą trzecio- i czwartorzędową, stanowiącym, że aktywność enzymu staje się zależna i od czynników, które wpływają na układ podjednostek w polimerze, i od tych, które modyfikują konformację podjednostki. Zawsze jednak ostatecznym skutkiem związania ligandu, substratu czy efektora jest kooperatywne wzmożenie lub zahamowanie katalizy enzymatycznej.

kaskady enzymatyczne

Allosteryczne regulacje polegają w istocie na niekowalencyjnych modyfikacjach struktury enzymu. Stąd też są one łatwo odwracalne i enzym może oscylować pomiędzy stanami odpowiadającymi wiązaniu rozmaitych ligandów. Często jednak potrzebna jest bardziej trwała zmiana własności enzymu — uruchomienie lub zahamowanie jego aktywności czy uzależnienie go (względnie uniezależnienie) od pewnych efektorów. Do tego celu służą kowalencyjne potranslacyjne modyfikacje struktury enzymu, z reguły nadrzędne w stosunku do działania efektorów allosterycznych. Enzym może być modyfikowany nieodwracalnie, gdy np. rozcięte zostaną w nim niektóre wiązania peptydowe (rys. 29), albo odwracalnie, gdy odpowiedni system enzymatyczny może usunąć grupy modyfikujące (rys. 30). W każdym wypadku — ściśle określony bodziec (jak np. uraz powodujący krzepnięcie krwi czy hormon) (rys. 29 i 30) wyzwala kaskadowe, wielostopniowe układy enzymatyczne o potężnym działaniu biologicznym.

Białka — sygnały i receptory sygnałów

W organizmach zwierzęcych funkcja biologiczna wszystkich komórek jest całkowicie podporządkowana potrzebom ustroju jako całości. Nadrzędnym integratorem jest system nerwowy, komunikujący się

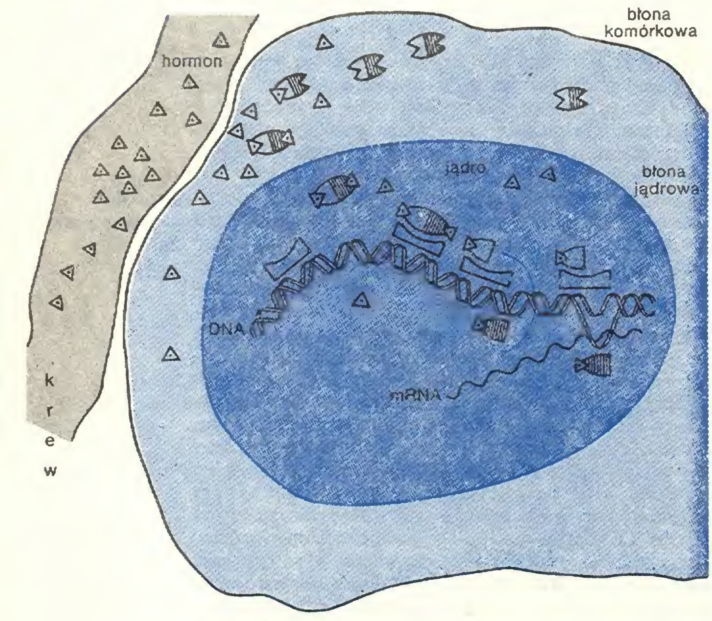
z komórkami i tkankami za pośrednictwem hormonów — grup różnorodnych chemicznie substancji, do których się dzisiaj zalicza zarówno hormony w sensie klasycznym (tj. produkty gruczołów wydzielania wewnętrznego), jak i tzw. neurotransmitery, czyli związki wydzielane przez zakończenia nerwowe. Większość hormonów stanowią pochodne aminokwasów oraz oligo- i polipeptydy, ale są wśród nich także związki nieaminokwasowe, jak sterydy. Hormony przekazują bodźce nerwowe od jednej komórki nerwowej do drugiej (np. acetylocholina, noradrenalina), kontrolują metabolizm (np. insulina — przemianę cukrową), dojrzewanie i różnicowanie komórek (np. hormony płciowe). Hormony są sygnałami chemicznymi, wyzwalającymi w komórkach ich utajone lub nie w pełni ujawnione możliwości biologiczne przez aktywację tych genów, które w nieobecności hormonu pozostają w stanie represji, a zawarta w nich informacja genetyczna nie znajduje wyrazu w produkcji swoistego białka.

hormony i neurotransmitery

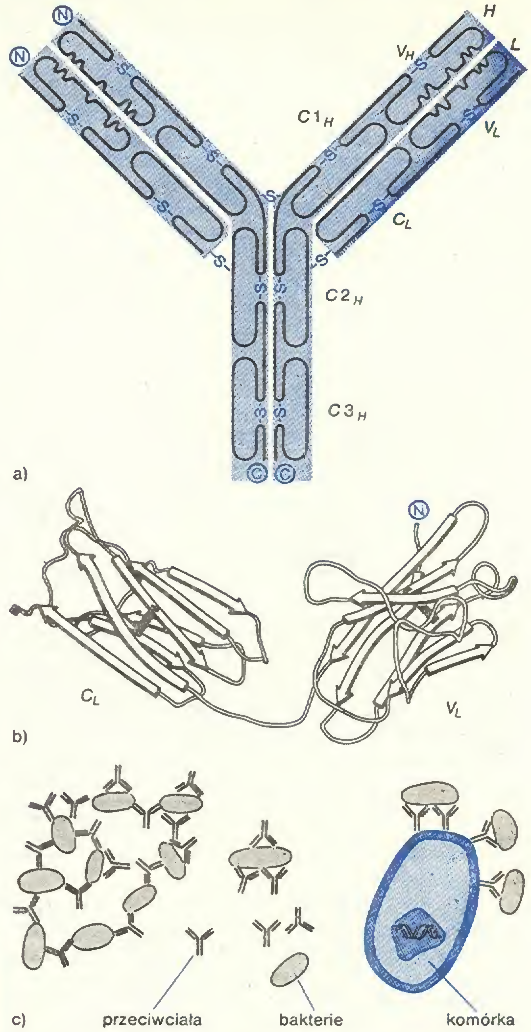
receptory

Hormony docierają do wielu (niektóre do wszystkich) komórek organizmu, ale tylko tzw. komórki docelowe dla danego hormonu odpowiadają na jego obecność charakterystyczną aktywacją swoich funkcji biologicznych. Komórki bowiem docelowe, w odróżnieniu od innych, zawierają białka zwane receptorami, zdolne do silnego i ściśle wybiórczego wiązania właściwego hormonu. Receptory hormonów są białkami o masach cząsteczkowych w przybliżeniu od ok. 100 000 do 300 000 daltonów (większe zbudowane są z podjednostek), a ich cząsteczka ma kształt wydłużony i charakter wyraźnie hydrofobowy. Prawie wszystkie receptory hormonów są trwale związane z błoną komórkową i wychwytują hormony z otaczającego komórkę płynu. Odmienne są tylko receptory hormonów sterydowych (i jak się zdaje, hormonów tarczycy), które się mogą przemieszczać w cytoplazmie i pobierać stamtąd sterydy, wnikające swobodnie do komórki poprzez błonę.

Receptor wiąże hormon, ulega przekształceniu konformacyjnemu, podobnie jak białka transportowe i enzymy. Receptor nie jest jednak transporterem doprowadzającym hormon do wrażliwego nań obszaru



Rys. 31. Receptory hormonów sterydowych. Hormony sterydowe (trójkąty), w odróżnieniu od innych hormonów swobodnie przenikają z krwi do komórki. W cytoplazmie napotykają na swoiste dla nich receptory, białka zbudowane z dwu podjednostek i mające dwa miejsca wiązania dla danego hormonu. Utworzenie kompleksu hormon-receptor modyfikuje cząsteczkę receptora tak, że może ona przeniknąć przez błonę jądrową i wiąże się z białkami stabilizującymi DNA. To związanie ponownie modyfikuje cząsteczkę receptora: dysocjuje ona, jedna jej podjednostka (lokalizacyjna) pozostaje związana, druga (aktywacyjna) przenosi się wzdłuż nici DNA i w bliżej nie poznany sposób rozluźnia strukturę DNA, ułatwiając transkrypcję określonego genu



Rys. 32. Immunoglobuliny — białka odpowiedzi obronnej organizmu. a) Immunoglobulina zbudowana jest z 4 łańcuchów polipeptydowych: dwóch „ciężkich" — *H* (od ang. *heavy*), złożonych z ok. 450 reszt aminokwasowych, i dwóch „lekkich" — *L* (od ang. *light*), złożonych z ok. 220 reszt. Cząsteczka immunoglobuliny ma kształt litery Y z dwukrotną osią symetrii między parami łańcuchów H_L. Łańcuchy *H* ciągną się przez całą długość cząsteczki, łańcuchy *L* — tylko wzdłuż ramienia litery Y; końce *N* łańcuchów znajdują się na końcu ramion Y. W sekwencji aminokwasowej łańcuchów powtarzają się sekwencje po około 110 aminokwasów, zw. domenami, spięte pośrodku wiązaniem dwusiarczkowym (—S—). b) Model utworzony na podstawie badań krystalograficznych. Widać, że każda z domen stanowi typową β-baryłkę (rys. 10) w układzie łańcuchów przeciwrównoległym (strzałki); widoczne jest również wiązanie dwusiarczkowe (czarny słupek). Sekwencja aminokwasowa w domenach *C* (od ang. *constant* 'stały') jest bardzo podobna u wszystkich osobników danego gatunku i mało się różni u rozmaitych gatunków kręgowców. Domeny *V* (od ang. *variable* 'zmienny'), przy pewnym podobieństwie ogólnym sekwencji do domen *C*, mają wyróżniające je bardzo krótkie odcinki sekwencji „hiperzmiennych" (linie faliste na rys. a). Te hiperzmienne sekwencje są wprowadzone do cząsteczki immunoglobuliny w odpowiedzi na obecność antygenu w organizmie (mechanizm genetyczny tego procesu jest zagadkowy) i formują w niej obszar wiązania antygenu. c) Cząsteczka immunoglobuliny ma dwa, zawsze identyczne, obszary wiązania antygenu. Wiążąc się z antygenem, immunoglobulina (przeciwciało — jeżeli znany jest swoisty dla niej antygen) może tworzyć rozległą sieć (np. z lewej strony — sieć, której węzły tworzą determinanty antygenowe na powierzchni bakterii), wypadającą ze środowiska (np. z krwi). Cząsteczki immunoglobuliny mogą także wiązać się na powierzchni bakterii, wirusów i in. (na dole pośrodku), a następnie — wiązać je z komórkami, które mają na swojej powierzchni receptory dla ostatniej czy ostatnich dwóch domen stałych immunoglobuliny. Komórki te (np. komórki żerne) mają zdolność niszczenia lub unieszkodliwiania rozmaitych ciał obcych, mogących przedostać się do organizmu

komórki ani też nie nabiera własności katalitycznych. W utworzonym kompleksie receptor staje się przetwornikiem otrzymanej z hormonem informacji i czyni ją zrozumiałą dla komórki. Receptor hormonów sterydowych (rys. 31) przekształca się w induktor (derepresor, → Organizacja procesów życiowych komórki) i uruchamia transkrypcję genów. Receptory trwale związane z błoną komórkową wywołują ten sam efekt, ale pośrednio: wiążą podjednostkę regulacyjną enzymu zwanego cyklazą nukleotydową (rys. 30), hamującą jej podjednostkę katalityczną, i uruchamiają kaskadę enzymatyczną, która aktywuje rozmaite białka komórkowe — enzymy, białka strukturalne (układ tubulina–mikrotubule), białka jądra komórkowego. Selektywność działania receptorów hormonów sterydowych polega prawdopodobnie na tym, że jedna ich podjednostka (rys. 31) rozpoznaje pewne obszary (tzw. miejsca akceptorowe) w chromatynie, a druga rozluźnia pobliski obszar zbitej, tj. spoczynkowej chromatyny. Nie wiadomo jednak, w jaki sposób zachodzące w obszarze błony komórkowej oddziaływanie przekształconego receptora innych niż sterydowe hormonów z cyklazą nukleotydową doprowadza do aktywacji wybranych genów.

przetwornik

nie tylko hormony...

Hormony są sygnałem chemicznym, ale analogiczny system receptorów funkcjonuje w przetwarzaniu informacji zawartej w sygnale fizycznym (takim receptorem jest np. rodopsyna, zwana purpurą wzrokową, w siatkówce oka, aktywująca cyklazę adenylową pod wpływem światła widzialnego). I nie tylko sygnały chemiczne czy fizyczne są w ten sposób rozpoznawane, ale również sygnały, których natury nie potrafimy jeszcze zdefiniować w pojęciach innych niż biologiczne. Należy do nich zdolność organizmów wyższych do rozróżniania między składnikiem (substancją, komórką) „swoim" i „obcym", leżąca u podłoża immunologicznych odczynów obronnych.

...ale i odporność

Ponieważ wniknięcie komórki obcej do organizmu wielokomórkowego stanowi dla niego potencjalne niebezpieczeństwo, w toku ewolucji organizmów wyższych wykształcone zostały mechanizmy zapobiegające temu zagrożeniu. Typowym ich przykładem jest produkcja przeciwciał — białek należących do grupy immunoglobulin (rys. 31) — w odpowiedzi na zakażenie bakteryjne czy szczepienie ochronne. Przeciwciała mogą również powstać po wprowadzeniu do organizmu prawie każdej substancji wielkocząsteczkowej — pod warunkiem, że nie jest ona naturalnym i prawidłowym jego składnikiem. Wytworzone przeciwciało wiąże czynnik, który wzbudził jego produkcję (tzw. antygen); jest to wiązanie niekowalencyjne, ale silne i o wybiórczości często przekraczającej wybiórczość wiązania substratu przez enzym.

Obszar wiązania antygenu (rys. 32) jest zagłębieniem w cząsteczce przeciwciała utworzonym wspólnie przez *N*-końcowe odcinki łańcucha ciężkiego i łańcucha lekkiego, czyli znajduje się w domenie zmiennej, a jego geometrię określają łańcuchy boczne aminokwasów hiperzmiennych (rys. 32a). Zagłębienie to (w cząsteczce przeciwciała) ma wymiary ok. $3,5 \times 1,5 \times 1,0$ nm, tak że może pomieścić tylko niewielką część antygenu wielkocząsteczkowego, jego tzw. determinantę, liczącą 5–7 reszt aminokwasów czy cukrów. Antygen ma zazwyczaj wiele determinant, a przeciwciało wiąże co najmniej dwa antygeny (rys. 32c), tak że powstający kompleks jest agregatem zbudowanym z wielu cząsteczek. Konformacja przeciwciała ulega w nim zmianie, wyrażającej się m.in. usztywnieniem połączeń między domenami stałymi (rys. 32b). W wyniku zaś tego ulega aktywacji druga funkcja przeciwciała — funkcja efektorowa: przeciwciało nabiera zdolności do wiązania się z błonami komórek (np. żernych, trawiących antygeny) i z czynnikami znajdującymi się w osoczu krwi, a wyzwalającymi kaskadę enzymatyczną (tzw. komplement, czyli dopełniacz), niszczącą spójność błony komórki obcej. Przeciwciało jest receptorem, który rozpoznaje sygnał obcości w swojej domenie zmiennej, a otrzymaną informację przetwarza w domenach stałych (rys. 32a,b) w postać zrozumiałą dla wyspecjalizowanych układów komórkowych organizmu, mogących wyeliminować komórkę czy substancję obcą.

Za zjawisko odporności odpowiedzialne są białe komórki krwi zwane limfocytami. Jedne z nich wytwarzają przeciwciała i uwalniają je do krwi i innych płynów ustrojowych; odrębny rodzaj stanowią limfocyty biorące udział w odczynach alergicznych w odrzucaniu przeszczepów tkankowych: rozpoznają one antygeny za pomocą swoich receptorów, związanych z błoną komórkową. Budowa ich jest odmienna (choć o wielu cechach wspólnych) niż przeciwciał i kontrolowane są przez odrębne zespoły genów. Powstaje w związku z tym pytanie, jedno z najważniejszych we współczesnej biologii: skąd przeciwciało czy inny receptor „wie", że dany antygen jest obcy? W jaki sposób organizm może wyprodukować tak ogromną ilość różnych receptorów (a są ich zapewne miliony), rozpoznających praktycznie wszystkie determinanty antygenowe, zarówno naturalne, jak syntetyczne? Niegdyś sądzono, że antygen funkcjonuje jak matryca wytwarzająca specyficzną geometrię obszaru wiązania w nie uformowanej jeszcze ostatecznie cząsteczce immunoglobuliny. Dziś wiemy, że te wszystkie różne receptory istnieją już przed pierwszym kontaktem organizmu z antygenem, w mało zróżnicowanych komórkach macierzystych, z których powstają immunologicznie czynne komórki, limfocyty. Antygen jest dla nich sygnałem wyzwalającym — tak jak się to dzieje przy pobudzeniu komórki przez hormony — podziały i różnicowanie, prowadzące do namnożenia populacji komórkowej zdolnej do swoistej odpowiedzi na dany antygen. Różnorodność przeciwciał (i innych tego rodzaju receptorów sygnału obcości) musi więc być wynikiem działania aparatu genetycznego, a genetyczna kontrola biosyntezy immunoglobulin wydaje się wyjątkową. Pojedynczy łańcuch polipeptydowy immunoglobuliny, zarówno lekki, jak ciężki, jest ciągły, ale jego dwie funkcjonalnie różne części, domena zmienna i domena stała, są pod kontrolą dwu różnych genów (ściślej: zespołów genów). Ogólną jednak prawidłowością biosyntezy białka jest to, że jeden łańcuch polipeptydowy jest kontrolowany przez jeden gen. W czasie biosyntezy immunoglobulin nie pojawiają się odrębne mRNA (→ Kwasy nukleinowe) domeny zmiennej i domen stałych ani nie występują krótsze peptydy, które by się mogły połączyć w jeden łańcuch (ciężki lub lekki) po zakończonej translacji. Można stąd wnosić, że już w obrębie genetycznego aparatu następuje połączenie genów domeny zmiennej z genem (genami) domen stałych. To zapewne dzięki tej szczególnej właściwości genów immunoglobulin do domeny zmiennej mogą się włączać krótkie sekwencje aminokwasowe, nadające specyficzną geometrię centrum wiązania antygenu. Nie wiadomo jednak, czy w genach domeny zmiennej zawarta jest informacja nabyta w czasie ewolucji organizmu, czy też powstaje ona w wyniku mutacji somatycznych, zachodzących w komórkach macierzystych układu odpornościowego.

rozpoznawanie sygnałów obcości

Kwasy nukleinowe

Edward Czuryło i Magdalena Fikus

Najmniejszym obiektem żywym jest komórka. Z teoretycznych rozważań biochemicznych i biofizycznych wynika, że najmniejsza żywa komórka powinna mieć średnicę co najmniej 50 nm, a jej sucha masa (masa pozostała po odparowaniu wody z komórki) powinna by się składać z ok. 1,5 mln atomów. Średnica najmniejszej żywej komórki znalezionej przez biologów wynosi ok. 100 nm i jej sucha masa zbudowana jest z ok. 12 mln atomów.

Ta ogromna liczba atomów połączona jest w cząsteczki, z których znaczna większość to cząsteczki organiczne, cząsteczki zawierające węgiel. Węgiel jest pierwiastkiem czterowartościowym i może tworzyć wiązania chemiczne z czterema innymi atomami. Dzięki temu cząsteczki organiczne mogą się ze sobą łączyć, tworząc tzw. makrocząsteczki. Makrocząsteczka, w której się powtarza stale określona grupa atomów (zwana monomerem lub merem), nazywa się polimerem. Taka makrocząsteczka przypomina łańcuch połączonych ze sobą ogniw, a proces tworzenia się takiego łańcucha nazywa się polimeryzacją. Można sobie wyobrazić łańcuch (makrocząsteczkę) zbudowany z kilku lub kilkunastu różnych (pod względem składu chemicznego, kształtu lub wielkości) ogniw. Makrocząsteczka zbudowana z różnych ogniw nazywako się polimerem, polimerem złożonym z kilku urodzajów monomerów.

biopolimery

W żywej komórce występuje kilka rodzajów makrocząsteczek — biopolimerów, a każda taka makrocząsteczka jest kopolimerem innej grupy związków organicznych; np. makrocząsteczka białka (→ Białka) zbudowana jest ze związków stanowiących grupę aminokwasów (istnieje 20 rodzajów aminokwasów), cząsteczka wielocukru składa się z cząsteczek tzw. cukrów prostych.

W dalszej części tego artykułu rozpatrzymy budowę, strukturę i funkcję, jaką spełniają kwasy nukleinowe — biopolimery zbudowane z merów zwanych nukleotydami.

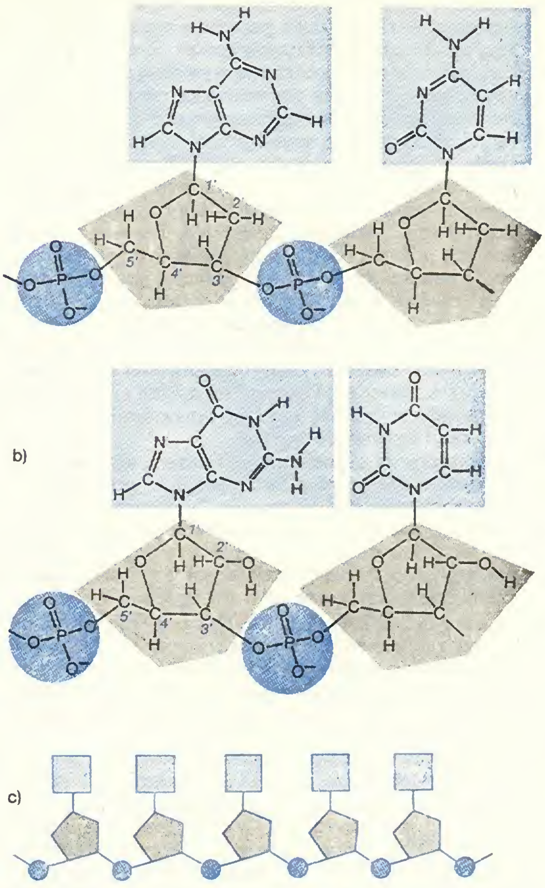
Budowa kwasów nukleinowych

W skład kwasów nukleinowych wchodzą atomy wodoru, węgla, azotu, tlenu i fosforu. Powtarzającymi się jednostkami kwasów nukleinowych są nukleotydy, w których skład wchodzi reszta kwasu fosforowego, cukier i zasada. W naturalnych kwasach nukleinowych występują dwie różne cząsteczki cukru: ryboza i dezoksyryboza. W zależności od tego, jaki cukier występuje w kwasie nukleinowym, mówimy o kwasach rybonukleinowych lub o kwasach dezoksyrybonukleinowych (rys. 1). Każdy z tych kwasów zawiera cztery zasady. Kwas rybonukleinowy (RNA) zawiera adeninę, cytozynę, guaninę i uracyl, kwas dezoksyrybonukleinowy (DNA) — adeninę, cytozynę, guaninę i tyminę. Odpowiednie nukleotydy oznaczamy pierwszymi literami (dużymi) nazw zasady, a dla całkowitego opisania często przed nią dodaje się małą literę r lub d, w celu wskazania, o jakim rodzaju nukleotydu mowa (np. rA, dC itp.).

nukleotydy

łańcuch polinukleotydowy

Podczas łączenia się nukleotydów w łańcuch polinukleotydowy reszta kwasu fosforowego łączy wiązaniami kowalencyjnymi atom węgla 3′ jednego nukleotydu z atomem węgla 5′ drugiego nukleotydu (rys. 1). W ten sposób powstaje łańcuch złożony z dwu rodzajów ogniw, które występują przemiennie, a zasada jest jak gdyby perełką wiszącą obok. Kolejność pojawiania się po sobie poszczególnych zasad nie jest dowolna i jak się dowiemy później, ma ona ogromne znaczenie dla życia całej komórki i organizmu.



Rys. 1. Łańcuchy polinukleotydowe; zasadę umieszczono w prostokącie, cukier w pięciokącie, a resztę kwasu fosforowego w okręgu. W pierścieniu cukrowym podano numerację atomów węgla. a) Fragment łańcucha kwasu dezoksyrybonukleinowego. b) Fragment łańcucha kwasu rybonukleinowego, który od łańcucha kwasu dezoksyrybonukleinowego różni się tym, że w pierścieniu cukrowym przy atomie węgla 2′ zamiast wodoru (H) występuje grupa hydroksylowa (OH). c) Schemat łańcucha polinukleotydowego (oznaczenia analogiczne jak w punkcie a)

Struktura kwasów nukleinowych

Jedną z podstawowych cech charakterystycznych struktury biopolimerów jest kolejność merów w łańcuchu makrocząsteczki (w wypadku kwasów nukleinowych kolejność nukleotydów). Jest to tzw. sekwencja biopolimeru lub struktura pierwszorzędowa. Następną cechą jest ułożenie poszczególnych części łańcucha w przestrzeni, czyli tzw. struktura przestrzenna. Strukturę przestrzenną dzielono na: strukturę drugorzędową — tworzenie kształtów śrubowych odpowiedniego typu, oraz strukturę trzeciorzędową — ułożenie odcinków śrubowych i ewentualnych odcinków nieśrubowych względem siebie. Najnowsze badania wykazały, że ten podział jest sztuczny; najczęściej tworzenie się spirali i ich wzajemna orientacja zależą tylko od sekwencji i są ściśle uwarunkowane odpowiednimi wymaganiami fizycznymi. Inaczej mówiąc, przy danej sekwencji i ściśle określonych warunkach tylko jedna struktura będzie najbardziej ekonomiczna i najbardziej trwała. Strukturą natywną biopolimeru nazywamy taką strukturę, która zapewnia jego aktyw-

ność biologiczną, a więc taką strukturę w jakiej występuje w żywej komórce.

Istnieją cząsteczki, które w komórce łączą się po dwie lub więcej i w takim stanie są zdolne spełniać swe funkcje; są to tzw. struktury nadcząsteczkowe.

Struktura DNA

Na podstawie wyników badań rentgenograficznych prowadzonych przez M. Wilkinsa oraz zebranych do tego czasu wiadomości chemicznych i biochemicznych J. D. Watson i F. Crick opracowali w 1953 r. model cząsteczki DNA. Zadanie było stosunkowo łatwe, ponieważ wiadomo było, że w skład cząsteczki DNA wchodzą cztery nukleotydy oraz że liczba nukleotydów *A* jest równa liczbie nukleotydów *T*, liczba zaś nukleotydów *C* jest równa liczbie nukleotydów *G*. Na tej podstawie stwierdzono, że tworzą się pary nukleotydowe utrzymywane za pomocą wiązań wodorowych (rys. 2). Zjawisko to nazwano komplementarnością nukleotydów.

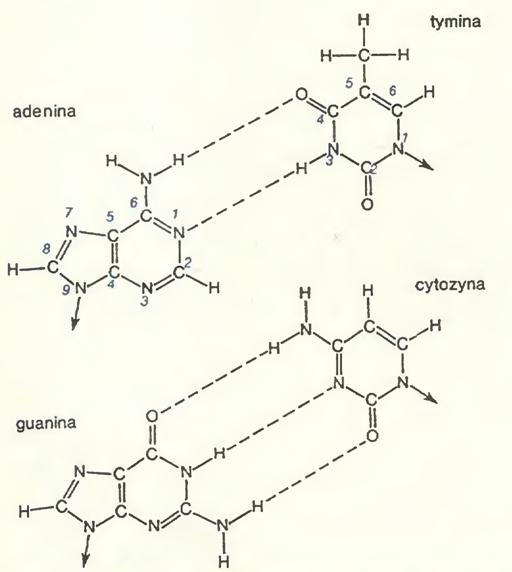
komplementarność nukleotydów

Cząsteczki DNA nie tworzą kryształów, gdyż są bardzo duże, a poza tym preparaty DNA najczęściej są niejednorodne pod względem długości i składu nukleotydowego. Ze stężonych roztworów DNA można jednak formować włókna. Włókno takie składa się z wielu cienkich i długich, równoległych do siebie cząsteczek DNA. Na takie włókna można skierować wiązkę promieni rentgenowskich, a z zarejestrowanego dyfrakcyjnego obrazu spróbować odtworzyć ułożenie nukleotydów w cząsteczce, ustalić jej strukturę. Otrzymana fotografia (il. 118, tabl. 29) różni się od rentgenogramu wszelkich kryształów. Nie można na niej znaleźć refleksów poszczególnych atomów, można jedynie stwierdzić, że w cząsteczce DNA jest powtarzający się element strukturalny. Należało zatem zastosować inną analizę takiego rentgenogramu niż w wypadku kryształów (→ Osiągnięcia krystalografii białek).

metoda porównywania obrazów dyfrakcyjnych

Przyjęto istnienie komplementarnych par zasad nukleinowych i poszukiwano takiej struktury, w której pary te powtarzałyby się periodycznie. Następnie odtwarzano dyfrakcyjny obraz dla modelu i porównywano go z dyfrakcyjnym obrazem DNA uzyskanym przez Wilkinsa (il. 118, tabl. 29). Podobieństwo dyfrakcyjnego obrazu modelu do dyfrakcyjnego obrazu cząsteczek DNA otrzymano wtedy, gdy przyjęto strukturę utworzoną z dwu nici polinukleotydowych skierowanych antyrównolegle i nawiniętych na powierzchnię walca wzdłuż prawoskrętnej linii śrubowej — heliksu. Strukturę taką nazywamy heliksem Watsona–Cricka. Heliks taki (il. 117, tabl. 29) ma średnicę 2 nm, skok 3,4 nm i skok ten utworzony jest przez 10 par nukleotydowych, a zatem na jedną parę nukleotydów przypada wycinek heliksu 0,34 nm. Płaszczyzny zasady są prostopadłe do osi heliksu. Ponieważ w obu komplementarnych parach występuje po jednej zasadzie „małej" i po jednej „dużej" średnica heliksu jest stała na całej długości cząsteczki.

heliks Watsona–Cricka



Rys. 2. Komplementarne pary zasad występujące w heliksie DNA; linią przerywaną zaznaczono wiązania wodorowe. Liczby obok atomów wchodzących w skład pierścieni oznaczają numerację tych atomów odpowiednio w zasadzie purynowej (dużej) — adeninie i guaninie, oraz w zasadzie pirymidynowej (małej) — tyminie i cytozynie

Widać więc, że model struktury DNA został odgadnięty a nie ustalony drogą analityczną. Dlatego też jego słuszność była przedmiotem wielu badań kontrolnych, których wyniki nie pozwoliły sformułować ważkich zarzutów podważających zgodność modelu z rzeczywistą strukturą DNA. Natomiast ustalono wiele faktów, które potwierdzały hipotezę Watsona i Cricka o helikalnej budowie makrocząsteczki DNA i mechanizmie przekazywania informacji genetycznej.

doświadczenie Meselsona i Stahla

Bardzo ciekawe doświadczenie przeprowadzili M. Meselson i E. W. Stahl, którzy stosując ciężki izotop azotu i prawo Archimedesa udowodnili hipotezę Watsona i Cricka. Poświęćmy temu doświadczeniu nieco więcej uwagi.

Kolonie bakterii *E. coli* hoduje się na pożywce, która zawiera azot ^{15}N, a nie zawiera atomów ^{14}N. Po pewnym czasie we wszystkich komórkach bakteryjnych DNA zawiera zasady, w których występują wyłącznie atomy azotu ^{15}N. Masa cząsteczkowa tego DNA jest większa od masy cząsteczkowej DNA bakterii hodowanych na normalnej pożywce. Różnica między nimi wynosi ok. 1% masy cząsteczkowej DNA. Nie można jej stwierdzić żadną znaną metodą pomiaru mas cząsteczkowych polimerów, gdyż jest ona bardzo mała, mniejsza od dokładności tych metod. Ponieważ przyrostowi masy *M* cząsteczkowej nie towarzyszy zmiana objętości, należy przypuszczać, że zmieniła się gęstość cząsteczki. Gęstość DNA zależy od procentowej zawartości par G–C i waha się w granicach 1,64–1,76 g/cm³. Dokładność pomiaru gęstości preparatów DNA jest nie mniejsza niż ±0,001 g/cm³. Gęstość DNA komórek *E. coli* jest następująca: gęstość ^{14}N-DNA wynosi 1,709 g/cm³, ^{15}N-DNA – 1,725 g/cm³. Gęstość DNA wyznacza się za pomocą wirowania w ultrawirówce roztworu tego DNA w stężonym (7,7*M*) roztworze chlorku cezu (CsCl). Po pewnym czasie wirowania takiego roztworu jego gęstość w różnych miejscach naczynia wirowniczego jest różna. Zmiany gęstości zależą w liniowy sposób od odległości od osi obrotu, tzn. wirowanie wytworzyło liniowy gradient gęstości. Jeśli się do takiego roztworu doda małą ilość roztworu DNA, to cząsteczki tego DNA będą się zbierały w takim miejscu naczynia wirowniczego, w którym gęstość roztworu chlorku cezu będzie równa gęstości DNA. Wynika to stąd, że na cząsteczkę DNA znajdującą się w roztworze o większej gęstości będzie działała siła wyporu Archimedesa, a na cząsteczkę znajdującą się w roztworze o mniejszej gęstości będzie działała siła odśrodkowa.

Badacze przenosili bakterie *E. coli* hodowane na pożywce zawierającej ^{15}N na pożywkę zawierającą ^{14}N i kontynuowali hodowlę przez czas równy wzrostowi jednego pokolenia lub dwu pokoleń. Z tych bakterii wydzielono DNA i wirowano go w roztworze CsCl. Okazało się, że DNA z bakterii, które rosły na pożywce zawierającej ^{14}N przez okres wzrostu jednego pokolenia, ma gęstość równą średniej arytmetycznej gęstości ^{14}N-DNA i ^{15}N-DNA. Natomiast DNA z bakterii, które rosły na pożywce zawierającej ^{14}N przez okres równy wzrostowi dwu pokoleń, rozdziela się na dwie warstwy, przy czym gęstość jednej z nich jest równa gęstości ^{14}N-DNA, a gęstość drugiej warstwy równa się gęstości DNA pierwszego pokolenia bakterii. Z wyników tych można wyciągnąć następują-

cy wniosek: w pierwszym pokoleniu bakterii ich DNA składa się z jednej nici zawierającej zasady z ^{14}N i jednej — z ^{15}N. W drugim pokoleniu część bakterii posiada DNA taki jak w pierwszym, pozostała zaś część bakterii posiada DNA, w którym obie nici zawierają tylko zasady z ^{14}N, jest to ^{14}N-DNA.

formy helikalnych DNA

Dalsze badania rentgenograficzne włókien DNA wykazały, że parametry heliksu mogą się nieco zmieniać. Wyróżnia się trzy formy helikalnych DNA we włóknach.

W przypadku włókien utworzonych z sodowej potasowej lub cezowej soli DNA o wilgotności względnej 75% stwierdzono formę *A*. Formę *A* charakteryzuje się skokiem heliksu 2,8 nm, zbudowanego z 11 par nukleotydów, których zasady są nachylone pod kątem 20° do osi heliksu. Podobną strukturę może tworzyć dwuniciowy RNA.

forma *A*

W przypadku soli sodowej DNA o 92% wilgotności względnej heliks DNA przyjmuje parametry identyczne jak w modelu Watsona i Cricka i tę formę nazywamy formą *B*.

forma *B*

Wyniki badań przeprowadzonych za pomocą innych metod pozwalają przypuszczać, że w roztworze o małych siłach jonowych DNA tworzy strukturę zbliżoną do formy *B*.

forma *C*

Wyróżnia się również formę *C* utworzoną przez sól litową DNA o wilgotności względnej 60% (skok heliksu 3,1 nm, który wypełniony jest przez 9,3 par nukleotydów). Zasady tworzą kąt 6° z osią heliksu. Heliks DNA przyjmuje formę *C* w roztworach o wysokich stężeniach soli lub w obecności glikolu etylenowego.

Zbadano również parametry heliksu DNA w hybrydzie DNA–RNA. Są one identyczne jak w przypadku formy *A*. Ponieważ helikalna struktura DNA utrzymywana jest głównie przez oddziaływania warstwowe i hydrofobowe oraz istniejące wiązania wodorowe, to można przypuszczać, że przejście od jednej formy heliksu do innej związane jest z różnym wkładem poszczególnych typów oddziaływań do całkowitej energii makrocząsteczki. W praktyce wkład ten zmieniamy przez stosowanie rozpuszczalników odpowiedniej kompozycji (zawartość i rodzaj soli, pH, zawartość substancji organicznych) lub też przez wprowadzenie do roztworu określonych białek. Można więc przypuszczać, że w roztworze cząsteczka DNA tworzy helikalną strukturę dynamiczną, a jej końcowymi formami mogą być forma *A* i *B*. Ta dynamiczna struktura zapewnia biologiczną aktywność DNA. W roztworach mogą się tworzyć struktury pośrednie; niektóre z tych struktur znaleziono doświadczalnie i stwierdzono, że są to struktury metastabilne i częściej występują w DNA bogatych w pary *A–T*.

metody wyznaczania masy cząsteczkowej

W zależności od rodzaju organizmu, z którego komórek pochodzi DNA, jego masa cząsteczkowa zawiera się w granicach od kilku milionów do kilkunastu miliardów jednostek masy atomowej (daltonów). Dokładne wyznaczanie masy cząsteczkowej natywnego DNA stanowi dość trudne zagadnienie.

Stosowanie tradycyjnych metod hydrodynamicznych jest niecelowe, gdyż łańcuchy DNA są bardzo wrażliwe na występujące napięcia ścinające. Inne metody takie jak mikroskopia elektronowa czy autoradiografia wymagają wstępnego wzorcowania. Z metod stosowanych najlepsza była metoda rozproszenia światła pod małymi kątami. Rozwój techniki laserowej stwarza możliwość zastosowania do badania współczynnika dyfuzji dużych cząsteczek dopplerowskiego przesunięcia długości fali rozproszonego światła laserowego. Współczynnik ten wraz ze stałą sedymentacji pozwala, podobnie jak rozproszenie światła, wyznaczyć bezwzględną masę cząsteczkową. Inną przeszkodą na drodze do wyznaczenia masy cząsteczkowej natywnego DNA są trudności w wyizolowaniu nieuszkodzonych cząsteczek. W czasie preparatyki DNA narażony jest zarówno na napięcia ścinające jak i na działanie enzywów z grupy nukleaz, które mogą trawić łańcuch DNA.

Niektórzy badacze uważają, że masa cząsteczkowa DNA w komórce eukariontów sięga $1{,}7 \cdot 10^{10}$ daltonów, a jego cząsteczka składa się z około 21 podjednostek o masie $8 \cdot 10^{8}$, które połączone są liniowo przez odcinki białkowe. Właśnie w miejscach łączenia się podjednostek cząsteczka bardzo łatwo ulega rozerwaniu. W tym przypadku zakłada się, że strukturą podjednostek jest właśnie heliks Watsona–Cricka.

heliksy liniowe i kołowe

Komórki prokariontów zawierają DNA o masach cząsteczkowych w granicach od 10^{6} do 10^{9} daltonów, a jego struktura przestrzenna może być różna. DNA tych organizmów może być dwuniciowym heliksem liniowym, dwuniciowym heliksem kołowym (kiedy końce jego łańcuchów są połączone ze sobą fosfodwuestrowym wiązaniem kowalencyjnym) bądź też występować w postaci jednoniciowej cząsteczki kołowej lub jednoniciowej cząsteczki liniowej.

Cząsteczki dwuniciowego DNA kołowego mogą tworzyć superheliks (tzn. heliks o odpowiednio dużej średnicy i dużym skoku linii śrubowej) utworzony z heliksu Watsona–Cricka (np. DNA niektórych wirusów, plazmidowy lub chromosom *E. coli*). W organizmach wyższych kołowy DNA dwuniciowy występuje w mitochondriach i chloroplastach. Struktura cząsteczek DNA jednoniciowego liniowego i kołowego prawdopodobnie jest zbliżona do struktury RNA.

Mimo, że struktura przestrzenna DNA była znana od dawna, to przez wiele lat nie udawało się ustalić sekwencji DNA z dwu podstawowych przyczyn. Po pierwsze nie znano tak specyficznych dezoksyrybonukleaz jak rybonukleazy, a po drugie otrzymane preparaty DNA były mieszaniną różnych fragmentów natywnej cząsteczki DNA. Dopiero w 1977 r. udało się ustalić sekwencję DNA bakteryjnego wirusa ØX 174. Ten DNA koduje 9 znanych białek wirusa ØX 174 i zbudowany jest z 5375 nukleotydów, z czego około 10% to sekwencje nie wykorzystywane w translacji. Jest to jednoniciowy DNA kołowy o masie cząsteczkowej $1{,}6 \cdot 10^{6}$ daltonów i długości około 0,6 µm, a zatem jest to jeden z najkrótszych znanych DNA.

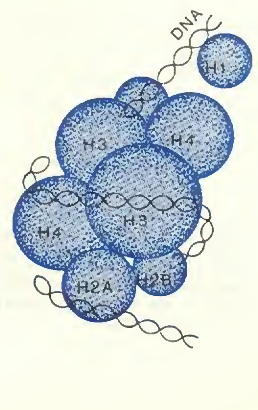
Cząsteczki DNA, których masa cząsteczkowa jest rzędu 10^{10} daltonów mają długość kilkunastu milimetrów. Zrozumiałe jest więc, że cząsteczki te powinny być w jakiś sposób „upakowane", aby mogły pomieścić się w komórce, a w przypadku eukariontów — w jądrze komórkowym.

W komórkach eukariontów DNA znajduje się w chromosomach, a te z kolei w jądrze komórkowym. U różnych gatunków liczba chromosomów jest różna. Problem w jaki sposób cząsteczka DNA rozmieszczona jest w chromosomie, pozostaje nadal otwarty. Wiemy, że dwuniciowy heliks ma średnicę 2 nm, a w komórce oddziałuje z białkami i średnica jego zwiększa się do 3 nm. Przypuszcza się, że cząsteczki białka oddziałując z DNA układają się w większej (głębszej) bruździe heliksu. Prócz tego DNA oddziałuje z grupą białek globularnych, zwanych histonami, tworząc z nimi bardzo specyficzne kompleksy o ściśle określonej strukturze. Kompleks helikalnego DNA, histonów i białek niehistonowych nazywamy chromatyną.

histony

chromatyna

Znamy pięć rodzajów białek histonowych (H1, H2a, H2b, H3 i H4) i ok. 20 białek niehistonowych, które



Rys. 3. Schemat budowy pojedynczego nukleosomu. Widoczne kule symbolizują globule poszczególnych białek histonowych

wchodzą w skład chromatyny. Sposób oddziaływania i rola białek niehistonowych, wchodzących w skład chromatyny, są jeszcze słabo poznane. Znacznie więcej wiemy o kompleksach DNA z histonami, które tworzą tzw. nukleosomy (il. 113, tabl. 28). W skład nukleosomu wchodzi po dwie cząsteczki histonów H2a, H2b, H3 i H4 oraz odcinek heliksu zawierający około 140 par nukleotydowych (rys. 3). Ten odcinek cząsteczki DNA tworzy 1,75 zwoju superheliksu o średnicy około 11 nm i skoku linii śrubowej około 2,8 nm. Nukleosom ma kształt płaskiego walca o średnicy ok. 11 nm i wysokości ok. 5,7 nm. Wyniki te uzyskano z analizy rentgenograficznej kryształów nukleosomów i są one zgodne z wcześniej uzyskanymi wynikami za pomocą rozproszenia promieni rentgenowskich pod małymi kątami lub rozproszenia neutronów przez roztwory nukleosomów.

nukleosomy

Histon H1 tworzy kompleks z nukleosomem, ale nie wiadomo, w jaki sposób i jaka jest jego rola. Przypuszcza się jedynie, że nie wpływa on na strukturę nukleosomu. Nie jest wykluczone, że kompleksuje on z odcinkiem heliksu nie biorącym udziału w superheliksie i składającym się z ok. 40 par nukleotydów. Przyjęcie takiej hipotezy tłumaczyłoby wyniki badań biochemicznych, z których wynika, że powtarzającą się jednostką w chromatynie jest odcinek heliksu złożony z 200 par nukelotydów.

Struktura RNA

Wydawało się, że po ustaleniu helikalnej struktury DNA wyznaczenie struktury kwasów rybonukleinowych nie będzie przedstawiało większych problemów. Okazało się jednak, że przestrzenna struktura RNA w zasadniczy sposób różni się od struktury DNA. Znamy dwuniciowe i jednoniciowe cząsteczki RNA. Struktura oraz wielkość cząsteczki zależą od funkcji, jaką spełniają kwasy rybonukleinowe w żywej komórce. Ze względu na spełnianą funkcję kwasy rybonukleinowe dzieli się na: informacyjne — mRNA, rybosomalne — rRNA i transportujące — tRNA.

Cząsteczki tRNA są bardzo małe, gdyż składają się z 75–85 nukleotydów, a ich masa cząsteczkowa zawiera się w granicach 30 000 daltonów, stała sedymentacji — ok. 4 S.

transportujący kwas rybonukleinowy tRNA

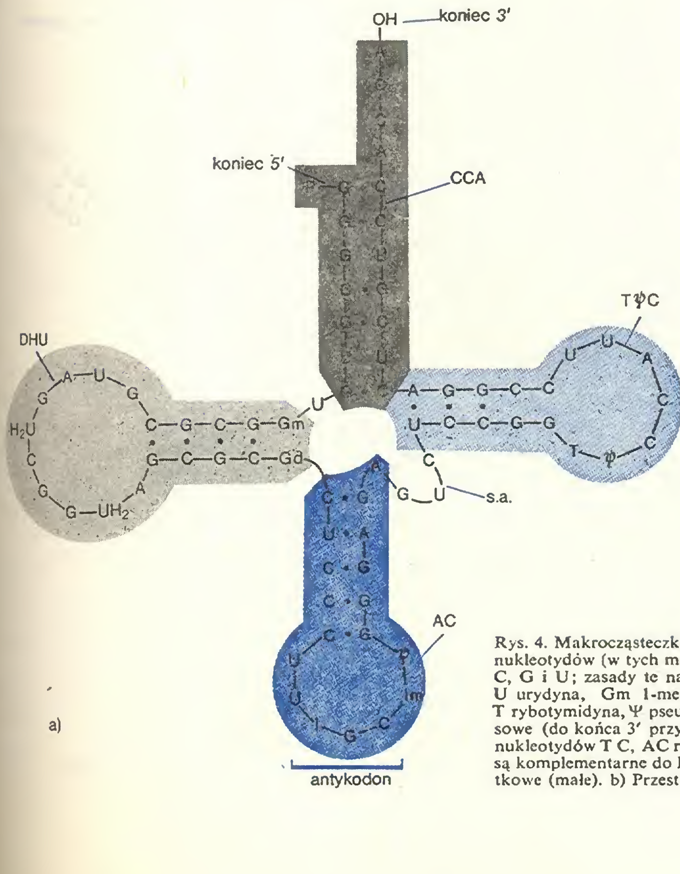
Stała sedymentacji jest to wielkość równa stosunkowi prędkości osiadania cząsteczek do przyspieszenia odśrodkowego wywołującego to osiadanie, a jednostką stałej sedymentacji jest swedberg (1 S = 10^{-13} s).

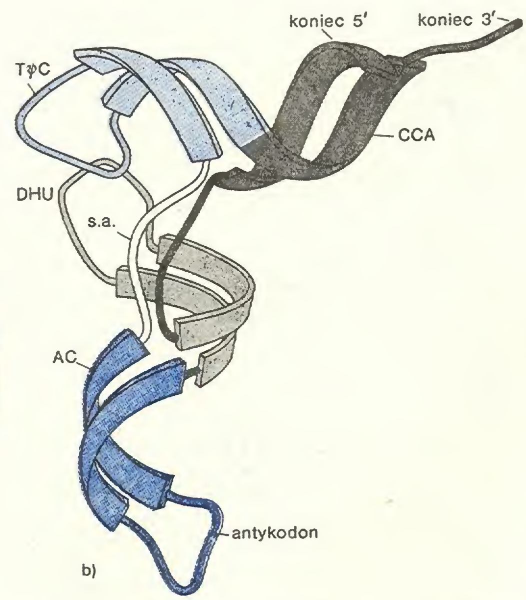
Tak małe cząsteczki jak tRNA nie tworzą włókien lecz kryształy i dzięki temu można, podobnie jak w przypadku białek globularnych, przeprowadzić analizę fourierowską rentgenogramów i ustalić położenie atomów w przestrzeni. Aby jednak tego dokonać otrzymywane monokryształy powinny być odpowiedniej jakości i trwałości. Otrzymanie takich kryształów wymagało około dziesięciu lat pracy naukowców z kilku wyspecjalizowanych laboratoriów na świecie. W międzyczasie przeprowadzono badania innymi metodami, a najcenniejsze wyniki otrzymano za pomocą metody rozproszenia promieni rentgenowskich pod małymi kątami przez roztwory tRNA. Dzięki tym wynikom można było wyznaczyć kształt cząsteczki, tzn. wyznaczyć taką najmniejszą pod względem objętości bryłę geometryczną, w której by się mieściła cała cząsteczka tRNA. Wprowadzając wielkość zwaną elektronowym promieniem bezwładności, analogiczną do promienia bezwładności układu punktów materialnych, można było przewidzieć zagęszczenie atomów w poszczególnych częściach cząsteczki. Metoda ta umożliwiła obliczenie masy przypadającej na jednostkę długości cząsteczki, co z kolei pozwoliło stwierdzić, że ta jednoniciowa cząsteczka powinna zawierać odcinki helikalne podobne do heliksu Watsona–Cricka, lecz utworzone z jednego łańcucha polinukleotydowego.

wyznaczenie kształtu tRNA

Pierwszy model struktury przestrzennej tRNA opracowano w 1968 r. W latach następnych udoskonalono stare lub proponowano nowe modele tej cząsteczki, pod koniec 1974 r. liczba modeli przekroczyła 20. Modele te różniły się głównie ułożeniem poszczególnych części nici polinukleotydowej wewnątrz cząsteczki, a więc różnice dotyczyły tych szczegółów

model tRNA





Rys. 4. Makrocząsteczka tRNA przenoszącego alaninę w komórkach drożdży: a) Kolejność nukleotydów (w tych makrocząsteczkach tRNA występuje wiele zasad różnych od zasad A, C, G i U; zasady te nazywamy zasadami rzadkimi); A adenina, C cytozyna, G guanina, U urydyna, Gm 1-metyloguanina, Gd N^2-dwumetyloguanina, H_2U-dwuhydrourydyna, T rybotymidyna, Ψ pseudourydyna, I inozyna, Im 1-metyloinozyna, CCA ramię aminokwasowe (do końca 3' przyłączany jest aminokwas), TC ramię posiadające w swej pętli trójkę nukleotydów T C, AC ramię antykodonowe (w pętli zaznaczono trójkę nukleotydów, które są komplementarne do kodonu mRNA), DHU ramię dwuhydrouracylowe, s.a. ramię dodatkowe (małe). b) Przestrzenny model makrocząsteczki tRNA wg Kima, 1973 r.

struktury, których nie można było ustalić na podstawie badań rentgenograficznych pod małymi kątami. Na rys. 4b przedstawiono najnowszy model tej cząsteczki, opracowany na podstawie badań rentgenograficznych kryształów tRNA. Większość modeli opracowano dla przypadku tRNA transportującego alaninę w komórkach drożdży, ponieważ w 1965 r. R.W. Holley ze współpracownikami ustalili sekwencję nukleotydów w tym tRNA (rys. 4a). Znajomość sekwencji w znacznej mierze ułatwiała badania strukturalne, które równocześnie nie traciły na ogólności, gdyż struktura wszystkich tRNA jest identyczna. Była to pierwsza cząsteczka wśród kwasów nukleinowych, w której ustalono kolejność nukleotydów. W końcu 1977 r. znano sekwencję ok. 80 cząsteczek tRNA przenoszących różne aminokwasy i pochodzących z różnych organizmów.

wyznaczenie sekwencji tRNA

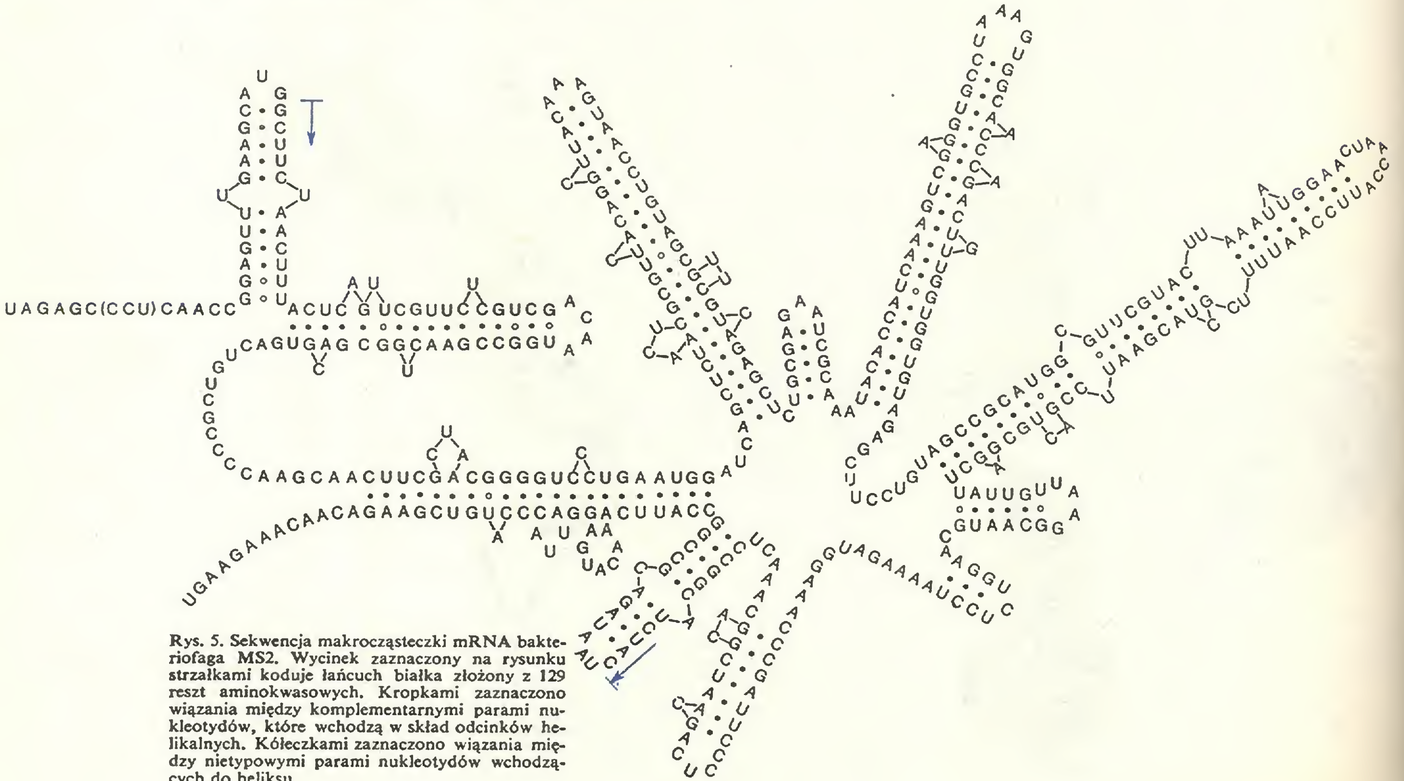
O strukturze pozostałych kwasów rybonukleinowych wiemy znacznie mniej, a główną przyczyną tego jest brak metod fizycznych umożliwiających badanie struktury przestrzennej jednoniciowych cząsteczek o dużych masach cząsteczkowych.

informacyjny kwas rybonukleinowy mRNA

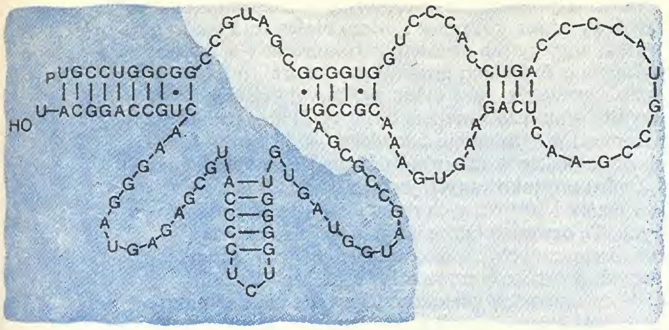
Cząsteczki mRNA stanowią 5–10% ogólnej zawartości kwasów rybonukleinowych w komórce. Ich masa cząsteczkowa jest bardzo różna i zawiera się w granicach $0{,}5\cdot10^6$–$2\cdot10^6$ daltonów, co jest związane z rolą biologiczną tych cząsteczek. Korzystając z doświadczeń w ustalaniu sekwencji tRNA rozpoczęto ustalenie sekwencji mRNA. Ustalono sekwencję mRNA z bakteriofaga MS2 (bakteriofag jest to wirus zakażający bakterie), który ma mRNA kodujący tylko trzy białka. Ustalono, że zawiera ok. 3500 nukleotydów. Na rys. 5 przedstawiono sekwencję wycinka mRNA bakteriofaga MS2, kodującego białko złożone z 129 aminokwasów. Wychodząc z założenia, że istnieje maksymalna liczba wiązań wodorowych między komplementarnymi parami zasad w łańcuchu polinukleotydowym (minimum energii konformacyjnej, objaśnione dalej), zaproponowano tzw. strukturę drugorzędową tego wycinka mRNA. Nie wiemy jednak, jak poszczególne odcinki heliksu są zorientowane względem siebie w przestrzeni oraz względem pozostałej (znacznie większej) części cząsteczki.

rybosomalny kwas rybonukleinowy rRNA

Najmniej wiemy o rybosomalnym kwasie rybonukleinowym, rRNA. Zdołano ustalić, że w jednej komórce istnieją cztery różne typy cząsteczek tego rodzaju kwasów nukleinowych. Ich nazwy wiążą się z szybkością osiadania w polu sił odśrodkowych — stałą sedymentacji. Poszczególne składniki rybosomalnego RNA to: 5S–RNA o masie cząsteczkowej ok. $3\cdot10^4$, 16S–RNA (u organizmów wyższych 18S–RNA) o masie cząsteczkowej ok. $0{,}55\cdot10^6$, 23S–RNA (u organizmów wyższych 28S–RNA) o masie cząsteczkowej ok. $1{,}2\cdot10^6$ daltonów. Ustalono, że 28S–RNA pewnej linii komórek ludzkich hodowanych w laboratoriach, zwanych komórkami HeLa, składa się z co najmniej dwóch łańcuchów; jeden krótki, składający się ze 140 nukleotydów, posiada stałą sedymentacji ok. 5S. Z powodu braku odpowiedniej nazwy ta czwarta z kolei cząsteczka rybosomalnego RNA nosi na razie w literaturze nazwę „RNA związane z 28S–RNA". Dotychczas prowadzono nad tymi kwasami głównie badania biochemiczne, dzięki którym ustalono sekwencje 5S–RNA w kilkunastu wypadkach. Na rys. 6 przedstawiono jedną z poznanych sekwencji cząsteczki 5S–RNA rybosomalnego oraz jedną ze 110 możliwych struktur drugorzędowych tej cząsteczki (wskazanie odcinków helikalnych i tzw. odcinków niestrukturowanych). Z tych 110 struktur tylko 9 można było uznać za mało prawdopodobne, wśród pozostałych 101 szukano dalej takiej, która by była prawdopodobna w odniesieniu do innych cząsteczek 5S–RNA o znanej sekwencji i która by tłumaczyła znane fakty chemiczne i fizyczne. Tak np. badania za pomocą jądrowego rezonansu magnetycznego o wysokiej rozdzielczości wykazały, że w cząsteczce powinno być ok. 28 par nukleotydów połączonych wiązaniami wodorowymi (na modelu 25 par), a wśród nich 4 pary A–U (na modelu 5 par A–U). W modelu cząsteczki 5S–RNA należało także uwzględnić krótkie sekwencje, które nie uczestniczą w wiązaniach wodorowych, lecz wchodzą w skład odcinków nieustrukturowanych, zwanych pętlami, oraz inne fakty doświadczalne.



Rys. 5. Sekwencja makrocząsteczki mRNA bakteriofaga MS2. Wycinek zaznaczony na rysunku strzałkami koduje łańcuch białka złożony z 129 reszt aminokwasowych. Kropkami zaznaczono wiązania między komplementarnymi parami nukleotydów, które wchodzą w skład odcinków helikalnych. Kółeczkami zaznaczono wiązania między nietypowymi parami nukleotydów wchodzących do heliksu



Rys. 6. Płaski model cząsteczki 5S-RNA (*E. coli*) oraz kolejność nukleotydów w jej łańcuchu. W modelu postuluje się tworzenie trzech par U–G; ich wiązania zaznaczono na rysunku kropkami. Istnienie par U–G jest możliwe, lecz są one mniej korzystne energetycznie od par G–C. Uważa się, że część zacieniowana oddziałuje z 50S składnikiem rybosomu

Jak się układają poszczególne części tej cząsteczki w przestrzeni, na razie nie wiemy; należy przypuszczać, że inne jest ułożenie wówczas, gdy w roztworze znajdują się tylko cząsteczki 5S-RNA, a inne, gdy wchodzą one w skład rybosomu.

Przejścia konformacyjne w cząsteczkach kwasów nukleinowych

Jeśli roztwór dwuniciowych heliksów DNA ogrzewać, zmieniać pH roztworu lub stosować jakikolwiek inny czynnik wpływający na strukturę, to wiązania wodorowe między komplementarnymi zasadami ulegają rozerwaniu, a nici polinukleotydowe ulegają rozdzieleniu. Proces ten zachodzi spontanicznie w przedziale ściśle określonych temperatur lub pH. W przybliżeniu jest to przejście fazowe drugiego rodzaju.

denaturacja

Rozdzielone nici polinukleotydowe przyjmują tzw. konformację kłębkową, w której łańcuch cząsteczki posiada dużą giętkość, a ułożenie poszczególnych jego części względem siebie jest dość dowolne. Przejście od struktury helikalnej do struktury kłębka nazywamy przejściem konformacyjnym heliks–kłębek lub denaturacją.

Tworzenie się kłębków obserwowano bezpośrednio za pomocą mikroskopu elektronowego. Natomiast rozdzielenie polinukleotydowych nici DNA obserwowano za pomocą wirowania, w gradiencie gęstości CsCl, natywnych i zdenaturowanych cząsteczek DNA zawierającego nici ^{15}N i ^{14}N (patrz wyżej). Przed denaturacją cząsteczka DNA, którego jedna nić zawiera izotopy azotu ^{15}N, a druga ^{14}N, posiada gęstość 1,717 g/cm^3, zaś po denaturacji otrzymujemy dwie warstwy, z których jedna posiada gęstość 1,724 g/cm^3 (nici zawierające izotop ^{14}N), a druga 1,740 g/cm^3 (nici zawierające izotop ^{15}N). Wzrost gęstości w obu przypadkach spowodowany jest utworzeniem kłębka — struktury o ściślejszym upakowaniu niż heliks.

Przejście konformacyjne heliks–kłębek jest przejściem kooperatywnym, tzn. że rozerwanie wiązań wodorowych między komplementarnymi parami zasad następuje na odcinku heliksu o określonej długości, nie mniejszej niż 4 pary zasad. W przeciwnym wypadku bilans energetyczny procesu byłby niekorzystny.

W sprzyjających warunkach w wielu miejscach heliksu mogą powstawać zalążki denaturacji — centra nukleacji, które z kolei ułatwiają denaturację sąsiadujących odcinków heliksu. Jest to więc proces samonasilający się. Powstawanie wielu zalążków denaturacji potwierdza niezwykle szybko zachodzący proces denaturacji — DNA o masie cząsteczkowej rzędu 10^8 daltonów ulega denaturacji w ciągu około 1 minuty, podczas gdy denaturacja postępująca tylko od obu końców trwałaby kilkadziesiąt dni.

Taka zmiana struktury wiąże się ze zmianą widma elektronowego, lepkości roztworu i stałej sedymentacji i innych wielkości fizykochemicznych charakteryzujących strukturę helikalną DNA. Obserwacje tego procesu prowadzi się przez pomiary odpowiedniej wielkości w zależności od natężenia czynnika wywołującego przejścia strukturalne. Otrzymuje się krzywą zależności podobną do litery S.

znaczenie badania przejścia konformacyjnego

Przejście konformacyjne zależy od wielu właściwości DNA, a jego badanie dostarcza informacji o samych cząsteczkach DNA. Na przykład wykazano, że natywna cząsteczka DNA posiada dwukrotnie większą masę cząsteczkową niż cząsteczka zdenaturowana, co z kolei świadczy o tym, że heliks utworzony jest z dwu a nie dowolnej parzystej liczby nici polinukleotydowych.

Proces przejścia konformacyjnego heliks–kłębek jest analogiczny do procesu topnienia kryształu, dlatego często mówi się o topnieniu DNA, zwłaszcza jeśli czynnikiem denaturującym jest temperatura. Jeśli czynnikiem wywołującym przejście konformacyjne jest temperatura, to można otrzymać następujące informacje.

Można wyznaczyć temperaturę przejścia (topnienia) T_m — temperaturę, w której 50% heliksu zostało stopione.

Można wyznaczyć szerokość przejścia konformacyjnego (ΔT_m) — czyli różnicę odpowiadającą stopieniu $^3/_4$ heliksu i $^1/_4$. Jeśli preparat DNA był jednorodny, to przejście będzie odbywało się w wąskim przedziale temperatur. Preparaty niejednorodne będą się charakteryzowały dużymi wartościami ΔT_m. Szerokość przejścia konformacyjnego zależy również od stężenia soli w roztworze. Przy małych siłach jonowych przejście odbywa się w szerszym zakresie temperatur, a przy dużych siłach jonowych w węższym.

Energia komplementarnej pary A–T jest mniejsza od pary G–C, dlatego też wartość T_m jest proporcjonalna do zawartości par G–C w heliksie. Zatem znając wartość T_m badanego DNA można wyznaczyć jaki procent stanowią w nim pary G–C. Gdyby zaś w heliksie były długie odcinki różniące się w znaczny sposób zawartością par G–C, to przejście może zachodzić wielostopniowo.

renaturacja

Proces odwrotny do procesu denaturacji nazywamy procesem renaturacji lub przejściem konformacyjnym kłębek–heliks. Proces ten następuje wtedy, gdy usuniemy czynnik powodujący denaturację. W omawianym powyżej przypadku będzie to obniżenie temperatury. Z zasady powinien to być proces fazowy drugiego rodzaju, a w rzeczywistości zależy od cech renaturującego DNA i warunków, w jakich zachodzi renaturacja. Łączenie dwu nici polinukleotydowych w procesie renaturacji zachodzi w sposób kooperatywny.

Renaturacja DNA uwarunkowana jest procesem dyfuzji, który określa prawdopodobieństwo spotkania się dwu komplementarnych nici DNA. Współczynnik dyfuzji zależy od temperatury i rozmiarów cząstek renaturujących. W niskich temperaturach (około 4°C) cząsteczki DNA będą poruszały się w roztworze najwolniej. W tym zakresie temperatur najczęściej spotykamy łączenie się fragmentów dwu łańcuchów DNA o nieodpowiednich sekwencjach, ponieważ nie ma źródła energii umożliwiającej rozerwanie powstałych wiązań wodorowych między kilkoma nukleotydami. Współczynnik dyfuzji będzie zmniejszał się wraz ze wzrostem ciężaru cząsteczkowego DNA i będzie hamował proces renaturacji.

Proces renaturacji będzie zależał również od siły jonowej roztworu, ponieważ cząsteczki DNA posiadają znaczne odpychające się ładunki elektryczne. Kationy znajdujące się w roztworze ekranują te ładunki, zmniejszając w ten sposób siły odpychania elektrostatycznego.

Prawdopodobieństwo spotkania się dwu komplementarnych nici DNA będzie zależało od stężenia DNA i czasu trwania renaturacji. Im większe jest stężenie DNA, tym mniejsza jest średnia przebyta droga konieczna do spotkania innej cząsteczki DNA, a im dłuższy jest czas trwania dyfuzji, tym dłuższa jest średnia przebyta droga.

Oprócz zjawisk dyfuzji na stopień renaturacji ma wpływ sekwencja DNA. Na przykład znacznie łatwiej będą renaturowały cząsteczki typu $d(G)_n \cdot d(C)_n$ niż DNA, w którym występują wszystkie cztery zasady i do tego połączone w określonej kolejności.

Aktywność biologiczna kwasów nukleinowych

Od czasu, gdy w 1944 r. O. T. Avery ze współpracownikami udowodnili, że żywe bakterie mogą wchłonąć DNA pochodzący z innych bakterii i uzyskać w ten sposób dziedzicznie utrwaloną nową cechę genetyczną, charakterystyczną dla bakterii — dawcy DNA, nagromadzono wiele bezpośrednich dowodów doświadczalnych na to, iż DNA jest tą substancją, w której się zawiera i w postaci której jest przenoszona informacja genetyczna żywych organizmów.

DNA — nośnik informacji genetycznej

Watson i Crick, postulując strukturę helikalną DNA, zaproponowali również hipotetyczny mechanizm wiernego powielania się tej struktury poprzedzającego podział komórki. Lata, które nastąpiły po ogłoszeniu hipotezy Watsona–Cricka, były latami intensywnych badań mechanizmów rządzących przekazywaniem informacji genetycznej oraz powielaniem, replikacją materiału genetycznego.

Warto może zdać sobie sprawę z fizycznych rozmiarów aparatu genetycznego żywych organizmów. W pojedynczej somatycznej komórce organizmów wyższych znajduje się wiele cząsteczek DNA, różnych pod względem kolejności nukleotydów i długości łańcucha. Gdyby się np. wszystkie takie cząsteczki zawarte w komórce ludzkiej rozciągnęło wzdłuż prostej, w postaci pojedynczego heliksu Watsona–Cricka, miałyby one łączną długość 174 cm, a ważyłyby 5 pikogramów. Te cząsteczki DNA zorganizowane są w pewnych okresach życia komórki w struktury zwane chromosomami; w każdej komórce somatycznej człowieka jest 46 chromosomów. Długość pojedynczej cząsteczki DNA składającej się na największy chromosom wynosi 7,3 cm. Sam chromosom ma długość 0,001 cm. Już zestawienie liczb wskazuje na to, że DNA musi tworzyć w chromosomie zwarte struktury, które jednocześnie dopuszczają wierną replikację DNA przed podziałem komórki i rozdział powielonego DNA między dwie potomne komórki. Niestety, mimo iż w tym kręgu zagadnień pracuje wielu badaczy, wiedza o tych strukturach jest jeszcze niepełna.

chromosomy

DNA różnych organizmów

Łączna długość zawartych w pojedynczej komórce danego organizmu cząsteczek DNA nie zawsze zależy od miejsca tego organizmu na drzewie ewolucyjnym. I tak np. ssaki, płazy i gady posiadają podobny zasób DNA, ale istnieją również płazy, których komórki zawierają 25 razy więcej DNA niż komórki ssaków. Ponieważ część łańcucha DNA w organizmach wyższych odgrywa rolę strukturalną, a nie uczestniczy w przekazywaniu informacji genetycznej, taka sytuacja jest możliwa. Jednak ogólnie biorąc, w komórkach ptaków i roślin wyższych jest średnio 3 razy mniej DNA niż w komórkach ssaków, w bakteriach 100–1000 razy mniej, w niektórych małych wirusach — milion razy mniej. Niemniej jednak w tej ostatniej grupie organizmów DNA musi się charakteryzować strukturami wyższego rzędu niż heliks Watsona–Cricka; pojedyncza cząsteczka DNA bakterii *E. coli*, stanowiąca jej cały aparat genetyczny, ma 0,12 cm długości, podczas gdy sama komórka jest tysiąc razy krótsza.

Informacja genetyczna komórki, zakodowana w jej DNA wyrażana jest przez syntezę białek tej komórki. Białka, enzymy lub elementy strukturalne (→ Białka) stanowią o funkcjach życiowych komórki, o jej fenotypie. Ocenia się, że w cyklu życiowym komórki bakteryjnej syntetyzowane jest ok. 3 tys. różnych białek, natomiast w organizmie człowieka — ok. 5 mln. Można obliczyć, że w gatunkach żyjących na Ziemi (ok. 1,2 mln) produkowanych jest ok. 10^{12} różnych rodzajów białek i 10^{10} różnych rodzajów kwasów nukleinowych. Tę ogromną liczbę warto porównać z liczbą ok. 10^6 organicznych związków chemicznych syntetyzowanych dotychczas przez człowieka.

W organizmach wielokomórkowych każda komórka określonego osobnika zawiera taki sam komplet cząsteczek DNA. Jednakże wyspecjalizowane tkanki produkują wybrany rodzaj białek, który nie jest syntetyzowany w innych komórkach tego samego organizmu. Tak więc w organizmach wielkokomórkowych muszą istnieć również mechanizmy wyboru tych informacji genetycznych, które powinny być realizowane w zależności od wieku organizmu lub specjalizacji danej grupy komórek, tkanki. Okazuje się również, o czym wspomniano wyżej, że nie wszystkie sekwencje DNA organizmów wyższych mają sens genetyczny (realizowane są jako białka), niemniej jednak odgrywają one określoną rolę fizjologiczną. Niektóre uczestniczą w powstawaniu struktur DNA wyższego rzędu niż heliks Watsona–Cricka.

Zadziwiającą cechą żywej komórki jest jej zdolność do bezbłędnej niemal samoreprodukcji przez setki i tysiące pokoleń. Stabilność DNA przewyższa większość zapisów informacji pozostawionych przez ludzkość; bakterie zachowały cechy biochemiczne i morfologiczne podobne do cech przodków sprzed milionów lat. Jednocześnie indywidualna, konkretna nić cząsteczki DNA, jest tak nietrwała poza organizmem, że ulega rozerwaniu nawet przy przechodzeniu przez rurki szklane o średnicy milimetra. Cząsteczka DNA może się stać również celem ataku różnych destrukcyjnych enzymów znajdujących się w komórce, jak również ulegać zmianom chemicznym pod wpływem takich czynników zewnętrznych jak promieniowanie, temperatura, niektóre związki przenikające do komórek. Zachowanie zatem genetycznej stałości gatunków jest uwarunkowane istnieniem w żywych organizmach skutecznych układów naprawczych, które szybko usuwają za pomocą enzymów każde lub prawie każde uszkodzenie powstające w DNA i zapewniają bezbłędne powielanie DNA, czyli bezbłędną replikację.

stabilność DNA

Replikacja DNA

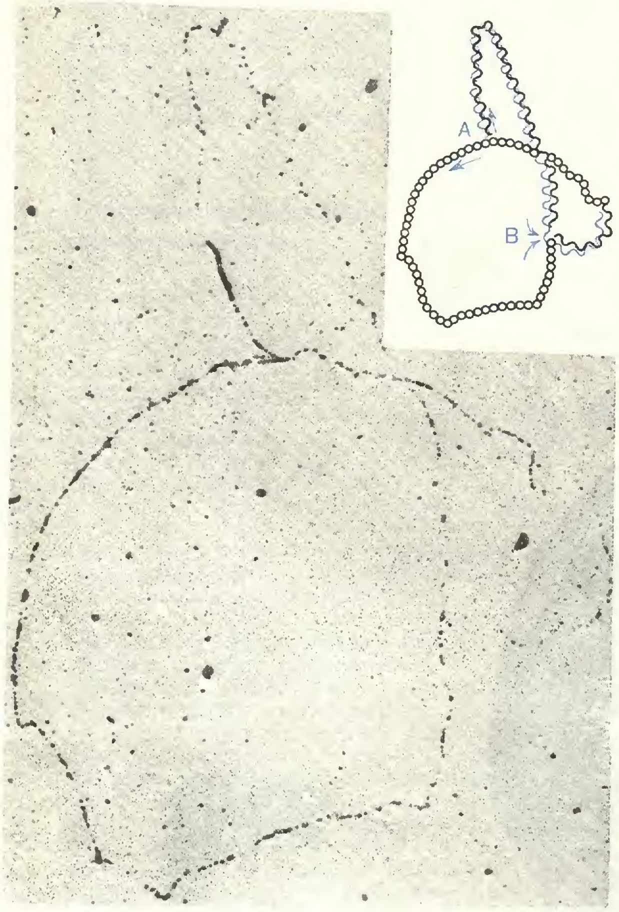
Replikacja DNA jest procesem złożonym i do dziś nie do końca zrozumianym. Najlepiej poznane są procesy replikacji DNA w bakteriofagach i w bakteriach, mniej wiadomo o nich w wyższych organizmach, posiadających materiał genetyczny zlokalizowany wewnątrz jądra (eukariontach). Nie ulega obecnie wątpliwości, że nawet w najprostszych organizmach replikacja przebiega w złożonym, wieloskładnikowym i wieloenzymatycznym kompleksie, prawdopodobnie związanym ze strukturą białkowo-lipidową błony komórkowej. Nie jest łatwo określić składniki takiego kompleksu, gdyż utrzymujące go w stanie funkcjonalnym oddziaływania, głównie między białkami, należą do oddziaływań słabych, kompleks replikacyjny ulega łatwo rozpadowi w czasie prób wydzielania i oczyszczenia jako całości. Z genetycznych badań replikacji w bakteriach wiadomo od dość dawna, że zależy ona od wielu białek, np. co najmniej od 13 w *E. coli*, jednak ich rola w replikacji jest w większości przypadków nieznana, lub w najlepszym razie nie do końca wyjaśniona. Zidentyfikowana i oczyszczona do homogenności jest polimeraza DNA, duże białko en-

polimeraza DNA



Rys. 7. Model replikacji DNA. Dwa komplementarne łańcuchy macierzystego DNA ulegają lokalnej denaturacji, tworząc widełki replikacyjne. Polimeraza DNA dobudowuje łańcuchy potomne do rozdzielonych łańcuchów matrycy. W dole rysunku ostatnia para nukleotydów (G–C) została narysowana nie symbolicznie, lecz w postaci odpowiednich wzorów chemicznych

zymatyczne, które syntetyzuje z nukleotydów potomne łańcuchy DNA, komplementarne do łańcuchów macierzystych (widać to na rys. 7 i 8). Kolejność nukleotydów łańcucha macierzystego wyznacza jednoznacznie kolejność nuekleotydów w łańcuchu



Rys. 8. Autoradiogram kolistej cząsteczki DNA bakteryjnego, w trakcie replikacji (powiększenie ok. 240 razy). Hodowane na nieradioaktywnym podłożu bakterie przeniesiono na pozywkę zawierającą znakowany promieniotwórczym izotopem (trytem) nukleotyd, który polimeraza DNA wbudowała do replikowanego DNA. Odcinki intensywnie zaczernione przedstawiają DNA znakowany w obu łańcuchach podwójnej spirali (dwie linie czarne na rysunkowym schemacie objaśniającym), odcinki jaśniejsze przedstawiają DNA znakowany w jednym łańcuchu (linie niebiesko-czarne na schemacie). Zdjęcie przedstawia sytuację przy końcu drugiej rundy replikacji po dodaniu izotopu

potomnym dzięki zachowaniu zasady komplementarności nukleotydów jednego łańcucha do nukleotydów drugiego. Przypomnijmy, że naprzeciw nukleotydu A zostanie wbudowany do nowego łańcucha nukleotyd T, a naprzeciw nukleotydu G — nukleotyd C.

replikacja półkonserwatywna

Z jednej cząsteczki macierzystej powstają zatem dwie identyczne cząsteczki potomne DNA, każda złożona z jednego łańcucha macierzystego i jednego nowo syntezowanego. Mechanizm ten nazwano replikacją półkonserwatywną. W wielu badanych systemach replikacyjnych wykazano istnienie białek specyficznie destabilizujących strukturę heliksu Watsona-Cricka, wiążących się kooperatywnie i specyficznie z pojedynczołańcuchowym DNA, ułatwiających denaturację macierzystego DNA w rejonie replikacji. Jedno z białek rozplątujących podwójnołańcuchowy DNA hydrolizuje jednocześnie ATP (→ Organizacja procesów życiowych komórki), jego funkcja życiowa jest nieznana. Odkryto również niezbędny w replikacji enzym, gyrazę (inaktywacja gyrazy powoduje natychmiastowe zatrzymanie replikacji), która badana *in vitro* podwyższa liczbę superhelikalnych zwojów DNA, ułatwiając w ten sposób w innym rejonie DNA rozplątanie skrętów heliksu Watsona-Cricka. Ten enzym również wymaga jednoczesnej hydrolizy ATP. Wśród pozostałych białek identyfikowanych jako uczestniczące w procesach replikacji wiele wywołuje również hydrolizę niektórych wysokoenergetycznych niskocząsteczkowych substancji. Wydzielająca się wówczas energia prawdopodobnie zużywana jest na wprowadzenie allosterycznych zmian w białkach kompleksu replikacyjnego, rozkręcenie podwójnego heliksu DNA, ułatwia utrzymywanie się replikacyjnych zespołów białkowych w całości. Żadne z odkrytych dotychczas białek nie dostarcza samo wg ilościowych ocen, wystarczającej energii do tych procesów.

starter

Interesującą cechą wszystkich znanych polimeraz DNA jest to, że nie potrafią one rozpocząć syntezy nowego łańcucha DNA. Prawdopodobnie syntezę taką rozpoczyna inny enzym, polimeraza RNA, budująca na wzorcu, łańcuchu DNA, jego komplementarną kopię RNA zwaną starterem. Nie wiadomo, co jest sygnałem do zmiany polimeraz, w pewnym stadium syntezy startera polimeraza RNA zostaje zastąpiona przez polimerazę DNA, która kontynuuje polimeryzację budując już potomny łańcuch DNA. Tak przebiegająca synteza DNA jest nieciągła, polimeraza DNA syntetyzuje krótkie fragmenty łańcucha, zwane od nazwiska odkrywcy tego zjawiska fragmentami Okazaki; w trakcie jednej rundy replikacji dochodzi zatem wielokrotnie do inicjacji fragmentów Okazaki przez polimerazę RNA i elongacji ich przez polimerazę DNA. Startery RNA są w późniejszym okresie replikacji usuwane przez inne enzymy, powstająca na ich miejscu luka zapełniana przez polimerazę DNA, a fragmenty Okazaki łączone w łańcuch ciągły enzymem, ligazą.

Replikacja DNA rozpoczyna się zawsze od ściśle zdefiniowanego odcinka macierzystego DNA zwanego punktem początkowym replikacji. Przyczyna wyboru przez kompleks replikacyjny punktu początkowego jest nieznana, w niektórych wypadkach został on zlokalizowany z dokładnością do kilkunastu lub kilkudziesięciu nukleotydów. Wiadomo również, że inicjacja replikacji jest kontrolowana genetycznie i że uczestniczą w niej specyficzne białka inicjatorowe. W niektórych przypadkach replikacja biegnie w jednym kierunku od punktu początkowego, w niektórych — jednocześnie w obie strony.

Szybkość replikacji w bakteriach (w temperaturze 37°) ocenia się na 10^5 nukleotydów polimeryzowanych w ciągu minuty. W przypadku, gdy do łańcucha potomnego włączony zostanie nukleotyd niekomplementarny istnieje możliwość bezpośredniego poprawienia tego błędu przez polimerazę DNA i towarzyszące enzymy drogą wymiany niewłaściwego nukleotydu na prawidłowy. Wierność replikacji jest bardzo wysoka, błąd popełniany jest nie częściej niż raz na 10^{10} włączonych nukleotydów. Jednak w pewnych warunkach do replikującego się DNA może zostać wbudowany nieprawidłowy nukleotyd, lub prawidłowy ulec modyfikacji chemicznej. Może to stać się przyczyną mutacji genetycznej, tzn. dziedziczonej zmiany, która może okazać się korzystna lub niekorzystna dla organizmu. Procesy ewolucyjne utrwalać będą mutacje korzystne dla gatunku, eliminować — mutacje obniżające żywotność gatunku.

mutacja genetyczna

Przyczynami mutacji mogą być również zmiany chemiczne w już syntetyzowanej cząsteczce DNA, wywołane przez czynniki zewnętrzne: światło, promieniowanie ultrafioletowe, jonizujące i inne. Nie zawsze przyczyna zmian jest nam znana, jeśli nie — mówimy o mutacjach spontanicznych. Zdarzają się zmiany obejmujące fragment łańcucha DNA, np. podwojenie pewnych sekwencji łańcucha, dodanie (addycja) lub wypadnięcie (delecja) części łańcucha DNA czy jednego nukleotydu. Jeśli przyczyną mutacji jest modyfikacja jednego nukleotydu, to wtedy

mutację tę nazywa się punktową. Mutacje punktowe, w odróżnieniu od delecyjnych są stosunkowo łatwo odwracalne.

Prawdopodobieństwo mutacji spontanicznej, zmieniającej funkcję białka u bakterii jest niewielkie (10^{-7}–10^{-9} na jedno pokolenie komórek), ale ze względu na dużą liczebność bakterii w środowisku jest ono źródłem pojawiania się mierzalnych ilości zmienionych osobników (mutantów). Przykładem mutacji niekorzystnych dla gatunku są tzw. choroby genetyczne.

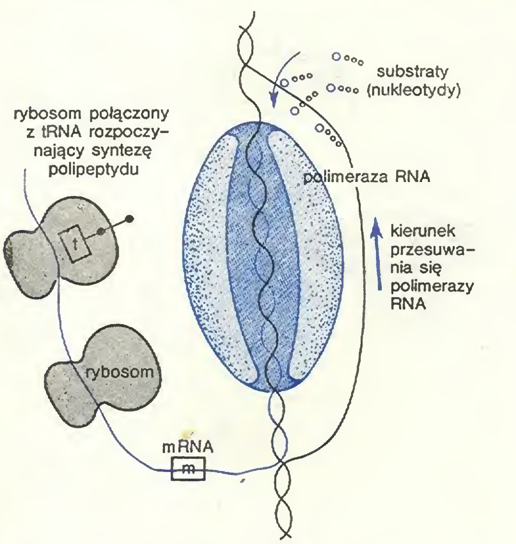
choroby genetyczne

Dobrze poznanym przykładem takiej choroby jest złośliwa anemia sierpowata, utrzymująca się wśród afrykańskich plemion murzyńskich. Białko hemoglobiny (masa cząsteczkowa ok. 67 tys. daltonów) wydzielone z krwinek ludzi chorych różni się od białka normalnej hemoglobiny tylko tym, że aminokwas — kwas glutaminowy zastąpiony został przez aminokwas — walinę. Różnica ta, wydawałoby się nieznaczna, powoduje, że ludzie dotknięci anemią sierpowatą umierają w wieku młodzieńczym. Opisano wiele innych nieprawidłowych hemoglobin i stwierdzono, że w każdym wypadku zmiana funkcji wynika z niewielkiej zmiany w składzie aminokwasowym globiny ludzi chorych.

Przekazywanie informacji genetycznej z DNA do komórki jest wynikiem dwu procesów: transkrypcji i translacji.

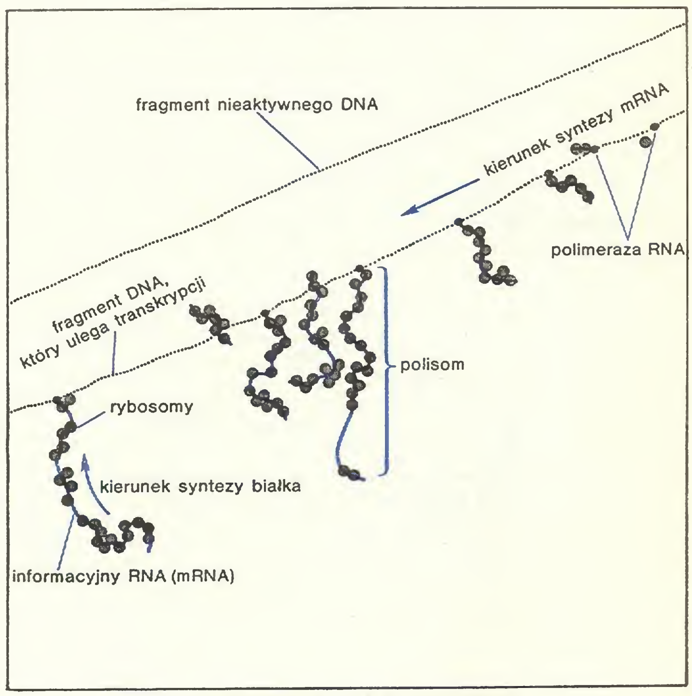
Transkrypcja informacji genetycznej

Określenie transkrypcja obejmuje procesy syntezy RNA wg wzorca DNA, a więc polega na przeniesieniu informacji genetycznej z jednego kwasu nukleinowego (DNA) na drugi (RNA). W procesach transkrypcji uczestniczą: DNA jako matryca, enzym polimeryzujący (polimeraza RNA), białka, niskocząsteczkowe substraty i substancje regulatorowe, tzw. efektory (rys. 9, 10 i il. 119, tabl. 29). Polimeraza RNA rozpo-



Rys. 9. Model transkrypcji RNA. Syntetyzowany na matrycy lokalnie zdenaturowanego DNA informacyjny RNA (mRNA) łączy się z rybosomami i tworzy kompleks translacyjny





Rys. 10. Dwa odcinki DNA bakteryjnego (zdjęcie w mikroskopie elektronowym, powiększenie 150 000 razy). W dolnym fragmencie DNA ulega transkrypcji, syntetyzowany jest mRNA, który jeszcze w trakcie transkrypcji łączy się z rybosomami, tworząc polisom. Rozpoczyna się synteza polipeptydu (translacja). W prawej górnej części zdjęcia widać globularne cząsteczki polimerazy RNA rozpoczynającej transkrypcję, w lewej dolnej części — najstarszy mRNA widoczny na zdjęciu, będący w trakcie intensywnego procesu translacji

znaje, nie wiadomo jeszcze jak, ten fragment łańcucha DNA, od którego się rozpoczyna transkrypcja. Rozpoczęcie transkrypcji oraz jej kontynuacja związane są z koniecznością lokalnego rozplątania heliksu matrycy DNA, gdyż transkrypcji ulega tylko jeden łańcuch DNA. W ogromnej większości poznanych przypadków polimerazy RNA są białkami składającymi się z podjednostek, w polimerazie bakteryjnej jedna z nich, tzw. białko sigma, ułatwia jedynie specyficzne rozpoczęcie transkrypcji, a następnie ulega odłączeniu od kompleksu transkrybującego. Polimerazy RNA, jak już wiemy, w odróżnieniu od polimerazy DNA,

mogą rozpoczynać transkrypcję *de novo*, nie wymagając do tego startera.

Nie wyjaśniono również mechanizmu kończenia transkrypcji z określonego odcinka DNA, nie wiadomo, jaki sygnał decyduje o rozpadzie transkrypcyjnego kompleksu i oddzieleniu się polimerazy i produktu (RNA) od matrycowego DNA. Być może w procesie tym odgrywają rolę dodatkowe białka. Jednym z takich białek jest np. tzw. czynnik rho, w obecności którego transkrypcja kończy się we właściwym dla danego genu rejonie DNA.

prekursor RNA

Ważną cechą procesu transkrypcji, poznaną stosunkowo niedawno, jest to, że jego pierwotny produkt, tzw. prekursor RNA, jest dłuższy niż odpowiadający mu RNA biologicznie aktywny. Prekursor taki podlega następnie procesowi dojrzewania, w którym uczestniczą liczne, często wysokospecyficzne enzymy. Proces dojrzewania polega na skracaniu łańcucha prekursora oraz na dodatkowych modyfikacjach, np. metylacji lub dodawaniu końcowych fragmentów, nie mających swoich komplementarnych odpowiedników w matrycowym DNA. Prekursory mRNA komórek eukariontów syntezowane w jądrze składają się w ok. 10% z właściwych sekwencji mRNA, który pełni swą funkcję następnie poza jądrem, w cytoplazmie; pozostałe 90% to sekwencje dodatkowe strukturotwórcze lub chroniące mRNA przed nukleazami, nigdy nie opuszczające jądra. Prawie we wszystkich mRNA organizmów eukariotycznych jeden z końców składa się tylko z nukleotydów adeniny (A) dodanych enzymatycznie do mRNA już po transkrypcji. Ta właściwość pozwoliła na łatwe oddzielenie preparatywne mRNA od całej puli innych kwasów nukleinowych komórki, umożliwiła w następstwie oczyszczanie mRNA pojedynczych genów, oznaczanie sekwencji ich podjednostek nukleotydowych, a także zastosowanie nowych technik inżynierii genetycznej (zob. niżej) do syntezy genów w probówce.

Warto tu również dodać, że zarówno prekursor mRNA w jądrze, jak i mRNA w cytoplazmie, nie występują nigdy w formie wolnego kwasu nukleinowego, a jedynie w kompleksach ze specyficznymi białkami tworząc tzw. informosomy. W kompleksie tym rola białek może polegać na blokowaniu pewnych odcinków mRNA, a więc może być rolą regulatora aktywności biologicznej mRNA. Jedną z większych sensacji genetyki molekularnej ostatnich lat stało się odkrycie, iż w organizmach eukariotycznych niektóre geny są podzielone na fragmenty, oddzielone jeden od drugiego odcinkami DNA o długości porównywalnej do długości genu, których nie odnajduje się następnie w dojrzałym transkrybowanym mRNA tego genu. Fakt ten sugeruje istnienie nowych, nieznanych dotąd sposobów regulacji wyrażania aktywności określonych genów organizmów eukariotycznych.

informosomy

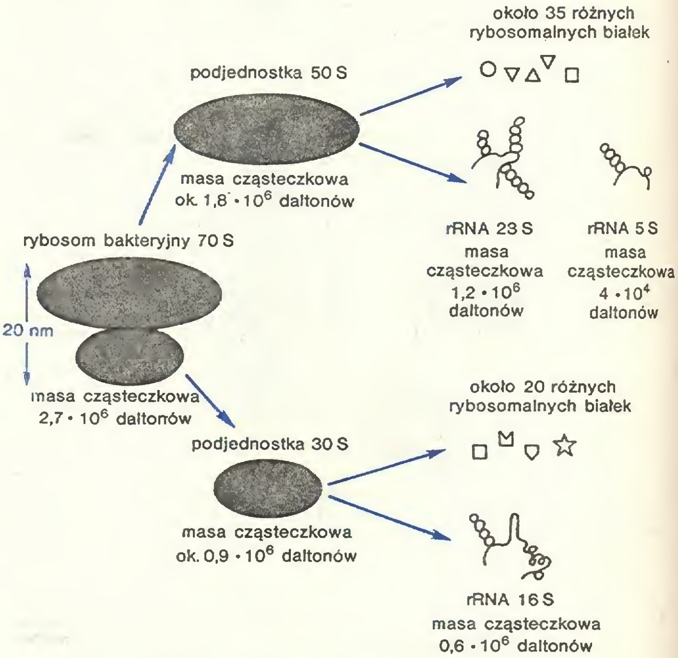
Wszystkie typy RNA syntetyzowane w procesie transkrypcji (mRNA, tRNA, rRNA) grają rolę w procesach translacji, czyli w procesie przeniesienia informacji genetycznej z kwasu nukleinowego (mRNA) na białko (rys. 9).

Translacja informacji genetycznej

Jak wyjaśniono uprzednio, kwasy nukleinowe składają się z 4 rodzajów podjednostek, nukleotydów, białka zaś — z 20 podjednostek, aminokwasów. Dlatego też translacja byłaby niemożliwa przy założeniu, że jednemu nukleotydowi odpowiada jeden aminokwas, jak również w wypadku sekwencji dwunukleotydowych, tzn. kombinacji, w której dwóm nukleotydom odpowiada jeden aminokwas, bo takim sekwencjom przypisać można tylko 16 różnych aminokwasów. Z prostych reguł kombinatoryki wynika, że dopiero trójnukleotydowe „litery" (64 możliwych kombinacji) składają się na „alfabet" bogatszy niż dwudziestoliterowy „alfabet" aminokwasowy. Okazało się, że istotnie w przyrodzie ustalił się powszechny kod trójkowy. Trójce nukleotydów w mRNA w określonej kolejności (kodonowi) odpowiada jeden i tylko jeden aminokwas. Badania te pozwoliły na określenie następujących cech kodu genetycznego: a) Kod jest uniwersalny, tzn. taki sam w całym świecie istot żywych, od wirusa do człowieka. Fakt ten świadczy o tym, że ewolucja kodu została bardzo dawno zakończona i że stanowi ona optymalne rozwiązanie problemu translacji genetycznej. b) Kod jest trójkowy, ciągły i nie zachodzący, tzn. odczytywane są kolejne trójki nukleotydowe w łańcuchu mRNA i każdy z nukleotydów wchodzi w skład tylko jednego kodonu. Dzięki tej właściwości kodu pojawienie się w łańcuchu białka danego aminokwasu nie zależy od rodzaju poprzedzającego i następującego po nim aminokwasu. c) Kod jest zwyrodniały, tzn. ten sam aminokwas może być zapisany więcej niż jedną trójką nukleotydów. Ta cecha kodu pozwala na syntezę dowolnych białek również i w organizmach, których DNA jest szczególnie bogate w niektóre nukleotydy, a ubogie w inne, tym niemniej wystarcza im kodonów do wyrażenia wszystkich aminokwasów składających się na białka konieczne do życia. d) Kod jest współliniowy, tzn. kolejność kodonów odpowiada kolejności wbudowanych do białka aminokwasów, odpowiadających kodonom. e) Kod jest odczytywany od określonego punktu startowego w sposób ciągły, znane są kodony początkujące i kończące odczytywanie danej informacji.

kod genetyczny

Czasem znajomość kodu genetycznego pozwala na poszukiwanie i izolowanie mRNA, matrycy w syntezie określonych białek. Tak np. produkowane przez poczwarki jedwabnika białko, fibroina, ma bardzo szczególny skład aminokwasowy; zawiera 45% aminokwasu glicyny, reszta zaś to tylko dwa aminokwasy: alanina i seryna. W kodonach odpowiadających tym aminokwasom przeważa nukleotyd guanylowy. Z poczwarek jedwabnika udało się wydzielić RNA bogaty



Rys. 11. Schemat struktury rybosomu bakteryjnego. W skład aktywnego biologicznie rybosomu (70S) wchodzą dwie różnej wielkości podjednostki (30S i 50S). Każda podjednostka składa się z rRNA i rybosomalnych białek. Różne geometryczne figury w prawej części rysunku symbolizują różnice w budowie między poszczególnymi białkami rybosomalnymi; skręcone nici — struktury przestrzenne rybosomalnych RNA
Stała sedymentacji jest proporcjonalna do masy cząsteczkowej makrocząsteczki podniesionej do potęgi α ($s \sim M^{\alpha}$, α — współczynnik zależny m.in. od kształtu makrocząsteczki)

w nukleotyd guanylowy; stanowił on aż 1,4% ogólnej puli RNA poczwarki i posiadał rzeczywiście cechy mRNA fibroiny.

Można przypomnieć tu również omawiany już przykład anemii sierpowatej, której przyczyną molekularną jest zmiana jednego aminokwasu w białku krwi, globinie. Dzięki ustaleniu zapisu kodu genetycznego wiemy, że kodonami kwasu glutaminowego są dwie trójki nukleotydów, GAA i GAG, natomiast waliny — GUA i GUG. Tak więc anemia sierpowata jest wynikiem zastąpienia w mRNA reszty A przez resztę U. Oznacza to również, że w odpowiadającym temu mRNA odcinku DNA nastąpiła zmiana nukleotydu A na nukleotyd T (mutacja).

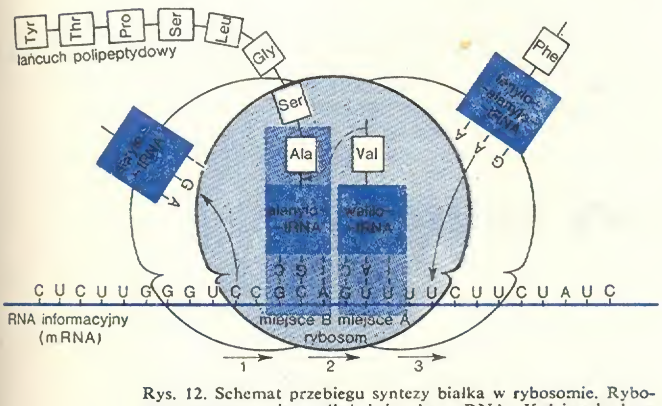
Tak więc wszystkie informacje genetyczne zostały zapisane językiem trójek nukleotydów DNA, następnie przepisane w procesie transkrypcji do RNA. Odczytanie kodu odbywa się w procesie translacji, czyli syntezy białka, zgodnie z kolejnością ułożenia kodonów w łańcuchu mRNA.

powstawanie kompleksów mRNA z rybosomami

Odczytywanie RNA informacyjnego wymaga powstania kompleksu mRNA z rybosomami (tzw. polisomu). Budowa i skład rybosomów bakteryjnych pokazana jest schematycznie na rys. 11. Rybosomalne białka i rRNA zostały wydzielone, oczyszczone i częściowo zbadane. Udało się również uzyskać funkcjonalnie aktywny rybosom po zmieszaniu w odpowiednich warunkach uprzednio rozdzielonych i oczyszczonych składników. Udało się również zbadać niektóre oddziaływania wzajemne składników rybosomu, scharakteryzować część sił utrzymujących strukturę rybosomu, prowadzone są intensywne badania roli biologicznej rybosomalnego RNA i poszczególnych białek rybosomalnych oraz jakim zmianom ulega struktura w czasie funkcjonowania rybosomu.

przyłączanie aminokwasów do tRNA

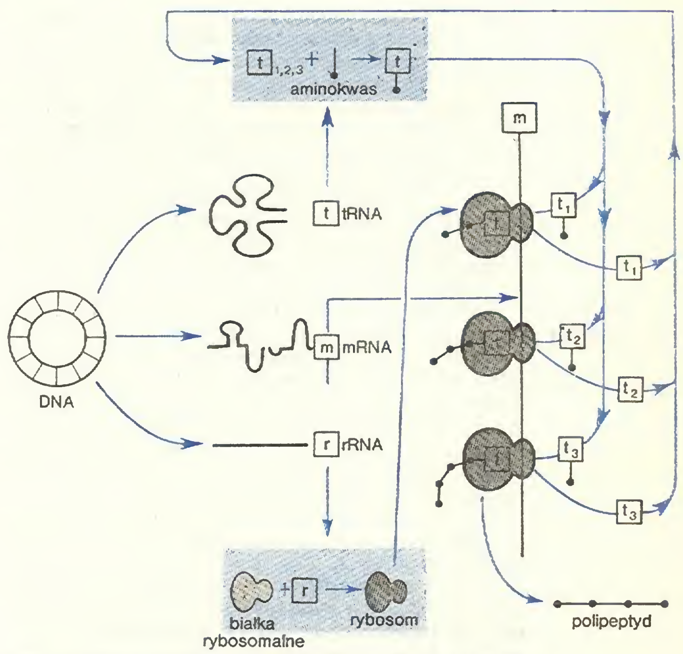
Aminokwasy nie zbliżają się do mRNA jako swobodne cząsteczki, lecz przyłączone do tRNA. Przyłączenie to katalizowane jest przez zespół enzymów, syntetaz aminoacylo-tRNA. Każdy aminokwas może być przyłączony do cząsteczki swojego specyficznego tRNA tylko przez taką syntetazę. Specyfika tRNA polega na tym, że w jego cząsteczce znajduje się trójka nukleotydów zwana antykodonem (rys. 4a). Antykodon i kodon mRNA są komplementarne, dzięki temu możliwe jest między nimi oddziaływanie takie jak w podwójnej spirali DNA. Kod, jak już wspomniano, jest zwyrodniały, a więc ten sam aminokwas może być przyłączony więcej niż do jednego tRNA, a mianowicie do tych tRNA, których antykodony są komplementarne do wszystkich kodonów danego aminokwasu. Dotychczas w komórce bakteryjnej znaleziono ponad 50 różnych tRNA.



Rys. 12. Schemat przebiegu syntezy białka w rybosomie. Rybosom przesuwa się wzdłuż łańcucha mRNA. Kolejne kodony mRNA odczytywane są przez komplementarne trójki nukleotydów (antykodony) w tRNA. W rybosomie znajdują się dwa miejsca wiążące, miejsce *A*, wiążące tRNA aminokwasu, który ma być włączony do łańcucha polipeptydowego, i miejsce *B*, do którego przyłączony jest poprzedzający tRNA połączony z syntetyzowanym łańcuchem polipeptydowym. Po utworzeniu wiązania peptydowego tRNA przesuwa się z miejsca *A* na miejsce *B*, a drugi tRNA zostaje uwolniony z miejsca *B* i z rybosomu. Aczkolwiek tRNA mogą się różnić między sobą sekwencją nukleotydów, to ich struktura przestrzenna musi być w zasadzie podobna, tak aby pasowała do struktury nadcząsteczkowej przedstawionego tu schematycznie kompleksu translacyjnego

synteza białka

Proces łączenia aminokwasów w łańcuchu białkowym tak zwanymi wiązaniami peptydowymi przebiega wewnątrz rybosomu (rys. 12). Kompleks rozpoczynający syntezę białka składa się z kilku białek o aktywności enzymatycznej, z mRNA, z rybosomów, z niskocząsteczkowego związku — dawcy energii, z tRNA połączonego z aminokwasem. Wyróżniono również białka, których obecność konieczna jest do zapoczątkowania translacji, inne uczestniczą w procesie przedłużania łańcucha polipeptydowego, w przesuwaniu rybosomu wzdłuż mRNA do następnych, jeszcze nie odczytanych kodonów, w kończeniu syntezy danego polipeptydu, w wiązaniu tRNA połączonego z aminokwasem do rybosomu. Jednym z białek rybosomalnych jest również enzym, transferaza peptydowa, syntetyzująca wiązanie peptydowe. Po zakończeniu syntezy białko zostaje odłączone od rybosomu i pozostaje w cytoplazmie.



Rys. 13. DNA przedstawiony został jako okrąg, w takim kształcie występuje w bakteriach i bakteriofagach. W wyniku transkrypcji na matrycy DNA powstają wszystkie typy RNA, o różnej strukturze przestrzennej, co schematycznie zaznaczono na rysunku. Transkrybowane RNA różnią się również funkcjonalnie: rRNA łączy się z białkami rybosomalnymi i tworzy aktywny biologicznie rybosom (schemat w ramce w dole rysunku), tRNA przyłącza aminokwas (schemat w ramce w górnej części rysunku) i w tej aktywnej formie dołącza się do rybosomu połączonego z mRNA. W wyniku translacji informacji zawartej w mRNA powstaje polipeptyd (białko), który uwalnia się od rybosomu. Po sfałdowaniu w strukturę natywną oraz po transporcie do odpowiedniego rejonu komórki białko to jest gotowe do wypełniania swej funkcji biologicznej

Ogólny schemat omówionych pokrótce procesów transkrypcji i translacji ze szczególnym uwzględnieniem roli RNA podany został na rys. 13.

Inżynieria genetyczna

Poznanie wzajemnych zależności między strukturą i funkcją kwasów nukleinowych w żywej komórce uzależnione jest od znajomości sekwencji nukleotydów w natywnej cząsteczce kwasu nukleinowego, DNA lub RNA. Nie można się spodziewać pełnego wytłumaczenia mechanizmu żadnego z omówionych

tu procesów: replikacji, mutagenezy, transkrypcji, translacji, jeżeli nieznana będzie budowa uczestniczących w tych procesach biopolimerów, a także zmiany zachodzące w strukturze w zależności od funkcjonalnego stanu badanego biopolimeru. Podczas gdy w latach 1965–1975 rozwinęły się znacznie metody badania sekwencji różnych RNA to w dziedzinie oznaczania sekwencji DNA postęp był niezwykle powolny. Sytuacja ta uległa radykalnej zmianie z chwilą odkrycia i scharakteryzowania (od 1972 r.) nowej grupy enzymów działających na DNA, nazwanych restryktazami. Enzymy te, izolowane z bakterii, nacinają łańcuch DNA wtedy, gdy zawiera on określoną sekwencję kilku nukleotydów, w taki sposób, że pozostawiają po obu stronach nacięcia krótki pojedynczołańcuchowy fragment DNA. I tak np. jedna z najlepiej poznanych restryktaz, Eco RI, nacina DNA w następujący sposób:

restryktazy

..... G ↓ AATTC ..... $\xrightarrow{\text{Eco RI}}$ ..... G   AATTC ....
..... GTTAA ↑ G .... CTTAA + G .....

Powstałe jednołańcuchowe końce, zwane lepkimi, mogą w odpowiednio dobranych warunkach odnaleźć się i dopasować, zgodnie z zasadą komplementarności nukleotydów. Jeżeli dodać wówczas enzym, który może zszywać przerwy w ciągłości łańcucha DNA, ligazę, to nastąpi połączenie dwu fragmentów DNA uprzednio przez restryktazę naciętych. Według tej strategii można połączyć ze sobą dwa łańcuchy DNA wydzielone z dowolnych organizmów, pod warunkiem, że były nacinane przez ten sam enzym restrykcyjny. To odkrycie stało się podstawą rozwoju nowego działu biologii molekularnej niekiedy zwanego inżynierią genetyczną.

ligaza

DNA — wektor

DNA, który chcemy badać, dołączamy w wyżej opisany sposób do innego DNA, zwanego DNA-wektorem. DNA-wektorem mogą być plazmidy lub DNA wirusów, które mogą wejść do żywej komórki i tam replikować przez dowolną liczbę pokoleń. W ten sposób można uzyskać dowolną ilość interesującego nas DNA połączonego ligazą z DNA-wektorem. Podstawowy schemat takich zabiegów przedstawiał się w 1977 r. jak następuje:

— Oczyszczano mRNA transkrybowane z określonego genu o znanej funkcji i o znanym produkcie białkowym.

— Specyficzną polimerazą DNA izolowaną z wirusów onkogennych syntetyzowano na matrycy tego mRNA dwułańcuchowy DNA komplementarny do mRNA, który nazwać można syntetycznym genem.

— Syntetyczny gen dołączono do DNA-wektora i wprowadzono do komórki bakteryjnej, która przez wiele generacji replikowała wektor razem z syntetycznym genem.

— Izolowano z powrotem DNA-wektor, wycinano z niego restryktazą syntetyczny gen i sprawdzano czy nie uległ on zmianie w czasie wielokrotnych replikacji w komórce gospodarza.

Jednocześnie i współzależnie od odkrycia i stosowania restryktaz opracowano dwie nowe i szybkie metody sekwencjonowania DNA. Jeżeli w 1974 r. oznaczenie sekwencji 20 nukleotydów DNA zajmowało dwa lata, to w cztery lata później wykonywano je w jeden dzień. Osiągnięty postęp jest tak wielki, że stało się łatwiejszym ustalenie sekwencji DNA niż sekwencji RNA, a nawet z sekwencji DNA przepowiada się już sekwencje aminokwasów w trudniejszych do oczyszczenia białkach, produktach danego genu (DNA).

Metody inżynierii genetycznej, aczkolwiek w dużej mierze oparte na procedurach *in vitro*, stanowią o tak wielkim postępie w rozumieniu struktury genu i jego transkryptu, mRNA, a także, co jeszcze ważniejsze, o sposobach regulacji ich aktywności, że nie mogą być pominięte przy omawianiu struktury i funkcji kwasów nukleinowych.

W krótkim artykule trudno wyłożyć wszystkie aspekty problemu wzajemnych zależności między strukturą i funkcją kwasów nukleinowych. Aby przedstawić w dostatecznie zwartej postaci wszystkie najważniejsze zagadnienia z punktu widzenia biologicznego, trzeba było zrezygnować z opisu wielu metod fizycznych, które niezmiernie wzbogaciły naszą wiedzę o kwasach nukleinowych (→ Przedmiot i problemy biofizyki molekularnej).

dlaczego istotne jest badanie kwasów nukleinowych

W przedstawionym schemacie funkcjonowania kwasów nukleinowych w komórce starano się głównie podkreślić pojawianie się w opisywanych procesach przejściowych oddziaływań kwasów nukleinowych między sobą i z białkami oraz to, że dopiero zespół biopolimerów może odgrywać określoną rolę biologiczną. Co więcej, cechą specyficzną organizmów żywych jest fakt tworzenia przez biopolimery funkcjonalnych struktur wyższych rzędów, takich jak chromatyna, kompleksy polimeraz z matrycami i substratami, rybosomy, informosomy i polisomy. Struktury te mogą powstawać w określonych warunkach doświadczalnych spontanicznie, bez udziału enzymów, a więc informacja o nich zawarta jest *a priori* w strukturze przestrzennej składników. Dlatego też badania struktur podstawowych i ich wzajemnych oddziaływań mają zasadnicze znaczenie w pojmowaniu procesu życia. W tej dziedzinie wiedzy należy się też spodziewać wciąż nowych, być może zaskakujących odkryć.

J. Ciernochowska i in. *Molekularne podstawy życia*, Warszawa 1976; J. N. Davidson *Biochemia kwasów nukleinowych*, Warszawa 1977; *Fizyczne metody badań białek i kwasów nukleinowych*, red. Ju. S. Łazurkin, Warszawa 1972; J. D. Watson *Biologia molekularna genu*, Warszawa 1975; M. W. Wolkensztejn *Na skrzyżowaniu dróg wiedzy*, Warszawa 1975.

Molekularne podstawy skurczu mięśnia

Hanna Strzelecka-Gołaszewska

Współczesne poglądy na temat molekularnego mechanizmu skurczu mięśni są rezultatem wieloletnich badań fizjologów, morfologów, biochemików i biofizyków. Najczęstszym obiektem dotychczasowych badań były mięśnie szkieletowe kręgowców. Mięśnie te, należące do grupy mięśni poprzecznie prążkowanych, mają strukturę bardziej regularną niż inne typy mięśni, toteż stanowią odpowiedniejszy materiał do badań mikroskopowych i rentgenograficznych, a zwłaszcza do śledzenia z zastosowaniem tych właśnie metod zmian strukturalnych zachodzących podczas skurczu. W badaniach biochemicznych, wymagających dużych ilości materiału, najczęściej używane są mięśnie królika, natomiast mięśnie sartorius i semitendinosus żaby okazały się — ze względu na ich wielkość i kształt — najodpowiedniejsze do doświadczeń nad mechaniką i energetyką skurczu. Niektóre zagadnienia badane są również porównawczo na różnych typach mięśni różnych zwierząt. Badania te wykazały, że mimo specyficznych różnic w sposobie działania, budowie makro- i mikroskopowej oraz w molekularnej strukturze i właściwościach białek

kurczliwych różnych typów mięśni podstawowe mechanizmy skurczu są wspólne. Co więcej, ostatnie badania nad organizmami jednokomórkowymi oraz wielokomórkowymi, lecz nie posiadającymi wyspecjalizowanej tkanki mięśniowej, wskazują na to, że podstawą wszelkiego rodzaju ruchliwości są te same lub bardzo zbliżone mechanizmy molekularne.

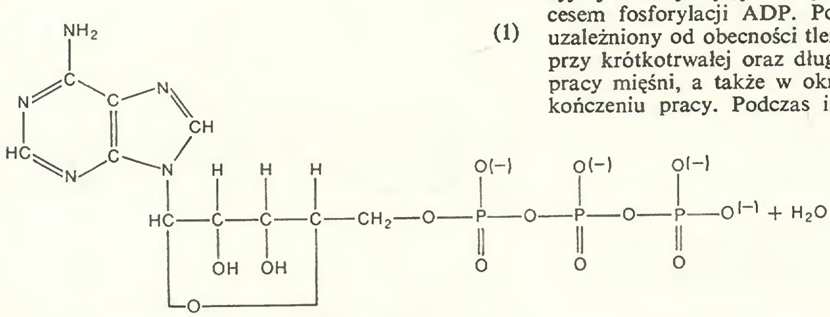
Źródła energii skurczu

Mięsień jest czymś w rodzaju biologicznego silnika przekształcającego energię chemiczną, zmagazynowaną w prostych substancjach chemicznych, w energię mechaniczną. Wiadomości o przemianach energetycznych związanych ze skurczem pochodzą z badań na mięśniach wyizolowanych z organizmu i drażnionych sztucznie impulsami elektrycznymi, imitującymi impulsy nerwowe w organizmie zwierzęcym.

Badając zmiany chemiczne zachodzące w kurczącym się mięśniu, ustalono ponad wszelką wątpliwość, że bezpośrednim źródłem energii mechanicznej i ciepła wydzielanego podczas skurczu jest reakcja hydrolizy kwasu adenozynotrójfosforowego, w skrócie ATP, w wyniku której powstają kwas adenozynodwufosforowy (ADP) i kwas ortofosforowy:

hydroliza ATP

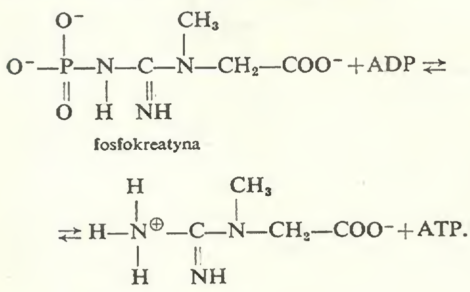
(1)



kwas adenozynotrójfosforowy (ATP) $\longrightarrow ADP^{2-} + HPO_4^{2-} +$ energia

Zawartość ATP w mięśniu jest stosunkowo niska i wystarcza najwyżej na 10 kolejnych skurczów, gdy tymczasem wyizolowany z organizmu mięsień może w warunkach tlenowych wykonać ponad 1000 skurczów. W każdym mięśniu zachodzi bowiem wiele procesów enzymatycznych, które odtwarzają rozłożony ATP. Najważniejszą z reakcji utrzymujących stały poziom ATP w pracującym mięśniu jest przeniesienie gupy fosforanowej z innego wysokoenergetycznego związku fosforowego, a mianowicie fosfokreatyny, na ADP (przeniesienie jest katalizowane przez enzym kreatynofosfotransferazę):

reakcje odtwarzające ATP



(2)

Reakcja ta jest odwracalna — nie towarzyszą jej zmiany swobodnej energii układu, lecz kierunek jej przebiegu uzależniony jest od stężeń reagujących ze sobą związków w taki sposób, że ADP, powstający w mięśniu w wyniku hydrolizy ATP, jest natychmiast przekształcany w ATP (fosforylacja ADP) i w rezultacie nie obserwuje się zmian zawartości ATP aż do momentu wyczerpania zapasu fosfokreatyny.

Inny enzym (miokinaza) katalizuje przeniesienie grupy fosforanowej z jednej cząsteczki ADP na drugą, z wytworzeniem jednej cząsteczki ATP i jednej cząsteczki kwasu adenozynomonofosforowego (AMP):

$$2\,ADP \rightleftarrows ATP + AMP. \qquad (3)$$

Ta reakcja przebiega w kierunku syntezy ATP dopiero wówczas, gdy stężenie ADP przewyższa stężenie ATP, toteż w warunkach normalnej pracy mięśnia ma ona mniejsze znaczenie, odgrywa natomiast ważną rolę przy intensywnej pracy, prowadzącej do wyczerpania zapasu fosfokreatyny.

Łączny zapas energii zawartej w ATP i fosfokreatynie w mięśniach żaby wystarcza na wykonanie 80–100 skurczów. Wykonywanie przez mięsień jeszcze większej pracy — nawet wówczas, gdy jest on wyizolowany z organizmu, a więc odcięty od dopływu substancji pokarmowych — możliwe jest dzięki procesom metabolicznym odtwarzającym ATP z ADP i nieorganicznego fosforanu. Źródłem energii, która w toku tych procesów magazynowana jest w wiązaniu końcowej grupy fosforanowej w cząsteczce ATP, jest utlenianie węglowodanów i tłuszczów. Wspólny dla przemiany węglowodanów i tłuszczów końcowy łańcuch reakcji składających się na proces tzw. oksydacyjnej fosforylacji jest najbardziej wydajnym procesem fosforylacji ADP. Ponieważ jednak jest on uzależniony od obecności tlenu, przeto odgrywa rolę przy krótkotrwałej oraz długotrwałej, lecz powolnej pracy mięśni, a także w okresie „odnowy", po zakończeniu pracy. Podczas intensywnej pracy szybkość dostarczania tlenu przez krew jest niewystarczająca dla zapewnienia sprawnego przebiegu reakcji tlenowych.

Odtwarzanie ATP z ADP i nieorganicznego fosforanu w warunkach beztlenowych zapewnia proces glikogenolizy, obejmujący wstępne etapy przemiany glikogenu. W obecności tlenu produkt glikogenolizy – kwas pirogronowy — wchodzi w cykl reakcji prowadzących do jego całkowitego spalenia; w warunkach niedotlenienia, np. przy dużym wysiłku fizycznym, glikogen ulega przemianie beztlenowej, której końcowym produktem jest kwas mlekowy, gromadzony w mięśniu podczas pracy, a następnie usuwany przez krwiobieg i utleniany w innych narządach.

Część ATP powstającego w procesach oksydacyjnej fosforylacji i glikogenolizy służy z kolei do odtworzenia zapasu fosfokreatyny (reakcja 2).

Wszystkie wymienione procesy i pojedyncze reakcje chemiczne umożliwiają przenoszenie energii chemicznej zawartej w różnorodnych substancjach pokarmowych na ATP — jedyną substancję zdolną do przekazywania energii chemicznej elementom kurczliwym, które przekształcają ją na pracę mechaniczną.

najprostszy model skurczu mięśnia

Badania nad mechanizmem przekształcania przez mięsień energii chemicznej w mechaniczną zostały zapoczątkowane odkryciem przez W. A. Engelhardta i M. N. Lubimową w 1939 r., że jedno z głównych strukturalnych białek mięśniowych — miozyna — jest enzymem katalizującym hydrolizę ATP, oraz o dwa lata późniejszą obserwacją A. Szent-Györgyiego, że miozyna połączona z drugim białkiem mięśniowym — aktyną — po wstrzyknięciu przez rurkę kapilarną do wody wytrąca się w postaci nici, które po

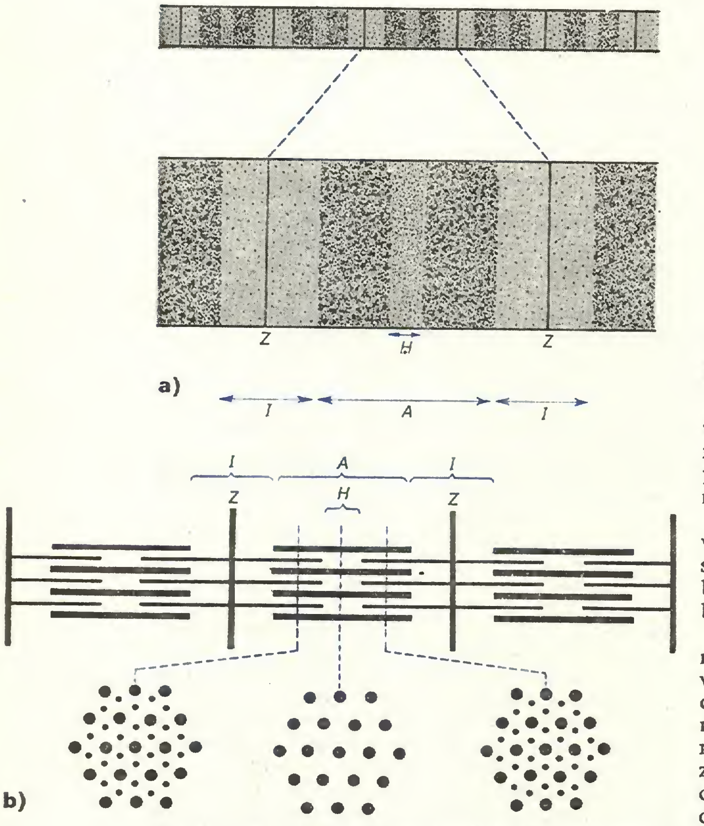
dodaniu ATP ulegają skurczeniu. W ten sposób uzyskano najprostszy modelowy układ złożony z miozyny, aktyny i ATP, umożliwiający badanie procesów, które zachodzą w żywym mięśniu podczas skurczu. Uzyskane tą drogą informacje, łącznie z wynikami jednocześnie prowadzonych badań nad molekularną organizacją miozyny i aktyny w mięśniu, ukształtowały współczesne poglądy na molekularny mechanizm skurczu.

Budowa komórek mięśniowych

Wszystkie typy mięśni zbudowane są z jedno- lub wielojądrowych komórek w kształcie włókien o średnicy ok. 100 µm. Włókna mięśni szkieletowych i mięśnia sercowego kręgowców wypełnione są przebiegającymi przez całą długość komórki, ułożonymi równolegle włókienkami, noszącymi nazwę miofibryli. Miofibryle wykazują regularne poprzeczne prążkowanie, widoczne na podłużnych przekrojach mięśnia w mikroskopie optycznym. Na podstawie obserwacji w mikroskopie polaryzacyjnym ustalono, że prążkowanie to jest skutkiem występowania wzdłuż miofibryli na przemian stref mających właściwość podwójnego załamywania światła, które w związku z tym otrzymały nazwę prążków anizotropowych (w skrócie prążki *A*), oraz stref nie mających tej właściwości, czyli prążków izotropowych (prążki *I*) (rys. 1a). Ponadto wyróżniono cienki, ciemniejszy prążek przechodzący w poprzek każdego prążka *I* i dzielący go na połowę, który nazwano linią *Z*, oraz jaśniejszą strefę pośrodku prążka *A*, zwaną strefą *H*. Powtarzające się wzdłuż miofibryli odcinki strukturalne ograniczone dwiema sąsiednimi liniami *Z* nazwano sarkomerami.

miofibryle

sarkomery

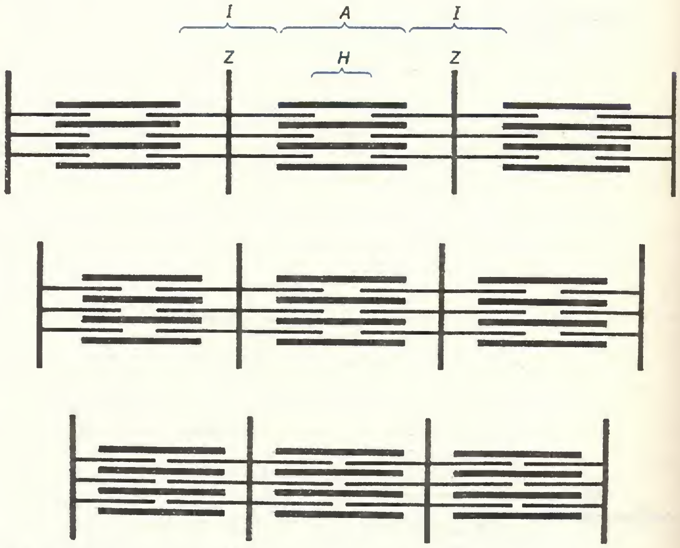


Rys. 1. Przekrój podłużny mięśnia poprzecznie prążkowanego: a) schemat obrazu miofibryli otrzymywanego w mikroskopie optycznym, b) schemat budowy pojedynczego włókna. U dołu ukazane są schematy przekrojów poprzecznych

Obserwacje w mikroskopie elektronowym wykazały, że poprzeczne prążkowanie miofibryli spowodowane jest obecnością i charakterystycznym ułożeniem w każdym sarkomerze dwojakiego rodzaju włókienkowatych elementów strukturalnych, które ze względu na ich różną średnicę nazwano filamentami grubymi i cienkimi (il. 120, tabl. 30). Filamenty grube, o średnicy 10–12 nm, przebiegają wzdłuż prążka *A*, natomiast filamenty cienkie, o średnicy 5–7 nm, biegną od linii *Z* przez całą długość prążka *I*, wnikają do prążka *A* i dochodzą w nim do granicy strefy *H*. Tak więc centralna strefa *H* zbudowana jest wyłącznie z filamentów grubych, boczne zaś strefy prążka *A* — z obu typów filamentów. Na przekrojach poprzecznych mięśnia widać regularne, heksagonalne ułożenie filamentów cienkich wokół każdego filamentu grubego, a filamentów grubych względem siebie (rys. 1b).

filamenty

W połowie lat 50-ych H. E. Huxley i J. Hanson oraz A.F. Huxley i R. Niedergerke jednocześnie ogłosili wyniki swoich obserwacji, które dały jednoznaczną odpowiedź na pytanie, jakie zmiany morfologiczne zachodzą w mięśniu podczas skurczu, powodując jego skracanie się. Stwierdzili oni, że następuje wówczas skrócenie prążków *I* oraz strefy *H*, całkowita zaś długość prążków *A* pozostaje niezmieniona. A zatem zmiany długości mięśnia są wynikiem przesuwania się grubych i cienkich filamentów wzdłuż osi miofibryli



Rys. 2. Schemat zmian układu filamentów we włóknie mięśniowym podczas skurczu

w przeciwnych kierunkach w każdej połówce sarkomeru oraz zwiększenia strefy zachodzenia filamentów pomiędzy siebie bez zmiany długości samych filamentów (rys. 2).

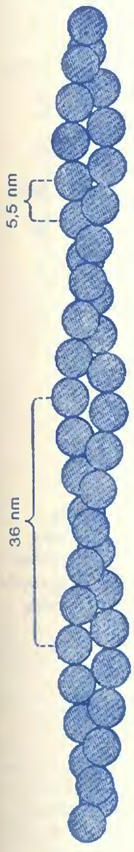
Zrozumienie mechanizmu powodującego przesuwanie się filamentów grubych i cienkich względem siebie wymaga dokładniejszej znajomości nie tylko budowy tych filamentów, lecz także właściwości białek, z których są one utworzone.

Po mechanicznym zniszczeniu struktury komórek mięśniowych można w stosunkowo prosty sposób wyizolować z nich miofibryle, z których działaniem odpowiednich roztworów soli można następnie kolejno wydzielić dwa główne strukturalne białka mięśniowe: miozynę i aktynę. Obserwacje mikroskopowe zachodzących przy tym zmian w strukturze miofibryli doprowadziły do wniosku, że filamenty grube zbudowane są z miozyny, filamenty cienkie — z aktyny. Materiałem, z którego utworzona jest linia *Z*, łącząca filamenty cienkie sąsiednich sarkomerów (rys. 1), jest inne białko strukturalne, zwane α-aktyniną.

Charakterystyka strukturalnych białek mięśniowych i ich molekularnej organizacji w mięśniu

aktyna

Aktyna wyizolowana z mięśnia może występować w dwóch postaciach: monomerycznej i spolimeryzowanej, odwracalnie przechodzących jedna w drugą. Cząsteczka aktyny monomerycznej ma kształt sferyczny. Zbudowana jest z pojedynczego łańcucha polipeptydowego o masie 42 300 daltonów (dalton — stosowana w biochemii nazwa jednostki masy atomowej u; 1 u jest równa $^{1}/_{12}$ masy jądra izotopu węgla ^{12}C). W tej postaci aktyna występuje w roztworach wodnych nie zawierających soli nieorganicznych. Dodanie soli w stężeniach fizjologicznych powoduje polimeryzację aktyny. Badając w mikroskopie elektronowym otrzymane w ten sposób polimery, stwierdzono, że mają one postać filamentów, takich samych jak cienkie filamenty bezpośrednio wyizolowane z mięśni po mechanicznym zniszczeniu struktury komórkowej i miofibrylarnej (il. 121, tabl. 30). Każdy filament zbudowany jest z dwóch śrubowo wokół siebie skręconych łańcuchów sferycznych jednostek o średnicy 5,5 nm, odpowiadającej średnicy monomeru obliczonej na podstawie fizykochemicznych badań roztworów aktyny monomerycznej. Punkty przecinania się obu łańcuchów monomerów w rzucie filamentu na płaszczyznę (rys. 3), powtarzają się w regularnych odstępach 36 nm. Ponieważ na rentgenogramach żywych mięśni znaleziono refleksy odpowiadające tej okresowości, można sądzić, że struktura filamentów aktynowych „widziana" w mikroskopie elektronowym odpowiada ich strukturze w żywym mięśniu.



Rys. 3. Schemat budowy polimeru aktyny

Podobnie jak aktyna, wyizolowana z mięśni miozyna może występować bądź w postaci monomerycznej, bądź w postaci agregatów odpowiadających kształtem filamentom grubym. Cząsteczki miozyny, o masie ok. 480 000 daltonów, mają kształt pałeczek zakończonych dwiema główkami (widoczne na rys. 4). Zbudowane są z dwóch łańcuchów polipeptydowych o masie 200 000 daltonów, zwanych łańcuchami ciężkimi, oraz czterech łańcuchów tzw. lekkich, o masach od kilkunastu do dwudziestu kilku tysięcy daltonów. Każdy z dwóch łańcuchów ciężkich ma w większej swej części strukturę heliksu α; na tym odcinku oba łańcuchy skręcone są jeszcze dodatkowo wokół siebie w superheliks, który tworzy pałeczkowatą część cząsteczki. W dalszej swej części oba ciężkie łańcuchy tracą wysoki stopień strukturalnego uporządkowania i przechodzą w jedną z dwóch główek, w których się znajdują również łańcuchy lekkie, luźno połączone z łańcuchami ciężkimi. Wiadomo obecnie, że dwa z czterech lekkich łańcuchów są niezbędne dla zachowania aktywności enzymatycznej miozyny, natomiast funkcja dwóch pozostałych, które można usunąć nie powodując utraty aktywności, jest dopiero przedmiotem badań. Charakterystyczną cechą miozyny z różnych typów mięśni oraz z mięśni różnych zwierząt jest znaczne zróżnicowanie masy, struktury i własności jej lekkich łańcuchów, niewątpliwie pozostające w związku z pewnymi różnicami własności całej cząsteczki.

miozyna

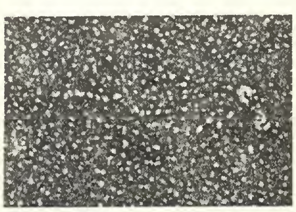
Ważnym elementem molekularnej budowy miozyny jest obecność w cząsteczce dwóch miejsc szczególnie wrażliwych na działanie enzymów trawiących białka. Dzięki temu przez trawienie trypsyną można podzielić cząsteczkę — jak zaznaczono na rys. 4 — na dwa fragmenty, z których jeden nazwano meromiozyną lekką (w skrócie *LMM*), drugi — ciężką (*HMM*). Przy dłuższym trawieniu następuje rozszczepienie *HMM* na dwa subfragmenty: *S1* i *S2*, z których pierwszy odpowiada główce, drugi zaś stanowi odcinek pałeczkowatej części cząsteczki. Po rozdzieleniu produktów trawienia okazało się, że *S1* zachowuje zdolność hydrolizy ATP, natomiast pozostałe fragmenty cząsteczki nie biorą udziału w aktywności enzymatycznej miozyny. Stwierdzono ponadto, że tylko subfragment *S1* posiada inną ważną właściwość miozyny — zdolność tworzenia kompleksu z aktyną.

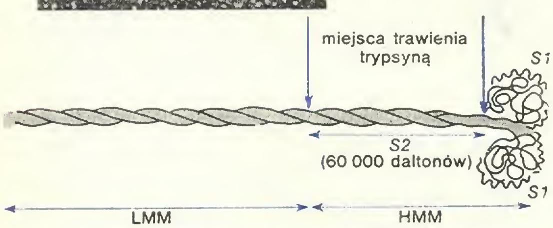
W roztworach soli o niskim, fizjologicznym stężeniu cząsteczki miozyny agregują, tworząc filamenty podobne do filamentów grubych bezpośrednio wyizolowanych z mięśni (rys. 5). W mikroskopie elektronowym na powierzchni filamentów widoczne są liczne „wypustki", których pozbawiony jest jedynie środkowy odcinek filamentu. W podobnych warunkach meromiozyna lekka również agreguje tworząc filamenty, lecz o powierzchni gładkiej, natomiast ani *HMM*, ani *S1* lub *S2* nie wykazują tendencji do agregacji. Z tych obserwacji H. A. Huxley wysnuł wniosek, że rdzeń filamentu grubego utworzony jest z ułożonych równolegle do osi filamentu odcinków cząsteczek miozyny odpowiadających meromiozynie lekkiej, natomiast części odpowiadające meromiozynie ciężkiej wystają na zewnątrz w formie widocznych w mikroskopie elektronowym luźnych wypustek (il. 123, tabl. 31). W centralnej, pozbawionej wypustek części filamentu zmienia się kierunek ułożenia cząsteczek, co nadaje filamentowi strukturalną polarność, której znaczenie zostanie omówione w dalszej części artykułu.



Rys. 5. Schemat budowy filamentu miozynowego

a)





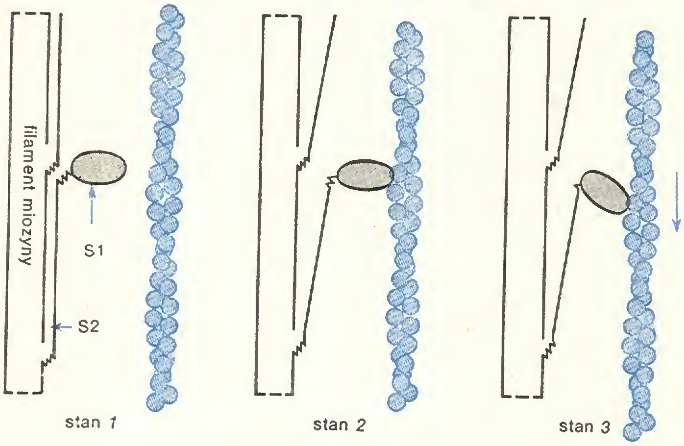
b) **Rys. 4. Budowa cząsteczki miozyny: a) obraz w mikroskopie elektronowym (preparat kontrastowany metodą cieniowania; powiększenie 175 000 razy (wg S. Lowey i in.), b) schemat. W nawiasach podano masy cząsteczkowe produktów trawienia cząsteczki trypsyną**

Molekularny mechanizm skurczu

Hydroliza ATP katalizowana przez miozynę w obecności jonów potasu, magnezu i wapnia w takich stężeniach, w jakich jony te występują w mięśniu, przebiega z małą szybkością, nie wystarczającą na pokrycie potrzeb energetycznych skurczu. W obecności aktyny szybkość ta znacznie wzrasta i osiąga wartość porównywalną z szybkością rozkładu ATP (mierzoną ubytkiem fosfokreatyny zużywanej na jego resyntezę) w kurczącym się mięśniu. Jak wynika z rozważań przedstawionych w poprzednim rozdziale, zarówno miejsce hydrolizy ATP, jak i łączenia się miozyny z aktyną znajduje się w tej części cząsteczki miozynowej, która wystaje z trzonu filamentu grubego w postaci wypustki. Według ogólnie dziś akceptowanej teorii H. E. Huxleya, dochodzący do mięśnia impuls nerwowy powoduje, że wypustki filamentów miozynowych (zwane częściej poprzecznymi mostka-

teoria Huxleya

mi) łączą się z filamentami aktynowymi. Połączenie z aktyną powoduje aktywację hydrolizy ATP przez miozynę i wyzwolenie energii wiązania końcowego fosforanu z ATP. Energia ta zostaje zużyta na zmianę kąta, pod jakim główka cząsteczki miozynowej (subfragment *S1*), stanowiąca część mostka, połączona jest z filamentem aktynowym. Zmiana kąta powoduje przesunięcie filamentu aktynowego wzdłuż miozynowego o pewien odcinek (rys. 6), nie większy niż 5–10 nm, po czym następuje rozerwanie międzyfilamentowego połączenia i ponowne jego utworzenie w dalszej części filamentu aktynowego. Opisany cykl zmian powtarza się wielokrotnie podczas skurczu, a każdy związany jest z hydrolizą jednej cząsteczki ATP.



Rys. 6. Schemat ilustrujący wahadłowy ruch mostków miozynowych oraz mechanizm przesuwania filamentów aktynowych podczas skurczu

Przedstawione wyżej zasadnicze elementy teorii Huxleya zostały sformułowane w końcu lat 50-ych. Wypada obecnie przedstawić nieco dokładniej założenia oraz dowody, na których teoria została oparta, i uzupełnić je wynikami późniejszych badań.

wahadłowy ruch mostków miozynowych

Dla wytłumaczenia wahadłowego ruchu mostków miozynowych w kierunku filamentów cienkich oraz zmiany kąta przyłączenia główek miozynowych do filamentu cienkiego Huxley oparł się na założeniu, iż w tej części łańcucha polipeptydowego, która tworzy mostek, istnieją dwa miejsca charakteryzujące się pewną giętkością i spełniające funkcję „zawiasów": jedno tam, gdzie mostek przechodzi w trzon filamentu, drugie — pomiędzy główką (subfragmentem *S1*) a liniową częścią mostka (subfragmentem *S2*), mającą charakter sztywnej pałeczki.

Cykliczne zmiany ustawienia mostków miozynowych podczas skurczu mięśnia muszą być rezultatem zmian w konformacji łańcucha polipeptydowego w obrębie wspomnianych wyżej giętkich obszarów mostka. Z zastosowaniem metod biofizycznych (np. badania zmian widma miozyny w ultrafiolecie, badania fluorescencji, rezonansu elektronowego itp.) uzyskano wiele dowodów na to, że w trakcie hydrolizy ATP rzeczywiście zachodzą zmiany konformacyjne w cząsteczce miozyny w pobliżu jej aktywnego centrum enzymatycznego.

Na schemacie ilustrującym zmiany ustawienia mostków miozynowych podczas skurczu (rys. 6) przedstawiono mostki w sposób uproszczony, z pojedynczą główką — subfragmentem *S1*, jakkolwiek wiadomo, że na jedną cząsteczkę miozyny, a co za tym idzie — na jeden mostek przypadają dwa *S1*. Wiele obserwacji wskazuje na to, że oba subfragmenty *S1* tego samego mostka nie działają jednocześnie.

Ruch filamentów aktynowych w każdym sarkomerze kurczącego się mięśnia odbywa się w sposób skoordynowany — w kierunku centralnej części sarkomeru, a więc, jak już o tym mówiono (rys. 2), w każdej połówce sarkomeru w kierunku przeciwnym. Taką koordynację zapewnia wspomniana w poprzednim rozdziale strukturalna polarność filamentów miozynowych, wynikająca z odwrotnego kierunku ułożenia cząsteczek miozyny w obu połówkach filamentu.

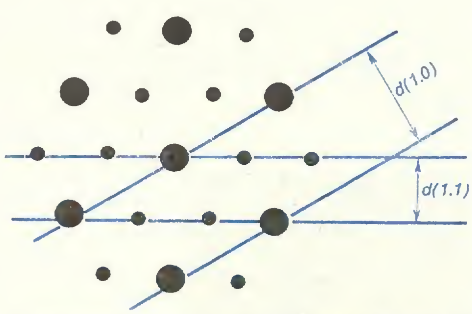
polarność filamentów miozynowych i aktynowych

Ponadto badania H. E. Huxleya i jego współpracowników dostarczyły również dowodów polarności struktury filamentów aktynowych. Przemawia za nią struktura kompleksów utworzonych przez przyłączenie cząsteczek meromiozyny ciężkiej lub subfragmentu *S1* do filamentów aktynowych (il. 122, tabl. 30). Kompleksy te wyglądem swoim przypominają długi szereg skierowanych w tę samą stronę grotów strzał długości 36 nm. Ponieważ długość ta odpowiada skokowi podwójnego heliksu, jaką tworzą łańcuchy monomerów w filamencie aktynowym, uznano, że każdy monomer aktyny wiąże jedną cząsteczkę subfragmentu *S1*. W ten sposób cały kompleks zachowuje okresowość śrubowej struktury filamentu aktynowego.

Aby powstała charakterystyczna struktura grotów strzał, przyłączone cząsteczki muszą tworzyć kąt ostry z osią filamentu i muszą być zwrócone w tym samym kierunku; w ukierunkowaniu tym ujawnia się właśnie strukturalna polarność filamentów aktynowych.

W odróżnieniu od sztucznie wytworzonych kompleksów — w mięśniu nie wszystkie monomery aktyny łączą się jednocześnie z mostkami miozynowymi. Maksymalna liczba możliwych połączeń wynika przede wszystkim z rozmieszczenia mostków na filamencie miozynowym. Do uzyskania informacji na ten temat wykorzystano obrazy dyfrakcji promieni rentgenowskich na żywym mięśniu. Kierując wiązkę promieni rentgenowskich prostopadle do dłuższej osi włókna mięśniowego, na rentgenogramie otrzymuje się w kierunku równoległym do osi włókna układ refleksów południkowych (il. 126, tabl. 31), będących wynikiem odbić od elementów powtarzających się wzdłuż osi włókna. Ich źródłem może więc być zarówno śrubowa struktura filamentów aktynowych, jak i regularne, periodyczne rozmieszczenie mostków wzdłuż filamentów miozynowych. Odległość refleksów południkowych od równika jest miarą odległości między powtarzalnymi strukturami w mięśniu. Natomiast rozkład refleksów na równiku — w kierunku prostopadłym do dłuższej osi włókna — odpowiada regularnościom przestrzennego rozmieszczenia filamentów (rys. 7).

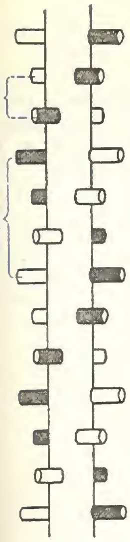
badanie struktury filamentów metodą rentgenowską



Rys. 7. Heksagonalna siatka grubych i cienkich filamentów w mięśniu poprzecznie prążkowanym kręgowców; na rysunku zaznaczono płaszczyzny będące źródłem refleksów równikowych (prostopadłych do dłuższej osi włókna mięśniowego) na obrazach dyfrakcji promieniowania rentgenowskiego na żywym mięśniu

Charakterystyczną cechą obrazów dyfrakcji promieni *X* przez mięsień w stanie spoczynkowym (il. 126, tabl. 31) jest ułożenie refleksów południkowych w linie warstwowe odpowiadające okresowości 42,9 nm oraz obecność o wyjątkowo dużym natężeniu refleksu południkowego odpowiadającego okresowości 14,3 nm. Po skonfrontowaniu tych danych z obrazami podłużnych skrawków mięśnia w mikroskopie elektronowym (na których przy dużych powiększeniach można zauważyć pary mostków odchodzące od filamentu miozynowego w odstępach ok. 40 nm)

model Huxleya i Browna



Rys. 8. Model rozmieszczenia poprzecznych mostków na filamencie grubym (wg H. E. Huxleya). Część mostków oznaczono szarym kolorem dla uwidocznienia ich śrubowego układu

H. E. Huxley i W. Brown zaproponowali model rozmieszczenia mostków na filamencie miozynowym (rys. 8). Śrubowe ułożenie mostków w taki sposób, że tworzą one jednocześnie sześć biegnących wzdłuż filamentu rzędów, w których odstępy między mostkami wynoszą 42,9 nm, najlepiej tłumaczy wyniki badań rentgenograficznych, a zarazem uzasadnione jest faktem, że każdy filament miozynowy otoczony jest sześcioma filamentami aktynowymi; takie rozmieszczenie mostków byłoby więc najkorzystniejsze jeśli chodzi o łatwość wytwarzania międzyfilamentowych połączeń.

Regularność budowy filamentów miozynowych uwidacznia się tylko na rentgenogramach mięśni w stanie spoczynkowym, kiedy to międzyfilamentowe połączenia są przerwane. Po wyczerpaniu zapasu ATP i fosfokreatyny w wyniku długotrwałej serii skurczów mięsień wyizolowany przechodzi w stan tężca, staje się sztywny i nierozciągliwy. Wskazuje to na wytworzenie trwałych połączeń między filamentami miozynowymi i aktynowymi, co jest zgodne z wynikami badań nad reagowaniem miozyny z aktyną *in vitro* (w nieobecności ATP oba białka łączą się w kompleks). Rentgenogramy mięśnia w tym stanie znacznie się różnią od obrazów jego w stanie spoczynkowym: południkowe refleksy odpowiadające okresowi 42,9 nm niemal całkowicie zanikają, a następuje wzmocnienie refleksów pochodzących z filamentów aktynowych (36 nm). Zmienia się także względne natężenie refleksów równikowych (osłabienie refleksu 1.0, a wzmocnienie 1.1: rys. 7). Ponieważ natężenie refleksów *klw* jest proporcjonalne do gęstości substancji rozpraszającej promienie *X*, zmiany te interpretuje się jako wynik przesunięcia większej części masy mostków miozynowych w kierunku filamentów aktynowych dla wytworzenia z nimi połączeń. Połączenie z filamentami aktynowymi nadaje konfiguracji mostków nową regularność, określoną sposobem ułożenia monomerów w filamencie aktynowym.

rentgenogramy mięśnia w stanie skurczu, spoczynku i tężca

Z porównania natężenia poszczególnych refleksów na rentgenogramach mięśni w stanie skurczu, tężca i spoczynku wynika, że podczas skurczu w każdym momencie dochodzi do skutku zaledwie ok. 20% możliwych połączeń międzyfilamentowych. Jest to eksperymentalny dowód słuszności jednego z głównych założeń teorii skurczu H.E. Huxleya — cykliczności wytwarzania międzyfilamentowych połączeń. Tak niski procent jednocześnie istniejących połączeń oznacza, że każdy mostek w następujących po sobie cyklach zmian jego stanu pozostaje w połączeniu z filamentem aktynowym przez stosunkowo krótki czas — zaledwie $^1/_5$ część czasu trwania całego cyklu.

Analiza rentgenogramów mięśnia dostarczyła również dowodów na to, że w mięśniu rozkurczonym, w którym międzyfilamentowe połączenia są rozerwane, globularne części mostków (główki cząsteczek miozyny) ustawione są pod kątem prostym do osi filamentów, natomiast gdy mięsień znajduje się w stanie tężca, są one połączone z filamentami aktynowymi pod kątem 45°. Są to spostrzeżenia, które nie stanowią wprawdzie bezpośredniego dowodu na występowanie tych dwóch konfiguracji mostków przejściowo w każdym cyklu tworzenia i rozrywania międzyfilamentowych połączeń, czynią jednak to istotne założenie teorii Huxleya wysoce prawdopodobnym.

W ostatnich latach wykazano, że hydroliza ATP przez miozynę jest procesem, na który się składa co najmniej 7 kolejnych reakcji:

hydroliza ATP przez miozynę

$$M + ATP \overset{1}{\rightleftarrows} M\cdot ATP \overset{2}{\rightleftarrows} M^*\cdot ATP \overset{3}{\rightleftarrows}$$

$$\overset{3}{\rightleftarrows} M^{**}\cdot ADP\cdot P \overset{4}{\rightleftarrows} M^*\cdot ADP\cdot P \overset{5}{\rightleftarrows} M^*\cdot ADP + P \overset{6}{\rightleftarrows}$$

$$\rightleftarrows M\cdot ADP + P \overset{7}{\rightleftarrows} M + ADP + P, \qquad (4)$$

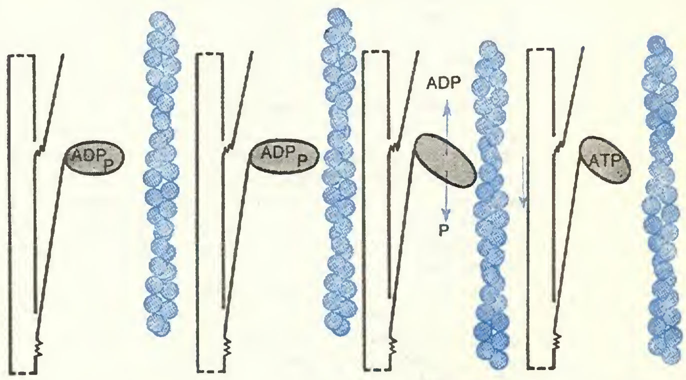
gdzie przez M, M* i M** oznaczono miozynę w różnych stanach konformacyjnych, które rozróżniono stosując techniki spektroskopowe. Jakkolwiek pod wpływem aktyny ogólna szybkość rozkładania ATP przez miozynę ulega znacznemu zwiększeniu, właściwa reakcja hydrolizy (3) przebiega w tym układzie z taką samą szybkością jak w nieobecności aktyny. Dzieje się tak dlatego, że przyłączenie ATP w centrum aktywności enzymatycznej miozyny powoduje dysocjację kompleksu miozyny z aktyną i dopiero po hydrolizie ATP, która zachodzi na cząsteczce wolnej miozyny, następuje ponowne połączenie się miozyny z aktyną, co przyspiesza uwalnianie produktów hydrolizy z centrum aktywności enzymatycznej miozyny i w ten sposób przyspiesza cały proces:

$$\begin{array}{ccc} M^*\cdot ATP & \rightarrow & M^{**}\cdot ADP\cdot P \\ \uparrow \nearrow & A & \searrow \downarrow \\ AM + ATP \rightarrow AM\cdot ATP \rightarrow AM^*\cdot ATP & & AM^{**}\cdot ADP\cdot P \rightarrow \end{array} \qquad (5)$$

$$AM^*\cdot ADP\cdot P \rightarrow AM^*ADP + P \rightarrow AM\cdot ADP + P \rightarrow AM + ADP + P.$$

Taka sekwencja reakcji implikuje rozdzielenie w czasie etapu uwalniania energii chemicznej z ATP i etapu jej przekształcania w energię mechaniczną, ponieważ hydroliza ATP (uwalnianie energii) zachodzi na wolnej miozynie, natomiast wykonanie pracy przesunięcia filamentu aktynowego wymaga istnienia międzyfilamentowego połączenia. Badania nad procesem hydrolizy ATP przez aktomiozynę *in vitro* wykazały że — zgodnie z przyjętym w teorii Huxleya założeniem — energia chemiczna końcowego wiązania fosforanowego w ATP zostaje podczas skurczu mięśnia zużyta na zmianę konformacji cząsteczki miozyny i w tej formie jest „przechowywana" do momentu przekształcenia w energię mechaniczną, co ma miejsce wówczas, gdy cząsteczka miozyny połączona z aktyną powraca do swojej konformacji wyjściowej. Nie jest jeszcze pewne, która z poznanych reakcji procesu hydrolizy ATP sprzężona jest z wykonaniem pracy mechanicznej, nie ulega jednak wątpliwości, że jest to jedna z reakcji prowadzących do uwolnienia produktów hydrolizy ATP z cząsteczki miozynowej.

zmiana konformacji — forma przechowywania energii



Rys. 9. Schemat zmian konfiguracji mostków miozynowych na różnych etapach hydrolizy ATP w układzie aktomiozynowym (wg H. G. Mannherza i in.).

Rysunek 9 ilustruje pierwszą próbę korelacji zmian konfiguracji mostków miozynowych z cyklem reakcji procesu hydrolizy ATP. W modelu tym najbardziej hipotetyczna jest konfiguracja mostków po przyłączeniu ATP (stan *4*); nie udało się dotąd sprawdzić jej doświadczalnie, ponieważ — jak już wspomniano — reakcja hydrolizy zachodzi szybciej niż uwalnianie jej produktów, można więc przypuszczać, że w żywym mięśniu, tak jak w roztworach białek, w obecności ATP w każdym momencie większość cząsteczek miozynowych występuje w połączeniu z produktami hy-

drolizy (stan 1). Stan 1 jest analogiczny do tego, w jakim się znajdują wszystkie lub większość mostków podczas rozkurczu (spoczynku) mięśnia, konfiguracja zaś mostków w stanie 3 jest typowa dla stanu tężcowego.

Regulacja cyklu skurczowo-rozkurczowego

Odpowiedź na pytanie, w jaki sposób dochodzący do mięśnia bodziec nerwowy odbierany jest przez białka kurczliwe i jakie wywołuje w nich zmiany, które prowadzą do skurczu mięśnia, nasunęła się w wyniku badania procesu odwrotnego, czyli rozkurczu. Skoncentrowanie uwagi na mechanizmie rozkurczu wynikło stąd, że wszystkie układy modelowe stosowane w badaniach nad skurczem wydawały się pozbawione tej ważnej właściwości żywego mięśnia, jaką jest zdolność do rozkurczu. Tak zw. glicerynowane włókna mięśniowe, tj. włókna pozbawione substancji niskocząsteczkowych i rozpuszczalnych białek przez wypłukanie ich wodnym roztworem glicerolu, po umieszczeniu w roztworze ATP i jonów magnezu ulegają skurczowi, który się utrzymuje nadal po całkowitym rozłożeniu ATP. Również nieodwracalna jest superprecypitacja — reakcja preparatów aktomiozyny na dodanie ATP w fizjologicznych warunkach jonowych, polegająca na wytrąceniu się żelu aktomiozynowego w postaci odwodnionego osadu; jest to najprostszy model skurczu mięśnia *in vitro*. Po rozłożeniu dodanego ATP nie obserwuje się rozpuszczenia wytrąconego białka.

rola jonów wapnia

Po bezskutecznych poszukiwaniach hipotetycznej substancji rozkurczającej udało się w latach 50-ych uzyskać rozkurcz glicerynowanych włókien mięśniowych przez potraktowanie ich substancjami silnie wiążącymi jony wapnia. Zwrócono wówczas uwagę na wcześniejszą obserwację fizjologów, że wstrzyknięcie soli wapnia do włókien mięśniowych wywołuje lokalny skurcz. Okazało się, że skurcz i rozkurcz inicjowane są za pośrednictwem tej samej substancji — jonów wapnia.

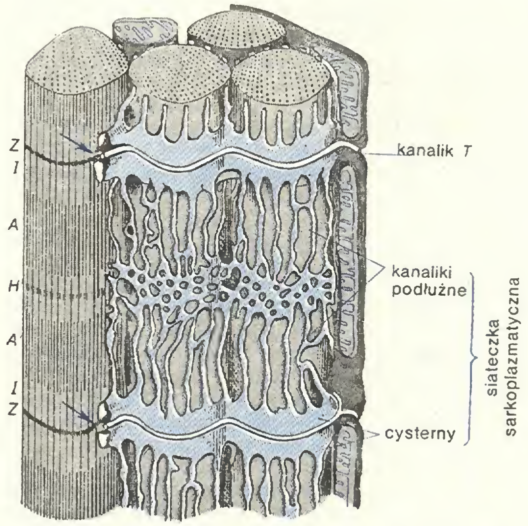
Badając zależność napięcia glicerynowanych włókien mięśniowych, stopnia superprecypitacji żelu aktomiozynowego, a także aktywacji ATPazy miozynowej przez aktynę od stężenia jonów wapnia, stwierdzono, że odpowiedź skurczowa wymaga obecności jonów wapnia w stężeniach powyżej 10^{-6} mol/l. Przy stężeniu 10^{-5} mol/l otrzymuje się maksymalną reakcję wszystkich badanych modeli mięśniowych. W tak niskich stężeniach wapń obecny jest zawsze jako zanieczyszczenie zarówno w preparatach białek, jak i w glicerynowanych włóknach — tym się tłumaczy stosunkowo późne zwrócenie uwagi na jego fizjologiczną rolę w mięśniu.

siateczka sarkoplazmatyczna

Całkowite stężenie wapnia w mięśniu jest rzędu 10^{-3} mol/l. Zaczęto więc z kolei poszukiwać fizjologicznej substancji lub struktury, która odwracalnie wiążąc i uwalniając jony wapnia mogłaby regulować ich stężenie w plazmie komórki mięśniowej. Okazało się, że funkcję taką spełnia wewnątrzkomórkowy układ strukturalny zwany siateczką sarkoplazmatyczną (rys. 10 i il. 124 z tabl. 31). Jest to zamknięty układ błon tworzących podłużne kanaliki wokół miofibryli, łączące się w pobliżu granicy między prążkami *A* i *I* (w mięśniach szkieletowych ssaków) lub w pobliżu linii *Z* (w mięśniach niższych kręgowców) w tzw. cysterny. Między sąsiadującymi ze sobą cysternami przebiegają kanaliki poprzeczne, tworzące tzw. układ *T* (skrót od angielskiej nazwy *transverse* 'poprzeczny'), który w odróżnieniu od błon siateczki sarkoplazmatycznej posiada łączność z błoną otaczającą komórkę mięśniową — sarkolemmą; kanaliki *T* są to wpuklenia sarkolemmy do wnętrza komórki mięśniowej.

badanie przemieszczania się jonów wapnia

Stosując różne metody pozwalające uwidocznić rozmieszczenie jonów wapnia na preparatach oglądanych w mikroskopie elektronowym, udało się zidentyfikować cysterny siateczki sarkoplazmatycznej jako miejsca magazynowania wapnia w stanie spoczynku mięśnia. Do zbadania przemieszczania się jonów wapnia w komórce mięśniowej w cyklu skurcz–rozkurcz posłużono się białkiem zwanym ekworyną, którego charakterystyczną cechą jest luminescencja pod wpływem wiązania wapnia. Wstrzyknięta do włókien mięśniowych ekworyna łatwo się w mioplazmie rozprzestrzenia, a jej wysokie powinowactwo do wapnia zapewnia wiązanie tego jonu przy jego niskich, fizjologicznych stężeniach. Rejestrując luminescencję nasyconych ekworyną włókien mięśniowych stwierdzono, że skurcz włókien pod wpływem elektrycznego pobudzenia poprzedzony jest wzrostem stężenia wapnia w mioplazmie. Udowodniono w ten sposób, że szybkość uwalniania wapnia z siateczki sarkoplazmatycznej jest wystarczająca na to, aby przekazywanie pobudzenia wytłumaczyć przemieszczeniem jonów wapnia do mioplazmy.

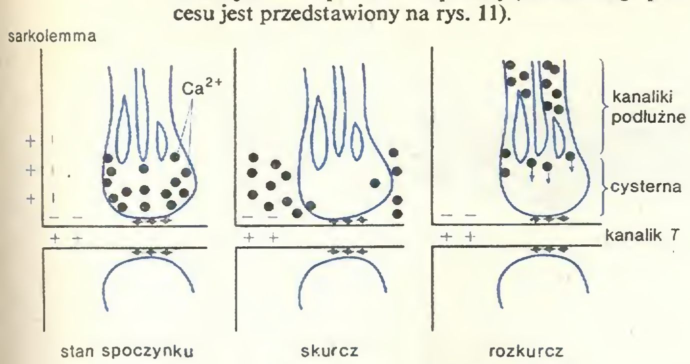


Rys. 10. Schemat budowy siateczki sarkoplazmatycznej: wg D. W. Fewcett, M. S. McNutt, J. Cell Biol. 42, 1 (1969)

rola kanalików *T*

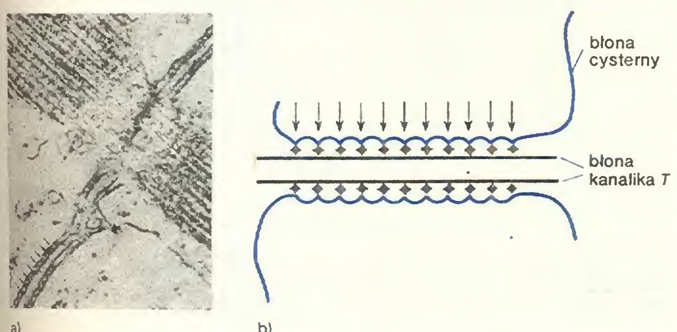
Błony siateczki sarkoplazmatycznej nie stykają się bezpośrednio z powierzchnią komórki mięśniowej w mięśniach szkieletowych kręgowców, wobec tego rolę ogniwa pośredniczącego w przekazywaniu bodźca nerwowego do wnętrza komórki przypisano kanalikom *T*, o czym zdecydowała ich łączność z sarkolemmą oraz bliskie sąsiedztwo z cysternami siateczki sarkoplazmatycznej. Z badań elektrofizjologicznych wiadomo, że dochodzący do mięśnia impuls nerwowy powoduje depolaryzację sarkolemmy, tj. obniżenie dodatniego potencjału elektrycznego występującego na jej zewnętrznej powierzchni w stanie spoczynku. Ten spoczynkowy potencjał jest wywołany różnicą stężeń niektórych jonów w komórce i w płynie pozakomórkowym, utrzymującą się wskutek selektywnej przepuszczalności sarkolemmy. Depolaryzacja związana jest ze zmianą przepuszczalności i przenikaniem jonów sodu do wnętrza komórki. Stopień depolaryzacji wzrasta ze wzrostem siły bodźca, przy czym do pewnej wartości progowej nie powoduje to zmiany stanu mięśnia; po przekroczeniu tej wartości następuje dalsza, gwałtowna zmiana przepuszczalności błony, powodująca dalszą dyfuzję jonów i wytworzenie potencjału o znaku przeciwnym, potencjału czynnościowego, któremu towarzyszy skurcz. Przyjmując założenie, że właściwości błon układu *T* są takie same jak sarkolemmy, łatwo sobie wyobrazić, że potencjał czynnościowy rozprzestrzenia się dostatecznie szybko na

błonę kanalików *T*. Sygnał ten, przekazywany cysternom siateczki sarkoplazmatycznej, powoduje uwolnienie jonów wapnia do mioplazmy (schemat tego procesu jest przedstawiony na rys. 11).



Rys. 11. Schemat ilustrujący sprzężenie procesu pobudzenia i skurczu mięśnia

Mechanizm przekazywania pobudzenia z kanalików *T* do siateczki sarkoplazmatycznej jest jeszcze przedmiotem dyskusji. Według jednej z hipotez depolaryzacja błony kanalików *T* wywołuje depolaryzację błony cystern, a ta z kolei zmienia stopień jej przepuszczalności dla wapnia. Inna hipoteza zakłada uwalnianie ze ścian kanalików *T* jakiejś nie znanej jeszcze substancji, która oddziałuje na błonę cystern. Dokładna obserwacja w mikroskopie elektronowym wykazała, że w miejscu zetknięcia z kanalikami *T* błona cystern jest pofałdowana, a pomiędzy nią a błoną kanalika *T* występują regularnie rozmieszczone zagęszczenia bezpostaciowego materiału (rys. 12). Te zagęszczenia mogłyby być miejscami przepływu



Rys. 12. Struktura zetknięcia się kanalików układu *T* z cysternami siateczki sarkoplazmatycznej: a) obraz w mikroskopie elektronowym, powiększenie 70 000 razy (wg C. Franzini-Armstrong), b) schemat, strzałkami zaznaczono zagęszczenia bezpostaciowego materiału między błonami cysterny i kanalika *T*

prądu elektrycznego albo mogłyby spełniać funkcję utrzymywania błon obu układów w stałej odległości, ewentualna zaś dyfuzja hipotetycznej substancji pobudzającej zachodziłaby w przestrzeniach pomiędzy nimi.

uwalnianie wapnia

Dokładne poznanie molekularnego mechanizmu uwalniania wapnia z siateczki sarkoplazmatycznej również wymaga dalszych badań. Obecnie już wiadomo, że absorpcja wapnia po ustaniu działania bodźca nerwowego jest procesem aktywnego transportu kosztem energii ATP, przy udziale enzymu hydrolizującego ATP, będącego głównym składnikiem białkowym błon siateczki sarkoplazmatycznej. Transport ten zachodzi do momentu obniżenia stężenia wolnych jonów wapnia na zewnątrz układu do wartości rzędu 10^{-8} mol/l. Na wewnętrznej powierzchni błon siateczki sarkoplazmatycznej rozmieszczone są białka o wysokim powinowactwie do wapnia, przypuszczalnie więc wapń — w stanie spoczynku mięśnia — zmagazynowany jest w postaci kompleksu z tymi białkami, a jego dyfuzję do mioplazmy pod wpływem impulsu nerwowego poprzedza zmiana konformacji białka, umożliwiająca uwolnienie wapnia w formie zjonizowanej.

Wkrótce po zwróceniu uwagi na rolę wapnia w procesie skurczu zauważono, że hydroliza ATP i superprecypitacja aktomiozyny wymagają obecności jonów wapnia tylko wówczas, gdy białka kurczliwe nie były poddawane dokładnemu oczyszczaniu. W latach 1963–66 S. Ebashi z uniwersytetu w Tokio wykazał, że funkcję „uczulania" aktomiozyny na zmiany stężenia jonów wapnia spełniają dwa inne białka: tropomiozyna, której obecność w mięśniu znana była już wcześniej, oraz białko odkryte przez Ebashiego i nazwane przez niego troponiną.

troponina wiąże wapń

Troponina jest jedynym białkiem miofibrylarnym, które wiąże i uwalnia wapń w zakresie jego fizjologicznych stężeń w mioplazmie. Sama troponina nie wywiera jednak żadnego wpływu na aktomiozynę, dopiero w połączeniu z tropomiozyną wykazuje zdolność uczulania wysoce oczyszczonych preparatów aktomiozyny na zmiany stężenia wapnia, tzn. powoduje hamowanie ich aktywności ATP-azowej i superprecypitacji gdy stężenie wolnych jonów wapnia nie przekracza 10^{-6} mol/l. Już te wstępne informacje wskazywały na to, że bodziec nerwowy jest — za pośrednictwem uwolnionego z siateczki sakroplazmatycznej wapnia — odbierany przez troponinę i przekazywany przez nią białkom kurczliwym przy udziale tropomiozyny.

model cienkiego filamentu

Zdolność wiązania tropomiozyny i troponiny, którą ma aktyna, a której miozyna nie posiada, była pierwszą wskazówką, że oba te białka są zlokalizowane w cienkich filamentach. Potwierdziły to badania immunochemiczne, w których się wykorzystuje specyficzną reakcję przeciwciała z jego antygenem. Oglądając w mikroskopie elektronowym podłużne skrawki mięśni potraktowane przeciwciałem na tropomiozynę, stwierdzono równomierną reakcję przeciwciała z filamentami cienkimi na całej ich długości. Natomiast obserwacje w mikroskopie elektronowym z zastosowaniem normalnej techniki kontrastowania preparatów białkowych nie wykazywały różnic w strukturze filamentów otrzymanych przez polimeryzację czystej aktyny i filamentów aktynowych zawierających tropomiozynę. Zaproponowany przez J. Hanson i J. Lowy'ego model rozmieszczenia tropomiozyny na filamencie aktynowym godzi te pozornie sprzeczne obserwacje. Cząsteczka tropomiozyny zbudowana jest z dwóch łańcuchów polipeptydowych o masie ok. 34 000 daltonów, posiadających konformację heliksu α i zwiniętych wokół siebie w superheliks (podobnie jak to ma miejsce w wypadku ciężkich łańcuchów miozyny w pałeczkowatej części cząsteczki miozynowej); długość cząsteczki wynosi 40 nm, a średnica — zaledwie 2 nm. W roztworach o niskiej zawartości soli cząsteczki tropomiozyny polimeryzują łącząc się „koniec z początkiem". Według sugestii Hanson i Lowy'ego dwa łańcuchy tak połączonych cząsteczek tropomiozyny, nawinięte na aktynowy trzon filamentu, mieszczą się w rowkach podwójnego heliksu, który tworzą łańcuchy monomerów aktyny (rys. 13), i dlatego nie są widoczne w mikroskopie.

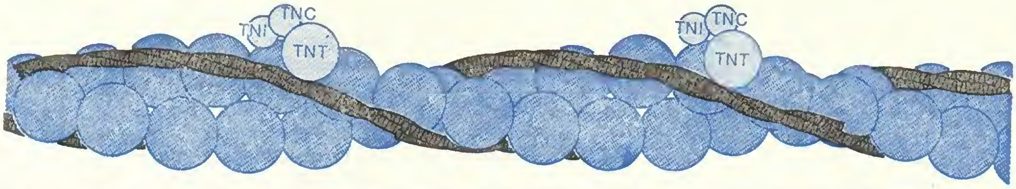
lokalizacja tropomiozyny

lokalizacja troponiny

Pierwsze informacje o rozmieszczeniu troponiny na filamencie cienkim uzyskano także przy użyciu przeciwciała na to białko. Przeciwciało przyłącza się do cienkich filamentów w regularnych odstępach, równych długości cząsteczki tropomiozyny, a znacznie przekraczających wymiary globularnej cząsteczki troponiny. Uznano więc, że regularne rozmieszczenie troponiny wzdłuż filamentu cienkiego (rys. 13) jest wynikiem wiązania się jej z tropomiozyną w jednym tylko miejscu, które według najnowszych badań znajduje się w odległości ok. 14 nm od karboksylowego końca cząsteczki tropomiozyny.

Bardziej szczegółowych informacji o strukturze cienkich filamentów dostarczyły badania włóknistych agregatów, powstających w pewnych warunkach

przez połączenie równolegle do siebie ułożonych filamentów. Agregaty filamentów czystej aktyny (il. 125a, tabl. 31) wykazują prążkowanie o okresie 36 nm, odnia z miozyną siedmiu monomerów aktynowych pozostających w kontakcie z tą samą cząsteczką tropomiozyny.



Rys. 13. Model cienkiego filamentu (C. Cohen). Symbolami TNI, TNC i TNT oznaczono trzy składniki białkowe wchodzące w skład cząsteczki troponiny

badanie agregatów filamentów aktyny

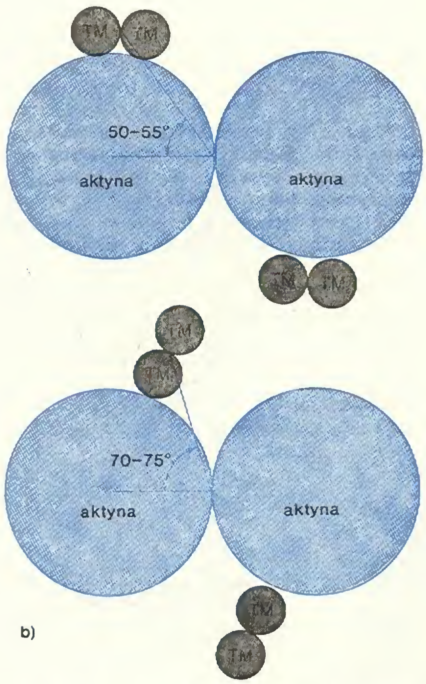
powiadające skokowi podwójnego heliksu łańcuchów monomerów w filamencie, co świadczy o połączeniu filamentów na odpowiednich poziomach ich spiralnej struktury. Na obrazach agregatów filamentów zawierających tropomiozynę i troponinę (il. 125b, tabl. 31) okres śrubowej struktury aktyny ulega zatarciu wskutek pojawienia się innego rodzaju prążkowania, o okresie 38,5 nm. W świetle wyników uzyskanych przy użyciu przeciwciał to nowe prążkowanie należy przypisać periodycznemu rozmieszczeniu troponiny. W ten sposób wyjaśniło się również nie znane dotąd pochodzenie południkowych refleksów o okresie 38,5 nm, występujących na rentgenogramach żywego mięśnia.

Dzięki regularności ułożenia filamentów w ich agregatach można badać ich strukturę metodą optycznej dyfrakcji na wykonanych w mikroskopie elektronowym zdjęciach agregatów. Analizując rozkład i natężenie refleksów na otrzymanych obrazach dyfrakcyjnych, można metodą transformacji Fouriera obliczyć funkcję rozkładu gęstości optycznej, a z niej — określić przestrzenny rozkład masy cząsteczek białkowych w filamencie i wykonać trójwymiarową rekonstrukcję filamentu. W ten sposób udało się nie tylko potwierdzić (w pewnych szczegółach nawet skorygować) przedstawiony wyżej model cienkiego filamentu, lecz również zauważyć zmiany w strukturze filamentu zachodzące pod wpływem wiązania wapnia przez troponinę. Okazało się, że w nieobecności jonów wapnia łańcuchy cząsteczek tropomiozyny przebiegają nie przez środek, lecz wzdłuż krawędzi rowków między łańcuchami monomerów aktyny (rys. 14a). Obecność jonów wapnia powoduje niewielkie przesunięcie cząsteczek tropomiozyny ku centrum rowka, z jednoczesnym odsunięciem środka ich masy od aktynowego trzonu filamentu.

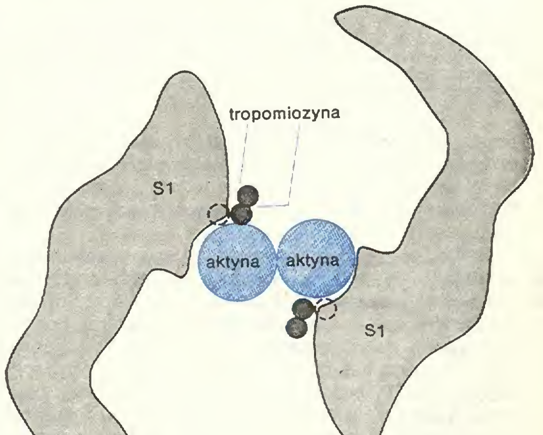
Z zaproponowanych dotychczas modeli najlepiej tę zmianę ilustruje schemat przedstawiony na rys. 14b. Przy konstruowaniu tego modelu uwzględniono dwułańcuchową budowę cząsteczki tropomiozyny i założono, że przesunięcie jest wynikiem rotacji cząsteczki wokół jej osi. Z przeprowadzonych analiz wynika, że sumaryczne przesunięcie tropomiozyny wynosi 1,5 nm. Do podobnego wniosku skłaniają wyniki analizy rentgenograficznej mięśni w stanach spoczynku i skurczu.

mechanizm regulacji skurczu

Na podstawie tych obserwacji przypuszcza się, że mechanizm regulacji cyklu skurczowo-rozkurczowego jest następujący: tropomiozyna w pozycji rozkurczowej (na krawędzi rowka między łańcuchami aktyny) blokuje przestrzennie miejsce reagowania aktyny z miozyną (rys. 15). Wapń uwolniony z siateczki sarkoplazmatycznej jest wiązany przez troponinę, co powoduje zmianę konformacji troponiny (obecnie jest już na to wiele dowodów eksperymentalnych). Wskutek tej zmiany następuje przemieszczenie tropomiozyny i odsłonięcie regionu niezbędnego dla wytworzenia połączenia aktyny z mostkami miozynowymi. W ten sposób troponina — mimo jej periodycznego rozmieszczenia na filamencie — może kontrolować stan wszystkich monomerów aktynowych: jedna cząsteczka troponiny reguluje zdolność do reagowa-



Rys. 14. Schemat przekroju poprzecznego przez cienki filament ilustrujący zmianę pozycji tropomiozyny *TM* w cyklu skurczowo-rozkurczowym: a) w nieobecności jonów wapnia, b) po związaniu wapnia przez troponinę (wg D. A. D. Parry'ego)

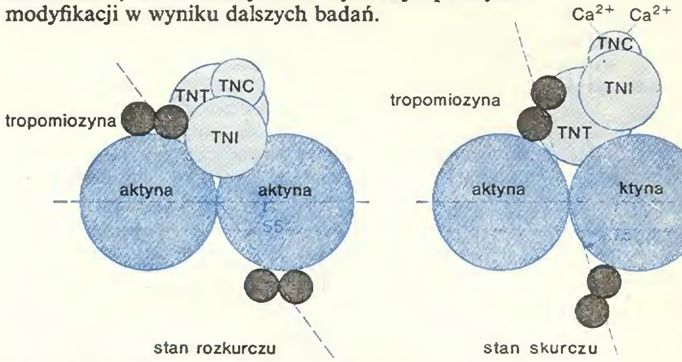


Rys. 15. Schematyczne przedstawienie wpływu pozycji tropomiozyny w filamencie cienkim na oddziaływanie aktyny z mostkami miozynowymi. Linią przerywaną zaznaczono pozycję „rozkurczową", blokującą miejsce interakcji z mostkiem miozynowym (modyfikacja schematu T. Wakabayashiego i in.)

W tej chwili nie wiadomo jeszcze, czy postulowana uprzednio konformacyjna zmiana w cząsteczkach tropomiozyny jest powodem ich przesunięcia na filamencie aktynowym. Istnieje alternatywna możliwość, że skurczowa pozycja tropomiozyny — bliżej centrum rowka — jest termodynamicznie bardziej korzystna od pozycji spoczynkowej. W tym wypadku zmiana konformacji troponiny w następstwie związania przez nią wapnia powodowałaby uwolnienie tropomiozyny z termodynamicznie mniej korzystnej pozycji, bez zmiany konformacji cząsteczki tropomiozyny. Druga z wymienionych możliwości wydaje się bardzo prawdopodobna w świetle obecnych wiadomości o mechanizmie reagowania troponiny w cienkim filamencie na zmiany stężenia jonów wapnia.

Dla uproszczenia pominięto w dotychczasowych rozważaniach fakt, że troponina zbudowana jest z trzech podjednostek o odrębnych własnościach, połączonych w stosunku molowym 1:1:1. Składnik zwany TNI, o masie 23 000 daltonów (w mięśniach szkieletowych królika; w różnych typach mięśni i u różnych zwierząt ciężary cząsteczkowe podjednostek troponiny nieco się różnią), w obecności tropomiozyny hamuje reakcję aktyny z miozyną niezależnie od stężenia jonów wapnia w środowisku. Składnik TNC, o masie 19 000 daltonów, jest tym składnikiem, w którym są zlokalizowane miejsca wiązania wapnia z troponiną, a w obecności trzeciego składnika, TNT, o masie 39 000 daltonów, powoduje zniesienie hamującego działania TNI, jeżeli stężenie wolnych jonów wapnia przekracza 10^{-6} mol/l. Wyniki szczegółowych badań nad zdolnością reagowania poszczególnych podjednostek troponiny z tropomiozyną i aktyną sugerują, że cząsteczka troponiny ma dwa miejsca przyłączenia do cienkiego filamentu (rys. 16): połączenie przez TNT jest trwałe niezależnie od stężenia jonów wapnia w środowisku, natomiast drugie połączenie, przez TNI, dochodzi do skutku tylko w nieobecności jonów wapnia. Ponieważ TNC nie wiąże się bezpośrednio ani z aktyną, ani z tropomiozyną, a jest jedynym składnikiem troponiny wiążącym wapń, zmiana stężenia jonów wapnia musi wpływać na wiązanie TNI z aktyną i (lub) tropomiozyną przez zmianę konformacji TNC. Jednoczesne istnienie obu połączeń (przez TNT i przez TNI) utrzymuje tropomiozynę w pozycji blokującej miejsce reagowania aktyny z miozyną. Rozerwanie połączenia przez TNI przy wzroście stężenia wapnia powoduje przemieszczenie tropomiozyny (być może właśnie wskutek zwolnienia jej z termodynamicznie niekorzystnej pozycji) i umożliwia reagowanie aktyny z miozyną. Taki mechanizm najlepiej tłumaczy dotychczas poznane fakty doświadczalne, nie można jednak wykluczyć pewnych modyfikacji w wyniku dalszych badań.

połączenie troponiny z filamentem



Rys. 16. Schemat regulacji cyklu skurczowo-rozkurczowego przez troponinę i wapń

Opisany mechanizm regulacji cyklu skurczowo-rozkurczowego jest charakterystyczny dla mięśni szkieletowych kręgowców. W mięśniach wielu zwierząt niższych nie znaleziono troponiny, najprawdopodobniej też białko to nie występuje w mięśniach gładkich kręgowców. Istnieją dowody przemawiające za tym, że w mięśniach gładkich rolę akceptora jonów wapnia podczas skurczu odgrywa jeden z lekkich łańcuchów miozyny.

Dokładniejsze omówienie różnic w budowie mikroskopowej wszelkich typów mięśni oraz molekularnej organizacji i funkcjonowania ich białek strukturalnych i regulujących przekracza ramy tego artykułu. Mimo różnic w charakterystycznych właściwościach poszczególnych typów mięśni, podstawą mechanizmu skurczu wszystkich mięśni jest reagowanie ze sobą filamentów miozynowych i aktynowych w trakcie hydrolizy ATP — źródła energii dla skurczu, oraz regulacja tego procesu poprzez zmiany stężenia jonów wapnia.

mechanizm skurczu wszystkich mięśni

Biomechanika mięśni

Kazimierz Fidelus, Krzysztof Kędzior i Adam Morecki

Przedmiotem biomechaniki jest badanie ruchu organizmów żywych, a w szczególności człowieka, przy korzystaniu z praw mechaniki. Zakres badań biomechanicznych obejmuje zarówno mechaniczne jak i biologiczne aspekty ruchu. W organizmach żywych poruszają się nie tylko poszczególne części ciała, lecz również organy wewnętrzne, ciecze w naczyniach krwionośnych i limfatycznych, powietrze w układzie oddechowym itp. Strona mechaniczna tych ruchów jest jeszcze mało zbadana. Stosowanymi metodami badawczymi są analiza i synteza ruchu prowadzone na podstawie pomiarów różnych jego parametrów i charakterystyk: kinematycznych, dynamicznych i regulacyjnych. Bardzo pomocne przy tych badaniach jest modelowanie układów ruchu i układów sterowania ruchem.

Rozwój nauk biologicznych, medycznych i cybernetyki wpływał i wpływa na rozwój biomechaniki. Ponieważ omawiana dziedzina ma charakter interdyscyplinarny, do rozwiązywania powstających zagadnień potrzebne są zespoły grupujące badaczy o różnym przygotowaniu zawodowym. Teoria biomechaniki formująca się od niedawna korzysta z jednej strony z ogromnego materiału doświadczalnego, zgromadzonego w ciągu ostatnich stukilkudziesięciu lat, a z drugiej strony — z wyników w dziedzinie modelowania matematycznego i symulacji systemów uzyskanych ostatnimi czasy.

analiza ruchu

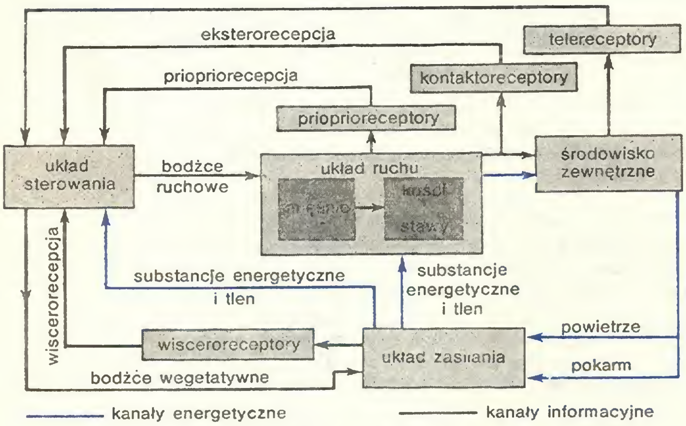
Analizę ruchu można przeprowadzić opierając się wyłącznie na podstawach anatomicznych i fizjologicznych oraz obserwując zewnętrzne przejawy ruchu. Metodami służącymi do takiej obserwacji są: technika filmowa i elektromiografia (rejestracja elektrycznych potencjałów mięśniowych). Ten kierunek badawczy otrzymał nazwę kinezjologii. Można także, korzystając z metod i praw mechaniki ciał sztywnych i odkształcalnych, opisywać ruch (np. ruch ciała ludzkiego lub jego części, czy przepływ cieczy w naczyniach) w postaci zależności matematycznych. Jest to podejście charakterystyczne dla współczesnych badań biomechanicznych. W razie potrzeby opis matematyczny obejmuje także towarzyszące zjawiskom mechanicznym zjawiska elektryczne, cieplne i inne. Na podstawie takich badań proponuje się różne modele organizmu, jego części, narządów wewnętrznych i zewnętrznych, które by wyjaśniły m.in. zagadnienia koordynacji ruchu.

W ostatnich latach biomechanika rozwija się

w trzech podstawowych kierunkach: mechanika człowieka traktowanego jako całość, mechanika kończyn człowieka i mechanika sterowania ruchem człowieka lub jego poszczególnych części ciała. Rozwijają się także zastosowania w medycynie, technice i sporcie. Stąd rozróżnia się biomechanikę medyczną, inżynierską i ćwiczeń sportowych.

główne układy ruchu człowieka

Systemowe ujęcie głównych układów człowieka uczestniczących w ruchu pokazano na rys. 1. Model przedstawiony na tym rysunku składa się z trzech podstawowych układów, a mianowicie: ruchu, sterowania i zasilania. Czwarty układ reprezentuje środowisko, z którym współdziała cały system. Ogólne sprzężenia oraz wpływy zewnętrzne i wewnętrzne w układach zaznaczono na rysunku odpowiednimi



Rys. 1. Schemat blokowy głównych układów współuczestniczących w ruchu organizmu, gdzie propriorecepcja oznacza informacje od receptorów czucia własnego mięśni, ścięgien i stawów, eksterorecepcja — informacje o stanie środowiska zewnętrznego (wzrokowa, słuchowa, powonienia, dotyku itp.), wisceroreccepcja — informacje płynące od trzew (viscera), ciśnienie krwi, stężenie CO_2 itp.

strzałkami. Układ ruchu składa się z aparatów: kostno-stawowego i mięśniowego. Układ sterowania we współczesnym ujęciu jest hierarchicznym, kilkupoziomowym układem, w którym, przy ogólnie obowiązującej centralizacji sterowania, poszczególne poziomy mają określony stopień autonomii. Bodźce ruchowe są przesyłane do poszczególnych tzw. aktonów mięśniowych wywołujących ruch w stawach. Wyjścia tego układu są jednocześnie wejściami do środowiska zewnętrznego. Informacje o stanie środowiska są przekazywane przez telereceptory i kontaktoreceptory, a informacje o stanie układu ruchu — przez proprioreceptory do układu sterowania. Układ zasilania pobiera pokarm i powietrze z otoczenia, przetwarzając pokarm na substancje energetyczne i powietrze, oraz rozprowadza substancje energetyczne i tlen. Tak ujęty model działalności ruchowej człowieka umożliwia przeprowadzenie różnych analiz jego działania w rozmaitych sytuacjach ruchowych. Szczególnym przypadkiem działania może być działanie informacyjne lub energetyczne układu. Należy podkreślić, że układy ruchu i sterowania są stosunkowo nieźle zbadane, natomiast znacznie mniej jest zbadany układ zasilania.

Układ ruchu

Aparat kostno-stawowy

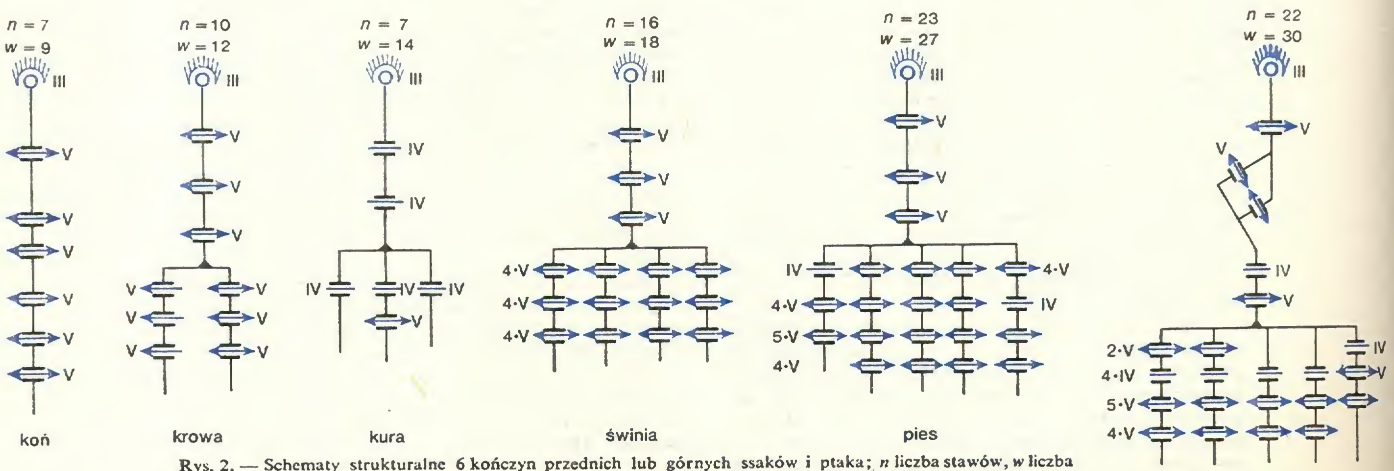
Podamy na początek kilka szczegółowych informacji o kościach i stawach, gdyż wiąże się to z głównym przedmiotem opisu, tj. mięśniami. Liczba kości u człowieka jest zmienna w zależności od wieku i wynosi od 270 u noworodka, przez 356 u ludzi młodych, do 206 u ludzi dorosłych i w miarę starzenia się organizmu nadal maleje. Każda z czterech kończyn u ludzi dorosłych ma 22 kości. Tak więc łączna liczba kości w kończynach wynosi 88, co stanowi ponad 40% ogólnej liczby kości u dorosłego człowieka. Na rys. 2 pokazano schematycznie kończyny przednie lub górne 5 ssaków i ptaka. Nietrudno zauważyć, że poszczególne kości — narysowane jako odcinki — tworzą łańcuchy kinematyczne. Są to łańcuchy otwarte, a ich zamknięcie następuje np. przy unieruchomieniu ręki względem otoczenia lub przy zetknięciu stopy z podłożem, tzn. podczas ruchów lokomocyjnych i stania. Tak samo jest w kończynach dolnych.

liczba stopni swobody

Liczba wszystkich ruchów (zwana liczbą stopni swobody), które możemy wykonać w poszczególnych stawach szkieletu, nie interesując się zakresami ruchów i przyjmując, że każdy z nich można wykonać niezależnie, wynosi 240–250. W tym kończyny górne i dolne umożliwiają wykonanie aż 120 ruchów, czyli ok. 50% wszystkich ruchów. W rzeczywistości niektóre ruchy np. palców u ręki oraz palców u stopy są sprzężone i trudno jest wykonywać niezależne ruchy poszczególnych części palca.

rozmieszczenie stawów

Stawy są tak rozmieszczone, że położone najbliżej tułowia stawy ramienny (rys. 2) i biodrowy umożliwiają wykonanie trzech niezależnych ruchów. Stopniowo przez stawy: łokciowy i promieniowo-nadgarstkowy w kończynie górnej oraz kolanowy i skokowo-goleniowy w kończynie dolnej, przechodzi się do stawów palców, które umożliwiają wykonanie tylko jednego ruchu w każdym stawie. Nasuwa się natu-



Rys. 2. — Schematy strukturalne 6 kończyn przednich lub górnych ssaków i ptaka; *n* liczba stawów, *w* liczba stopni swobody. *III, IV, V* oznacza stawy, o trzech, dwóch i jednym ruchu obrotowym (czyli pary biokinematyczne klasy *III, IV, V*)

ralne pytanie, dlaczego spośród różnych możliwych rozwiązań ukształtował się właśnie taki a nie inny schemat strukturalny aparatu kostno-stawowego. Składa się na to wiele czynników, z których podstawowy to omówiony niżej sposób napędu aparatu kostno-stawowego przez zespół mięśni.

Aparat mięśniowy

O ile układ kostno-stawowy wydaje się, przynajmniej na pierwszy rzut oka, dość przejrzysty i zbudowany wg określonych zasad, to zespół mięśni wywołuje wrażenie bardzo złożonego i nieprzejrzystego. Przyczynia się do tego bardzo duża liczba mięśni, człowiek ma ich ok. 640, które składają się łącznie z ok. 6 mld włókien mięśniowych i ok. 6 bilionów włókienek mięśniowych.

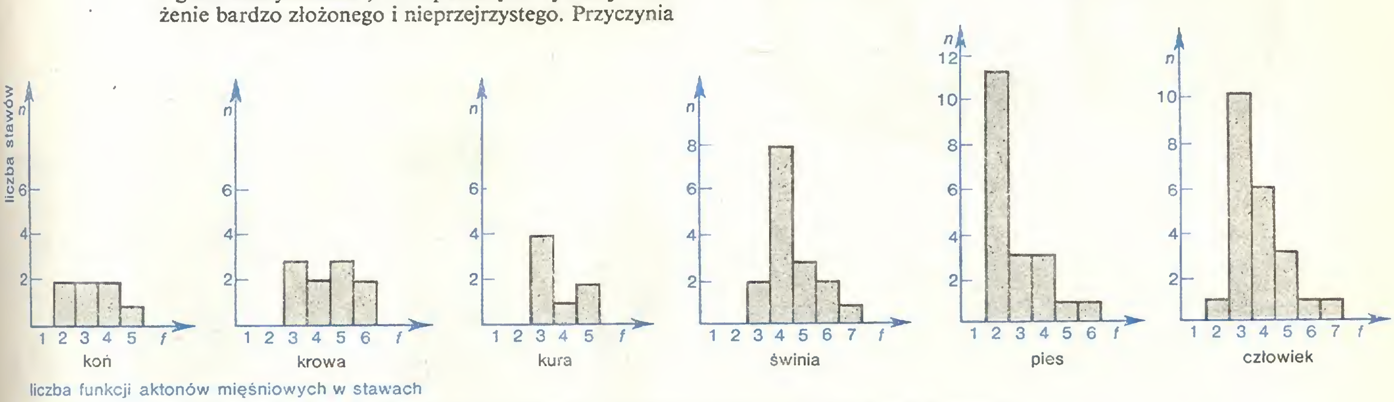
Mięśnie dzielą się na trzy rodzaje: mięśnie szkieletowe, czyli poprzeczne prążkowane, mięśnie gładkie i mięsień sercowy. Układ ruchu jest napędzany mięśniami szkieletowymi, których człowiek ma ponad 440. Do ich sterowania służy ponad 420 tys. komórek nerwowych (tzw. komórek ruchowych) znajdujących się w rdzeniu kręgowym. Każda z tych komórek steruje działaniem zespołu włókien mięśniowych. Zespół taki, zwany jednostką motoryczną (ruchową) zawiera od kilku (4–6 w mięśniach poruszających gałką oka) do kilkuset (640 w dużych mięśniach nóg) włókien mięśniowych.

jednostka motoryczna

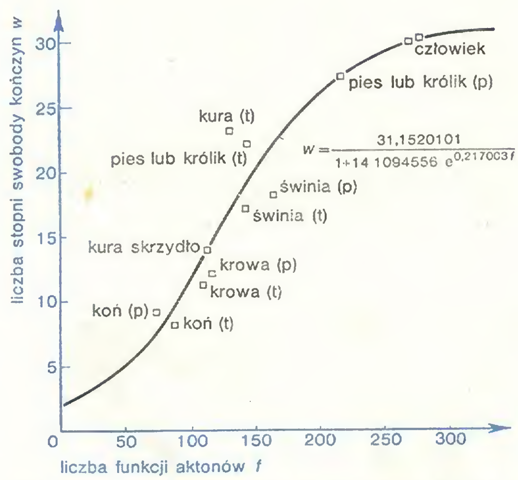
Mięśnie szkieletowe, w zależności od przebiegu włókien mięśniowych względem osi podłużnej mięśnia, dzielą się na mięśnie obłe, pierzaste, płaskie, okrężne i inne. Niektóre mięśnie mają kilka głów lub części i dzieli się je wówczas na tzw. aktony, ponieważ każda z części mięśnia wykonuje inne funkcje w stawach, ponad którymi przebiega. Tak więc przez pojęcie aktonu rozumie się tę część mięśnia, której włókna są tak usytuowane, że rozwijają (wytwarzają) siłę w jednym określonym kierunku, co jest równoważne rozwijaniu momentu siły względem jednej z osi obrotu w stawie. Wiele mięśni składa się z jednego tylko aktonu. W zależności od sposobu działania aktonu względem osi stawu wyróżnia się jego następujące funkcje: prostowanie i zginanie (ruch w płaszczyźnie strzałkowej), supinację i pronację (ruch w płaszczyźnie poprzecznej) oraz przywodzenie i odwodzenie (ruch w płaszczyźnie czołowej). W stawach o trzech stopniach swobody (np. staw ramienny człowieka) występuje wszystkie sześć funkcji.

aktony

Badanie rozmieszczenia mięśni szkieletowych nasuwa wniosek, że natura zastosowała tutaj bardzo skomplikowany system, który nie poddaje się łatwo analizie. Jak wynika z zależności przedstawionych na rys. 3, w organizmach o strukturach prostych występuje niewielkie zróżnicowanie funkcji. W strukturach rozwiniętych rozkład możliwych funkcji jest bardziej zróżnicowany. Świadczy to o przystosowaniu danej kończyny do wypełniania różnych funkcji. W ostatnich latach udało się otrzymać ogólny związek opisujący współzależność pomiędzy liczbą stopni swobody kończyn (suma liczby stopni swobody stawów kończyny) a liczbą funkcji mięśni napędzających daną kończynę (suma funkcji wszystkich mięśni). Na rys. 3 podano kilka charakterystycznych liczb dotyczących kończyn wybranych ssaków i ptaka. Przyjmując, że kończyna górna człowieka jest najbardziej zaawansowana pod względem wszechstronności rozwoju ustalono, że związek pomiędzy liczbą stopni swobody w a liczbą funkcji aktonów mięśniowych f może być aproksymowany podanym na rys. 4 równaniem, którego wykres ma postać tzw. krzywej logistycznej. Ta prawidłowość pozwala na wyciągnięcie różnych wniosków natury biologicznej i technicznej.



Rys. 3. Zależność pomiędzy liczbą stawów a liczbą różnych funkcji aktonów mięśniowych w tych stawach



Rys. 4. Zależność pomiędzy liczbą stopni swobody kończyn w a liczbą funkcji f aktonów różnych zwierząt; p kończyna przednia, t kończyna tylna. Zależność ta ma postać tzw. krzywej logistycznej

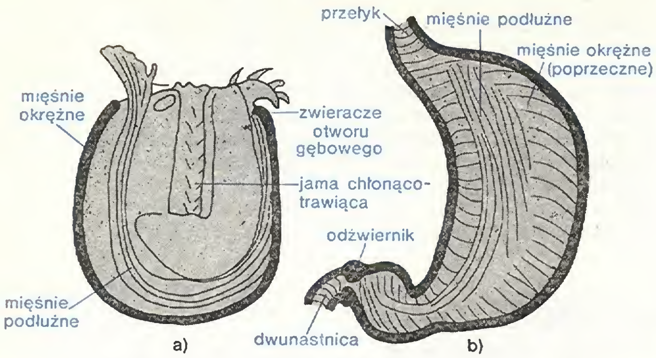
Rodzaje mięśni i ich możliwości

Ewolucja mięśni

witki, rzęski i mięśnie

U zwierząt jednokomórkowych narządem ruchu są różnego rodzaju witki i rzęski, wewnątrz których znajdują się nici aktyny i miozyny (→ Molekularne podstawy skurczu mięśnia). Zwierzęta wielokomórkowe mają bardziej urozmaicony aparat ruchowy, w którym nitki aktyny i miozyny otoczone sarkoplazmą grupują się, tworząc włókna mięśniowe i mięśnie. Mięsień,

zgodnie z III zasadą dynamiki Newtona, może rozwijać siłę (akcja), jeśli na końcach jego przyczepów lub wewnątrz jam ciała, które otacza, będzie występować opór (reakcja).

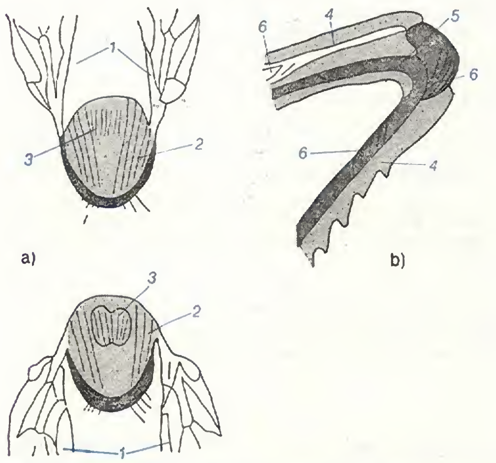


Rys. 5. Rozmieszczenie mięśni gładkich: a) ukwiału końskiego, b) w żołądku człowieka

aparaty ruchowe różnych zwierząt

Niżej rozwinięte zwierzęta, np. jamochłony i płazińce poruszają się całym ciałem, które tworzy wór skórno-mięśniowy. Pasma włókien mięśniowych biegną u nich w różnych kierunkach, powodując wydłużanie się lub grubienie całego ciała, zwieranie otworu gębowego lub wciąganie do wewnątrz niektórych części ciała albo pokarmu (rys. 5). U skorupiaków mięśnie są już bardzo dobrze rozwinięte, do czego przyczynia się niewątpliwie obecność twardej skorupy. Małże np. posiadają silne zwieracze dwu połówek muszli, która je otacza; ślimak natomiast posiada „nogę" przytwierdzoną pośrednio do skorupy, dzięki której może się, wprawdzie bardzo powoli, poruszać.

U stawonogów obserwuje się duże różnice w budowie aparatu ruchowego w porównaniu ze zwierzętami niższymi. Po pierwsze, niższe od nich formy poruszają się bez udziału specjalnych części ciała zwanych kończynami (nogami). Występujące u nich witki, rzęski a nawet „noga" ślimaka są tylko pseudonogami, ponieważ nie mają własnego szkieletu. Po drugie, poruszają się one za pomocą mięśni gładkich, które są powolne, choć bardzo wytrzymałe na zmęczenie. Mięśnie te nie potrzebują twardego szkieletu, ale równocześnie nie potrafią rozwinąć dużej mocy mechanicznej, niezbędnej np. do latania w powietrzu. U stawonogów za to pojawiają się: twardy zewnętrzny szkielet chitynowy oraz najbardziej sprawne żywe silniki — mięśnie poprzecznie prążkowane (rys. 6).

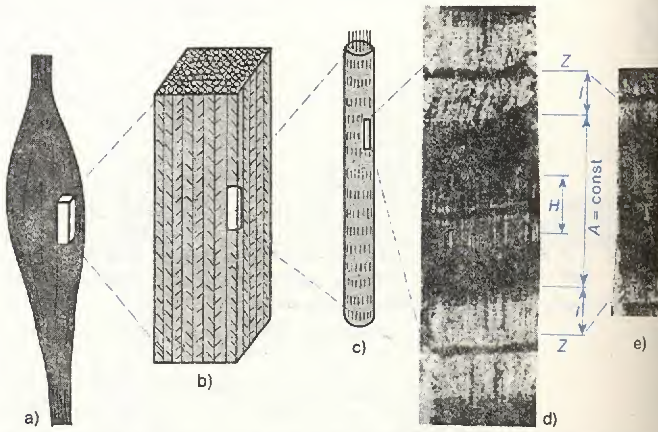


Rys. 6. Schemat działania mięśni u owadów: a) mięśnie poruszające skrzydłami *1*. W rzeczywistości poruszają one tułowiem i znajdującą się na nim płytką grzbietową. Mięśnie poprzeczne *2* powodują ruch skrzydeł ku górze a mięśnie podłużne *3* — do dołu. b) Budowa stawu nogi owada. Człony chitynowe *4* połączone tkanką wiotką tworzą staw *5*, poruszany znajdującymi się wewnątrz mięśniami *6*

Mięśnie szkieletowe

budowa mikroskopowa

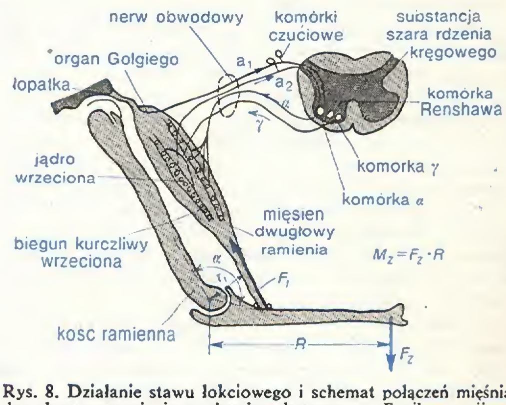
Mikroskopowa budowa mięśnia szkieletowego wskazuje, że składa się on z setek lub tysięcy równoległych komórek — włókien mięśniowych otoczonych wspólną osłoną. Na obu końcach mięsień przechodzi w ścięgna, które łączą się z kośćmi szkieletu (rys. 7).



Rys. 7. Budowa mięśnia poprzecznie prążkowanego: a) mięsień, b) wycinek mięśnia, c) włókno mięśniowe, na którym widać prążki poprzeczne oraz włókienka, d) sarkomer (który ograniczają linie *Z*) przed skurczem, e) sarkomer w stanie skurczu

włókno mięśniowe

Włókno mięśniowe z kolei składa się z zespołów włókienek mięśniowych (miofibryli) osłoniętych owłóknią mięśniową, która jest cienką (ok. 2,5 μm) przejrzystą błoną (rys. 7b). Miofibryle są elementami kurczliwymi włókien mięśniowych, które występują we włóknach mięśni gładkich, mięśnia sercowego i mięśni szkieletowych. W tych ostatnich i w mięśniu sercowym mają one budowę prążkowaną nadającą charakterystyczny wygląd całemu włóknu obserwowanemu pod mikroskopem. Włókno mięśniowe ma średnicę



Rys. 8. Działanie stawu łokciowego i schemat połączeń mięśnia dwugłowego ramienia z rdzeniem kręgowym; F_t siła rozwijana przez mięsień, r_t promień siły mięśniowej, F_z obciążenie zewnętrzne stawu, *R* promień siły F_z

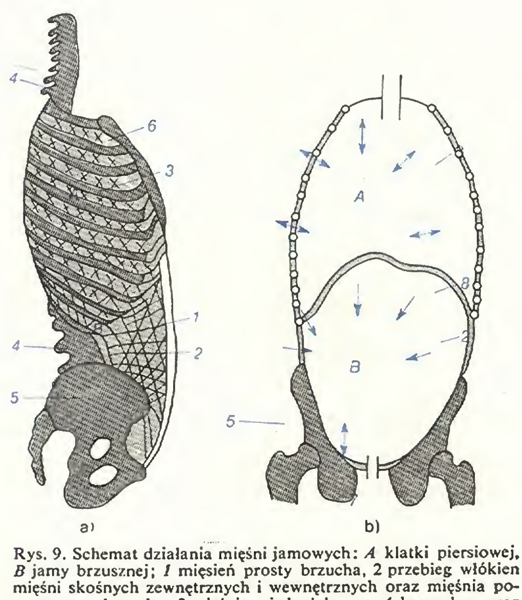
10–100 μm i długość 3–5 a nawet 12 cm. Miofibryle mają grubość ok. 0,5–2 μm (rys. 7c) i składają się z setek elementów strukturalnych, tzw. filamentów, które z kolei składają się z łańcuchów cząsteczek białkowych rozmieszczonych wg ściśle określonego planu

(rys. 7d) (→ Molekularne podstawy skurczu mięśnia). Badania ultracienkich preparatów mięśniowych metodą mikroskopii elektronowej wykazały, że miofibryle są zbudowane z dwóch rodzajów filamentów, a mianowicie grubszych (miozynowych) o średnicy ok. 1000 nm i długości 1,5 µm i cienkich (aktynowych) o średnicy 400–550 nm i długości 2 µm. Jak wynika z rys. 7d, filamenty w mięśniach poprzecznie prążkowanych częściowo pokrywają się i w rezultacie prążki *A* są widoczne w strefach nakładania się cienkich i grubych filamentów, prążki *I* odpowiadają nakładaniu się tylko cienkich, a strefy *H* — nakładaniu się grubych filamentów. Wąskie pasmo rozdzielające na połowę cienkie filamenty odpowiada tzw. linii *Z*.

Mięśnie szkieletowe kręgowców mają bardzo różne kształty: obłe, czworoboczne, trójkątne, wachlarzowate, okrężne itp. Poruszane przez nie kości wykonują ruchy obrotowe w stawach. Na rys. 8 pokazany jest dla przykładu staw łokciowy człowieka poddany obciążeniu siłą F_z.

mięśnie wywołujące zmianę ciśnienia

Jest jednak pewna odrębna grupa mięśni poprzecznie prążkowanych, która podczas skracania się wywołuje zmianę ciśnienia w obrębie jam naszego ciała. Jama ustna jest otoczona mięśniami, których praca wywołuje ciśnienie ujemne (zassanie) lub dodatnie (dmuchanie). Podobnie działają mięśnie otaczające klatkę piersiową oraz jamę brzuszną (rys. 9). Pierwsze zapewniają wdychanie i wydychanie powietrza z płuc, ale oddychanie może się również odbywać przy aktywnym udziale mięśni brzucha. Łatwo to wyczuć, np. podczas kaszlu. Jednak główne zadanie mięśni brzucha polega na utrzymaniu narządów na właściwym miejscu oraz usztywnianiu tułowia podczas przekazywania pędu z kończyn dolnych lub górnych na resztę ciała człowieka. Narządy wewnętrzne, takie jak: serce, wątroba, żołądek lub jelita wypełnione płynami mają masy rzędu 1 kg (np. serce ok. 0,5 kg, wątroba ok. 2 kg). Przyspieszenie ciała wynosi np. przy skoku wzwyż 2–3 *g* (*g* — przyspieszenie ziemskie), co wywołuje pojawienie się dużych sił bezwładności działających na te narządy. Siły bezwładności są równoważone przez zwiększone ciśnienie w jamach ciała.



Rys. 9. Schemat działania mięśni jamowych: *A* klatki piersiowej, *B* jamy brzusznej; *1* mięsień prosty brzucha, 2 przebieg włókien mięśni skośnych zewnętrznych i wewnętrznych oraz mięśnia poprzecznego brzucha, *3* mięśnie międzyżebrowe, *4* kręgosłup wraz z żebrami, *5* miednica, *6* mostek, *7* przepona moczopłciowa, *8* przepona piersiowa

mięśnie usztywniające szkielet

Szkielet tułowia człowieka ma 51 kości i 49 stawów, które mają łącznie 104 stopnie swobody. Przy wielu ruchach cały ten układ musi być usztywniony przez mięśnie poprzecznie prążkowane w celu przenoszenia sił działających na kończyny. Odbywa się to zarówno podczas odbicia nogami (skok, bieg, chód), jak i podczas rozwijania dużych sił za pomocą kończyn górnych (zwisy, podnoszenie ciężarów itp.).

powiązanie mięśnia z układem nerwowym

Przebieg włókien nerwowych między rdzeniem kręgowym a mięśniami szkieletowymi, na przykładzie mięśnia dwugłowego ramienia pokazano na rys. 8. Włókna mięśniowe są pobudzane z komórek ruchowych rogów rdzenia czyli komórek i włókien α. Informacja o długości i prędkości skracania mięśni przebiega do rdzenia od środkowego odcinka wrzeciona mięśniowego (włókna a_1 i a_2). Układ nerwowy uzyskuje informację o sile rozwijanej przez mięsień od organów Golgiego, które znajdują się w ścięgnach mięśni. Wrzeciona mięśniowe biegną równolegle do włókien mięśniowych. Na biegunach mają one części kurczliwe, które skracają się pod wpływem bodźców wysyłanych z komórek γ. Skurcz części biegunowej powoduje rozciąganie części środkowej wrzeciona i wytwarzanie impulsów przekazywanych do rdzenia kręgowego.

Jednostki motoryczne mięśnia szkieletowego działają zgodnie z zasadą „wszystko albo nic" (prawo Bowditcha), co oznacza, że rozwijają one maksymalną siłę na jaką je aktualnie stać, lub że ich naprężenie aktywne jest równe zeru. Takie działanie czyni mięsień szkieletowy mało przydatnym do utrzymania ciała w położeniu statycznym. Szkielet zewnętrzny stawonogów utrzymuje ciało w bezruchu bez udziału mięśni, natomiast wewnętrzny szkielet kręgowców wymaga stałego napięcia mięśni dla utrzymania postawy. Jeśli uśpić pająka za pomocą chloroformu, to będzie on stał nadal. Uśpienie kręgowca spowoduje natomiast jego przewrócenie się. Ze względu na warunki statyki ciała, mięsień szkieletowy jest zatem lepiej dostosowany do funkcji pełnionych u stawonogów, u których pierwotnie się wykształtował, niż u kręgowców. Wspomniana wyżej, niejednoczesna praca jednostek motorycznych u kręgowców występuje również u stawonogów, u których jednak cały mięsień jest unerwiony przez jedną komórkę ruchową i dlatego proces koordynacji (sterowania) ruchu jest u nich znacznie ułatwiony.

właściwości mięśni szkieletowych

Prawie połowa masy ciała człowieka składa się z mięśni i wszystkie nasze czynności zewnętrzne wykonywane są za ich pośrednictwem. Mięsień jest maszyną biochemiczną pracującą w stałej temperaturze. Nawet izolowany, utrzymywany sztucznie poza organizmem mięsień może podnieść ciężar o masie 2000 razy większej od jego własnej. Mięśnie dwugłowy ramienia (*biceps*) i ramienny (*brachialis*) u dobrze zbudowanego mężczyzny mają masę 200–300 g. Podczas pojedynczego skurczu wykonują one pracę równą pracy podniesienia masy 1 kg na wysokość 90 m. Przy pojedynczym skurczu temperatura tych mięśni podnosi się o ok. 0,003°C. Przy każdym skurczu znaczna część energii chemicznej zostaje zamieniana na ciepło i rozproszona. Należy podkreślić, że energia skurczu przewyższa ponad 1000 razy energię pobudzenia w mięśniu. Na ogół mięśnie pracują w niekorzystnych warunkach mechanicznych, przy przełożeniach dźwigni kostnych w stosunku ok. 10:1 (rys. 8). Tracąc na sile, wywołują one w stawach duże prędkości kątowe. Pomimo to można przez okres kilku sekund utrzymać w dwóch rękach, przy poziomej pozycji przedramienia, przedmiot o ciężarze 900 N. Siła rozwijana przez mięśnie musi być w tym położeniu 10 razy większa, czyli wynosić 9000 N. Podnoszona masa jest prawie 2000 razy większa od masy pracujących mięśni. Jako ciekawostkę można podać, że największy mięsień człowieka pośladkowy wielki (*gluteous maximus*) rozwija siłę 12 000 N. Sprawność mechaniczna tej znakomitej maszyny wynosi 25% lub więcej podczas pracy w odpowiednich warunkach. Mięsień osiąga więc sprawność silnika spalinowego. Siła rozwijana przez pojedyncze włókno mięśniowe przy stosunku jego średnicy do długości jak 1:5000, wynosi 1–3 N, a naprężenie mięśnia wynosi 70–

100 N/cm² przekroju poprzecznego. Masa 1 kg mięśnia może w ciągu 8 h rozwinąć moc ok. 15 W.

Człowiek może zaangażować jednocześnie $^{1}/_{7}$ swoich mięśni. Gdyby natomiast te wszystkie mięśnie miały ten sam punkt przyczepu, to mogłyby działać z siłą 250 000 N. Niezależnie od wymienionych zalet mięsień szkieletowy ma również swoje wady. Jego skurcz jest bardzo gwałtowny, co powoduje konieczność amortyzacji całego ruchu. Dlatego w mięśniu szkieletowym występują podatne elementy, takie jak błona *Z*, a zwłaszcza włókna łącznotkankowe tworzące ścięgna mięśni. Rolę amortyzatorów u stawonogów pełni również szkielet chitynowy, a u kręgowców chrząstki stawowe. Mięsień szkieletowy, w odróżnieniu od gładkiego, musi po skurczu odpocząć, co uwzględnia układ nerwowy kręgowców, pobudzając włókna mięśniowe nie równocześnie, lecz po kolei — część z nich pracuje, inne odpoczywają. Równoczesne pobudzenie prawie wszystkich włókien danego mięśnia w ruchach pojedynczych u kręgowców występuje niezmiernie rzadko, np. podczas gwałtownego skoku.

wady mięśni szkieletowych

Mięśnie gładkie

Mięśnie szkieletowe są przyczepione do szkieletu kostnego i chrząstkowego. Mięśnie gładkie natomiast nie przyczepiają się do elementów sztywnych. Znajdują się w ścianach przewodu pokarmowego, naczyń krwionośnych, układu moczowego, narządów rodnych i dróg oddechowych. W pewnym sensie „szkieletem" dla mięśni gładkich jest ciśnienie płynów przemieszczanych przez te mięśnie.

budowa włókien

Włókna mięśni gładkich (miocyty) mają kształt wrzecionowaty. Długość ich wynosi 30–500 μm, a średnica 5–10 μm. Z zewnątrz otoczone są błoną, a wnętrze włókna jest wypełnione sarkoplazmą, którą wypełniają prawie całkowicie miofibryle o średnicy 0,2–0,3 μm. W obrębie miofibryli włókien mięśni gładkich występują mniejsze elementy strukturalne, tzw. miofilamenty o średnicy ok. 5 nm. Kilka lub kilkanaście włókien mięśniowych tworzy pęczek. Niekiedy zakończenia włókien mięśni gładkich kończą się sprężystymi, splatającymi się ze sobą ścięgnami, dzięki czemu powstaje elastyczno-kurczliwy zespół, który reguluje wielkość przekroju przepływu płynu przez naczynie.

właściwości mięśni gładkich

Brak sztywnego szkieletu, do którego byłyby przyłączone mięśnie gładkie oraz ich funkcja w organizmie, determinują mniejszą dynamikę ich działania. Nici aktyny i miozyny w mięśniach gładkich nie są uporządkowane przestrzennie. Dlatego nie występuje w nich poprzeczne prążkowanie. Brak uporządkowania przestrzennego miofilamentów powoduje również mniejszą zmienność siły w zależności od długości mięśni gładkich. Ich wydłużenie względne jest większe niż mięśni szkieletowych. Na przykład mięśnie gładkie macicy w okresie ciąży wydłużają się kilkunastokrotnie w stosunku do długości spoczynkowej. Zmiana siły mięśni gładkich w funkcji ich długości zależy bardzo znacznie od prędkości ich rozciągania. Powolne rozciąganie ma niewielki wpływ na zmianę ich siły, natomiast szybkie ich rozciąganie powoduje istotny wzrost siły.

Czynność tych mięśni jest bardziej powolna i charakteryzują ją automatyczne skurcze rytmiczne. Skurcze te odbywają się niezależnie od świadomości. Czas skurczu wynosi ok. 3 s, a cały cykl z okresem utajonym i zluźnieniem trwa do 20 s. Pracują więc one 200 razy wolniej niż mięśnie szkieletowe.

Mięsień sercowy

Trzeci typ mięśnia to mięsień sercowy. Włókna mięśnia sercowego wykazują prążkowanie podłużne i poprzeczne, przy czym prążki poprzeczne przypominają analogiczne prążki występujące we włóknach mięśni szkieletowych. Skurcz pojedynczy mięśnia sercowego reagującego na pojedynczy bodziec ma przebieg podobny do skurczu mięśni szkieletowych. Jednakże czas trwania skurczu mięśnia sercowego jest znacznie dłuższy. Prądy czynnościowe mięśnia sercowego są zazwyczaj rejestrowane za pomocą elektrod umocowanych w wybranych miejscach powierzchni ciała. Ich zapis w funkcji czasu nosi nazwę elektrokardiogramu (EKG).

skurcze serca

W odróżnieniu od mięśni szkieletowych w mięśniu sercowym nie występuje zależność siły skurczu od wartości pobudzenia. Bodziec progowy wywołuje od razu maksymalny skurcz mięśnia sercowego. Skurcze serca są automatyczne. Częstość skurczów (zwana również częstością tętna) wynosi u dorosłego człowieka ok. 70/min. U dzieci jest ona znacznie większa (noworodek — 135/min, dziecko pięcioletnie — 105/ /min) i w miarę dojrzewania zmniejsza się stopniowo. W czasie dużych wysiłków częstość skurczów może wzrosnąć do 180/min, a u zawodników nawet do 220/min.

wydolność serca

Serce pompuje bez przerwy krew — w spoczynku ok. 4 l/min, co stanowi ok. $2 \cdot 10^6$ l/rok lub ok. $130 \cdot 10^6$ l w ciągu 65 lat. Podczas wysiłku wydolność serca jest większa, w czasie intensywnej pracy może wzrosnąć ośmiokrotnie w stosunku do wydolności spoczynkowej serca. Ciśnienie w tętnicach jest wysokie, równe w przybliżeniu ciśnieniu słupa wody o wysokości 1,5–1,8 m. Tak więc, w ciągu roku serce człowieka pozostającego w spoczynku wykonuje pracę równą podniesieniu ok. $2 \cdot 10^6$ l krwi na wysokość 1,5–1,8 m lub podniesieniu 340–410 l krwi na wysokość Mount Everestu. U człowieka wykonującego przeciętną pracę serce podnosi rocznie ok. 1 t krwi na wysokość Mount Everestu.

Badanie właściwości biomechanicznych mięśni szkieletowych

Badaniami różnych właściwości mięśni zajmują się fizjologowie od przeszło 100 lat, a w ostatnich dwudziestu kilku latach do badań tych włączyli się również biomechanicy i inżynierowie. Wynika to ze znaczenia mięśni dla utrzymywania przy życiu organizmów żywych. Niemałe znaczenie ma również fakt, że mięsień jest bardzo wdzięcznym obiektem do badań. Starannie wypreparowany, trzymany w odpowiednim roztworze i odżywiany, może całymi tygodniami funkcjonować po oddzieleniu go od właściciela.

Badania właściwości mechanicznych mięśni prowadzone są na mięśniach szkieletowych całkowicie izolowanych, częściowo izolowanych i w organizmie. Każde z tych badań wymaga stosowania odpowiedniej metodyki i odpowiednich stanowisk pomiarowych oraz dostarcza innych informacji.

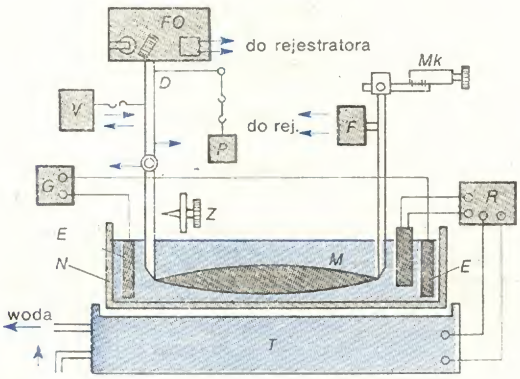
Badanie mięśni izolowanych

Większość badań podstawowych jest prowadzona na mięśniach (lub nawet pojedynczych włóknach mięśniowych) całkowicie izolowanych (*in vitro*). Warunki te umożliwiają bowiem wykonanie selektywnego pomiaru najważniejszych parametrów biomechanicznych mięśni i otrzymanie informacji o związku między nimi, co nie jest możliwe w organizmie żywym ze względu na wpływ czynników anatomicznych (np. mięśnie wielostawowe), a także psychicznych (motywacja), które zakłócają pomiar. Badania takie mają na celu ustalenie związku między siłą rozwijaną przez mięsień a jego długością, prędkością skracania lub wydłużania i wartością pobudzenia (bodźcem stymulacyjnym) przy obciążeniu statycznym lub dynamicz-

nym. Ponadto dąży się do ustalenia związków między wymienionymi parametrami biomechanicznymi a strukturą tkanki mięśnia, zjawiskami energetycznymi i biochemicznymi w mięśniu oraz do matematycznego opisu tych związków w postaci możliwie ogólnego modelu.

Badania te prowadzi się na specjalnie przystosowanych stanowiskach, umożliwiających stworzenie środowiska i warunków zbliżonych do naturalnych, które mięsień znajduje w organizmie. Jak już wspomniano, mięsień wypreparowany z organizmu, a pochodzący od zwierząt o zmiennej ciepłocie ciała (np. od żaby) umieszczony w roztworze fizjologicznym, przez który przetłaczany jest tlen, pobudzany bodźcami elektrycznymi o odpowiednio dobranych parametrach (amplituda, kształt, czas trwania, czas przerw) po pewnym czasie „przyzwyczaja się" do nowego środowiska i stabilizuje swe właściwości biomechaniczne. Na przykład mięsień krawiecki żaby (*sartorius*), który jest mały (przeciętna masa kilkadziesiąt mg) i cienki, a przez to łatwo pobierający tlen z roztworu, ma wystarczające zasoby energetyczne, aby nawet po kilku tygodniach przebywania w roztworze (wolnym od bakterii!) i po tysiącach skurczów nadal zachować swoje właściwości. Rysunek 10 przedstawia schemat

badanie mięśnia żaby



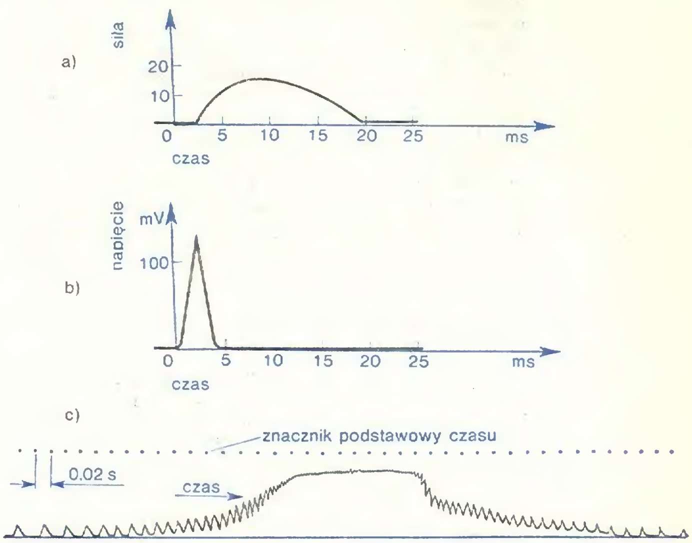
Rys. 10. Stanowisko pomiarowe do badania charakterystyk biomechanicznych mięśnia izolowanego żaby stosowane przez badaczy japońskich. Badany mięsień M umieszczony jest w naczyniu N z roztworem fizjologicznym. Wymiennik ciepła T sterowany przez termoregulator R utrzymuje stałą temperaturę roztworu. Do stymulowania mięśnia służą generator impulsów elektrycznych G oraz dwie elektrody platynowe E zanurzone po obu stronach mięśnia. Jeden koniec mięśnia przyłączony jest do czujnika siły F, a drugi koniec do ruchomej dźwigni D z obciążnikiem P, której ruch jest śledzony przez urządzenie z fotokomórką FO. Do ustalania długości mięśnia służą zderzak Z i urządzenie mikrometryczne Mk. Regulator prędkości V, który może być dołączony do dźwigni utrzymuje stałą, zadaną prędkość skracania lub wydłużania mięśnia

stanowiska do badania właściwości biomechanicznych mięśni żaby, stosowanego przez badaczy japońskich.

Pomiary podstawowych parametrów biomechanicznych mięśni izolowanych prowadzi się zarówno podczas tzw. skurczu izometrycznego (stała długość mięśnia), jak i tzw. skurczu izotonicznego (stała siła rozwijana przez mięsień). W żywym organizmie najczęściej występuje skurcz auksotoniczny, w którym zarówno długość mięśnia, jak i jego siła zmieniają się równocześnie w czasie trwania skurczu.

skurcz włókna mięśniowego

W odpowiedzi na otrzymany pojedynczy bodziec (naturalny — doprowadzony drogą nerwową, lub sztuczny — z generatora), o czasie trwania kilku milisekund i napięciu wyższym od progowego, pojedyncze włókno mięśnia szkieletowego reaguje skurczem pojedynczym, po czym wraca do spoczynku. Rysunek 11a przedstawia zależność siły od czasu po zadziałaniu bodźca stymulacyjnego na utrzymywane w warunkach izometrycznych włókno mięśniowe kręgowca. Przebieg pokazany na rys. 11a nie jest stopniowany, ale jak już wspomniano zgodny z zasadą „wszystko albo nic". Gdy bodziec jest wystarczająco silny, aby spowodować reakcję, to jest ona maksymalna. W czasie kilku milisekund po zadziałaniu bodźca nie obserwuje się odpowiedzi w postaci siłowej, rejestruje się natomiast zmiany potencjału błony komórkowej (rys. 11b).



Rys. 11. Różne charakterystyki włókna mięśniowego kręgowca utrzymywanego w warunkach izometrycznych: a) siła rozwijana w rezultacie zadziałania w chwili $t = 0$ bodźca stymulacyjnego, siła w jednostkach względnych, b) zmiany potencjału na błonie komórkowej włókna, c) przykład oscylogramu przedstawiający sumowanie się bodźców pojedynczych i powstanie skurczu tężcowego

czas trwania skurczu

Czas trwania skurczu pojedynczego zarówno włókna jak i całego mięśnia zależy od temperatury i od rodzaju mięśnia. Podobnie jak w wielu procesach biologicznych i chemicznych wzrost temperatury o 10°C powoduje dwu-, trzykrotny wzrost prędkości przebiegu tego zjawiska. U ssaków czas trwania skurczu pojedynczego jest na ogół mniejszy od 0,1 s, u owadów wynosi ok. 0,01 s, a u zwierząt o zmiennej ciepłocie ciała wydłuża się niekiedy (w niskich temperaturach) aż do kilkudziesięciu sekund. Tak np. mięsień uruchamiający zęby żołądkowe raka lub mięśnie nóg żółwia budzącego się ze snu zimowego, potrzebują ok. 30 s na wykonanie pojedynczego skurczu. Mięśnie skrzydeł pospolitej muchy (*Phormia*) skracają się 120 razy na sekundę, a częstość uderzeń skrzydeł dochodzi do kilkuset na sekundę. Rozpiętość jest więc ogromna i mięśnie owadów poruszają się ok. 3000 razy częściej niż mięśnie raka i żółwia. Na tym tle człowiek wypada korzystnie w porównaniu z żółwiem, ale niezbyt korzystnie w porównaniu z komarem, którego skrzydła uderzają ok. 900 razy na sekundę. Nie potrafimy bowiem poruszać palcami szybciej niż 8–10 razy na sekundę.

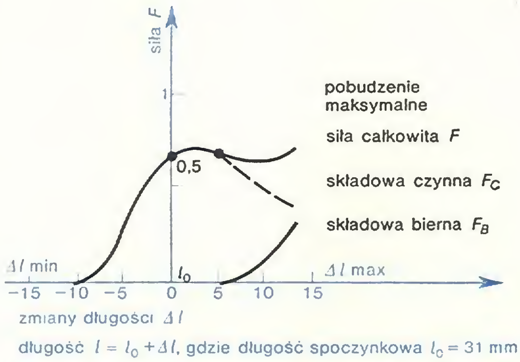
siła rozwijana przez mięsień

Wielkość siły rozwijanej przez cały mięsień, w przeciwieństwie do pojedynczego włókna mięśniowego, zależy od amplitudy bodźca stymulacyjnego. Gdy przekracza ona wartość progową, notuje się słabą odpowiedź mięśnia, która powiększa się aż do momentu osiągnięcia przez bodziec największej wartości. Słabe bodźce pobudzają kilka włókien mięśniowych, natomiast bodziec maksymalny pobudza wszystkie włókna mięśniowe. Podobnie przy pobudzaniu naturalnym siła rozwijana przez mięsień zależy od liczby jego włókien zaangażowanych w wysiłek.

Jeżeli kolejny bodziec jest doprowadzany do włókna mięśniowego w trakcie trwania poprzedniego skurczu, to występuje zjawisko sumowania odpowiedzi siłowych mięśnia. Przy określonej częstości bodźców zjawisko sumowania doprowadza do wystąpienia pełnego (tzw. gładkiego) skurczu tężcowego. Rysunek 11c

przedstawia przebieg siły rozwijanej przez włókno mięśniowe kręgowca w warunkach izometrycznych w trakcie stymulacji o częstości najpierw rosnącej, a następnie — malejącej. Częstość pobudzania wywołująca skurcz tężcowy zależy od rodzaju mięśnia i od temperatury. Dla mięśni szkieletowych żaby w temperaturze 0°C częstość ta wynosi ok. 18 Hz, w temperaturze 20°C ok. 30 Hz, a dla mięśni człowieka ok. 50 Hz. Skurcz tężcowy trwa tak długo, jak długo są dostarczane bodźce lub dopóki nie nastąpi zmęczenie mięśnia. Aby siła maksymalna nie malała, mięsień izolowany wymaga przeciętnie 1 min odpoczynku na 1 s skurczu tężcowego.

Podstawową właściwością mięśnia jest rozwijanie (wytwarzanie) siły. Na rys. 12 przedstawiono podstawową charakterystykę biomechaniczną mięśni na przykładzie charakterystyki uzyskanej dla izolowanego mięśnia krawieckiego żaby (*sartorius*). Charakterys-



Rys. 12. Charakterystyka statyczna mięśnia izolowanego

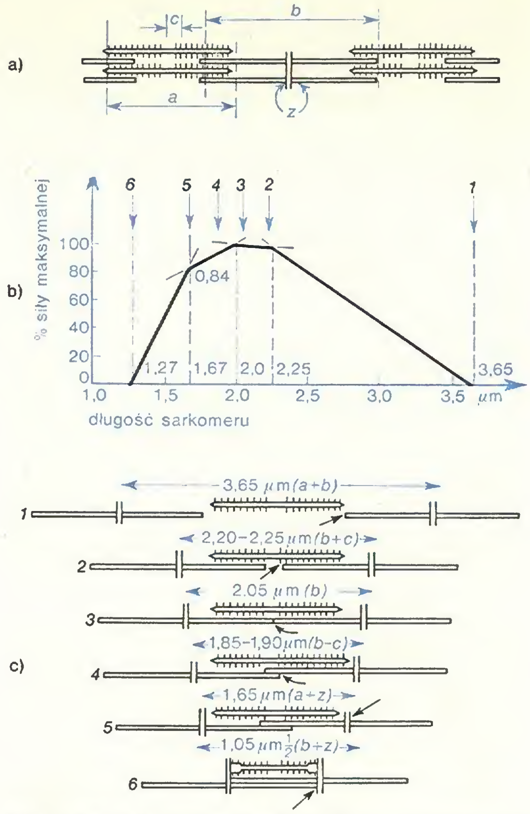
zależność rozwijanej siły od długości

tyka ta zwana charakterystyką statyczną przedstawia zależność siły całkowitej F rozwijanej przez mięsień (podczas izometrycznego skurczu tężcowego i przy stałym pobudzaniu) od długości l. Jest ona sumą składowej biernej F_B i składowej czynnej F_C, czyli $F = F_B + F_C$. Mięsień nie pobudzony, rozciągany powyżej długości spoczynkowej l_0 przeciwstawia się biernie sile rozciągającej z siłą F_B. Natomiast składowa czynna F_C pochodzi od aktywnie kurczących się, pod wpływem pobudzenia, elementów struktury mięśnia i jej wartość zależy od wartości bodźców pobudzających.

Doświadczalnie można mierzyć składową bierną (rozciągając mięsień nie pobudzony) i siłę całkowitą (mierząc siłę rozwijaną przez mięsień pobudzony). Przebieg składowej czynnej F_C dla $l > l_0$ wyznacza się odejmując od siły całkowitej składową bierną. Długość spoczynkowa l_0 w pewnym przybliżeniu odpowiada długości mięśnia w pośrednim położeniu kąta zakresu (amplitudy) ruchu w stawie. Krzywe typu $F = F(l)$, dla których siła osiąga maksimum w okolicy l_0, a długość mięśnia może zmieniać się o kilkadziesiąt procent l_0 w obie strony, są typowe dla prawie wszystkich rodzajów mięśni: szkieletowych, sercowego i gładkich. Wyjątkiem są mięśnie owadów, które rozciągnięte o więcej niż kilka procent ulegają uszkodzeniu. Siła rozwijana przez te mięśnie maleje w miarę ich rozciągania, dzięki czemu mogą one wytwarzać ciągłe drgania, jeżeli są połączone z tułowiem owada i jego skrzydłami, których bezwładność i sztywność są odpowiednio dostrojone do sztywności i częstości drgań mięśni.

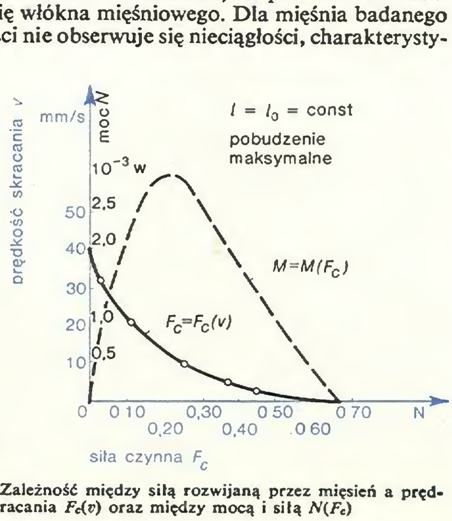
ślizgowa teoria skurczu

Przebieg zależności składowej czynnej F_C od długości mięśnia można wyjaśnić na podstawie ślizgowej teorii skurczu zaproponowanej przez H. E. Huxleya i współpracowników. Na rys. 13 przedstawiono wyniki bardzo precyzyjnych pomiarów przeprowadzonych przez H. E. Huxleya i innych na preparatach pobranych z pojedynczych włókien mięśniowych żaby. Na rys. 13a i c przedstawiono schemat mikrostruktury



Rys. 13. Sarkomer: a) schemat mikrostruktury z podstawowymi wymiarami: a = 1,60 μm, b = 2,05 μm, c = 0,15–0,20 μm, z = = 0,05 μm), b) krzywa siły w funkcji długości (schematyczne przedstawienie wyników, strzałki nad wykresem wskazują różne etapy zachodzenia na siebie nitek, przedstawione na rys. c; c) kolejne etapy zachodzenia na siebie grubych i cienkich nitek podczas skracania się

(por. z rys. 7d) i podano podstawowe wymiary sarkomeru (odległość między dwiema kolejnymi liniami Z). Na rys. 13b podano zależność siły czynnej F_C rozwijanej przez pojedyncze włókno mięśniowe w funkcji długości sarkomeru. Dla pojedynczego włókna mięśniowego zależność rozwijanej przezeń siły czynnej $F_C = F_C(l)$ jest liniowa odcinkami. Nieciągłości w punktach *1*, *2*, *3*, *4*, *5* i *6* wynikają z przejść między kolejnymi fazami współpracy nitek aktyny i miozyny. Dzięki wypustkom, jakie mają nitki miozyny (tzw. mostkom) następuje połączenie miozyny z aktyną z równoczesnym wzajemnym przesuwaniem się — ślizganiem — obu filamentów, co prowadzi do skracania się włókna mięśniowego. Dla mięśnia badanego w całości nie obserwuje się nieciągłości, charakterysty-



Rys. 14. Zależność między siłą rozwijaną przez mięsień a prędkością skracania $F_c(v)$ oraz między mocą i siłą $N(F_c)$

ki są zaokrąglone, ponieważ nawet w obrębie jednego włókna mięśniowego sarkomery nie kurczą się równomiernie. Rysunek 13c przedstawia schemat procesu ślizgania.

charakterystyka dynamiczna mięśni

Na rys. 14 pokazano przykładowy przebieg podstawowej charakterystyki dynamicznej mięśni, czyli zależności między siłą czynną F_C rozwijaną przez mięsień a prędkością skracania mięśnia v przy określonej długości i przy stałym pobudzeniu. Dla innych wartości $l = \text{const}$ i ustalonego pobudzenia przebieg charakterystyk ma podobny kształt. Krzywe tego typu otrzymuje się dla wszystkich rodzajów mięśni, nie tylko szkieletowych, lecz także sercowych i gładkich, a nawet dla preparatów z nitek aktomiozyny. Jedynym wyjątkiem są znowu mięśnie skrzydeł owadów pracujące na zasadzie niewielkich drgań, a nie dużych zmian długości.

Hiperboliczny przebieg zależności $F_C = F_C(v)$ tłumaczono początkowo, zakładając że siła rozwijana przez mięsień jest w rzeczywistości stała przy wszelkich prędkościach, lecz że część tej siły jest zużywana na pokonanie wewnętrznych sił tarcia i lepkości w mięśniu, dlatego nie ujawnia się ona na zewnątrz. Obecnie wydaje się raczej, że na charakter tej krzywej decydujący wpływ ma prędkość zachodzenia reakcji biochemicznych w skracającym się mięśniu. Praktyczny wniosek wypływający z przebiegu zależności $F_C = F_C(v)$ jest taki, że moc mechaniczna rozwijana przez mięsień $N = Fv$ osiąga maksimum w punkcie odpowiadającym mniej więcej jednej trzeciej wartości maksymalnej siły i maksymalnej prędkości. Moc maksymalna rozwijana przez mięsień wynosi ok. $0{,}1 F_{max} v_{max}$. Aby wykonać mechaniczną pracę przy użyciu mięśni w sposób najefektywniejszy, należy dobrać obciążenie zgodne z tą zasadą. Przerzutka rowerowa jest przykładem urządzenia pozwalającego dopasować obciążenie i prędkość do właściwości mięśni, bez względu na pochylenie drogi.

Mierząc energię doprowadzaną i odbieraną można określić sprawność mechaniczną mięśnia, która wynosi ok. 25%. Należy podkreślić, że nawet w warunkach izometrycznych, gdy mięsień nie skraca się, można mówić o wykonywaniu pracy (rozwijaniu mocy) przez mięsień, w odróżnieniu od urządzeń technicznych.

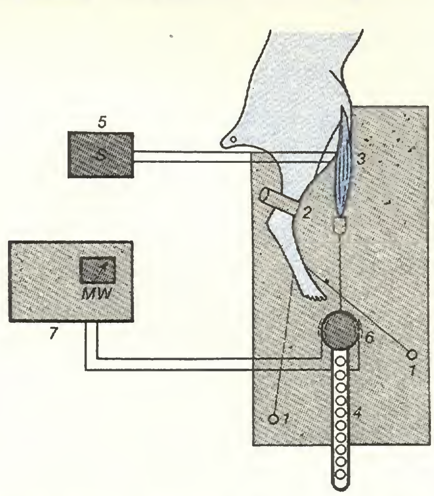
Badanie mięśni częściowo izolowanych i mięśni bezpośrednio w organizmie

Pomiary właściwości biomechanicznych mięśni częściowo izolowanych są pośrednim rodzajem badań. Dostarczają one informacji o mięśniach naturalnie odżywianych i przez to znajdujących się w warunkach bardziej zbliżonych do naturalnych niż mięśnie izolowane. Z drugiej strony konieczność wykonywania pomiarów w ograniczonym czasie, wpływ narkozy i obrzęku mięśni na skutek częściowego ich obnażenia, bardzo utrudnia pomiary i prowadzi zazwyczaj do rozrzutu wyników.

Na rys. 15 przedstawiono dla przykładu zestaw pomiarowy stosowany do badań zależności w rodzaju $F = F(l)$ dla mięśni obłych i pierzastych królika pod narkozą (*in vivo*). Badania na mięśniach pozostających w organizmie (*in situ*) prowadzi się między innymi w celu poznania zjawisk bioelektrycznych towarzyszących skurczowi mięśni. Duża liczba włókien mięśniowych i ich mikroskopijne wymiary skłaniają badaczy do rozpatrywania i mierzenia sumarycznego efektu elektrycznego zespołu włókien, który można traktować jako miarę działania całego mięśnia. Pomiar miopotencjałów mięśni powierzchniowych (znajdujących się tuż pod skórą) nie wymaga bezpośredniego dotarcia do mięśnia. Wystarczy umieszczenie odpowiednich elektrod na skórze w miejscu położonym nad mięśniem. Uzyskany w ten sposób zapis, zwany elektromiogramem (EMG), jest sumą potencjałów pewnej liczby włókien badanego mięśnia znaj-

elektromiogram

dujących się w pobliżu elektrod. Napięcie miopotencjałów zawiera się w przedziale od kilku µV do kilku mV. Za pomocą bipolarnych elektrod igłowych



Rys. 15. Zestaw pomiarowy stosowany do badań częściowo izolowanych mięśni królika. Łapa badanego królika jest umocowana linkami *1* i zaciskiem *2*. Badany mięsień *3* częściowo wypreparowany, jednym końcem odcięty od kości i przyłączony do czujnika siły *6* pobudzany jest bodźcami ze stymulatora *5* doprowadzonymi do mięśnia parą elektrod powierzchniowych. Długość mięśnia ustalona jest przy pomocy płytki z otworami *4*. Wartość siły odczytywana jest na wskaźniku wzmacniacza *7*

o niewielkich wymiarach, które nie powodują odczuwalnych uszkodzeń mięśnia, można otrzymać EMG nawet pojedynczych włókien mięśniowych. Okazuje się, że zamiany potencjałów mięśniowych są w pewnych zakresach proporcjonalne do zmian naprężeń włókien mięśniowych, a tym samym do siły rozwijanej przez mięsień.

Na il. 127a (tabl. 32) pokazano widok ogólny stanowiska pomiarowego stosowanego do badań. Przedmiotem badania jest współdziałanie aktonów mięśniowych obsługujących staw ramienny w warunkach statycznych. Aktonów tych jest siedemnaście i w różny sposób współdziałają one podczas ruchów w stawie. Badany człowiek wykonuje ruchy z różnym wysiłkiem, np. 25%, 50% i 100% wysiłku maksymalnego, a w tym czasie aparatura wzmacniająca i pisząca rejestruje zarówno rozwijaną siłę (uzyskuje się tzw. mechanogramy), jak i przebieg potencjałów mięśniowych (elektromiogramy). Przykład jednoczesnego zapisu mechanogramu i elektromiogramu jednego z aktonów pokazano na il. 127b (tabl. 32). W celu określenia udziału każdego z badanych aktonów mięśniowych należy prowadzić jednoczesny zapis mechanogramów i elektromiogramów dla każdego z nich.

mechanogram

Modelowanie mięśni izolowanych

Badania biomechaniczne mięśni izolowanych dostarczyły bogatego materiału doświadczalnego; nie udało się jednak sformułować na jego podstawie uniwersalnego modelu matematycznego mięśnia. Znane obecnie modele są typu fenomenologicznego i obowiązują tylko w określonych zakresach parametrów i stanów mięśnia.

modele reologiczne

Do najbardziej rozpowszechnionych należą modele typu reologicznego, co wynika z faktu, że właściwości mechaniczne tkanki mięśniowej są zbliżone do właściwości kauczuku i innych elastomerów. Podobieństwo to jest wynikiem podobieństwa budowy chemicznej — długich łańcuchów cząsteczek. W modelu reologicznym materiał rzeczywisty zastępuje się układem

sprężyn, tłumików i innych elementów nie mających odpowiedników w naturze. Jednakże, w przeciwieństwie do innych tworzyw, właściwości żywego mięśnia zmieniają się w zależności od stopnia pobudzenia. Poza tym mięsień może się samodzielnie kurczyć i rozwijać siłę. Fakty te muszą być uwzględnione przy budowie modelu.

Sformułowanie uogólnionego modelu

Sformułowanie uogólnionego modelu jest oparte na przyjmowanym powszechnie założeniu, że siła F rozwijana przez mięsień w funkcji długości mięśnia l, podczas skurczu tężcowego i przy stałym pobudzeniu jest sumą niezależnej od pobudzenia składowej biernej F_B i zależnej od wartości pobudzenia składowej czynnej F_C (rys. 12). Schemat uogólnionego modelu reologicznego mięśnia opracowany w Polsce przedstawiony jest na rys. 16, gdzie *1, 2, 3* oznacza zespoły składowe modelu, E_0, E_1, E_2, E_3 — wartości modułów Younga elementów sprężystych modelu, E_s — moduł Younga charakteryzujący element siłowy modelu (jest on funkcją czasu t napięcia stymulacji U i częstości f), η_1, η_2, η_3 — współczynniki lepkości elementów tłumiących modelu, σ_1, σ_2, σ_3, σ — naprężenia odpowiednio dla zespołów *1*, *2*, *3* i całego modelu.

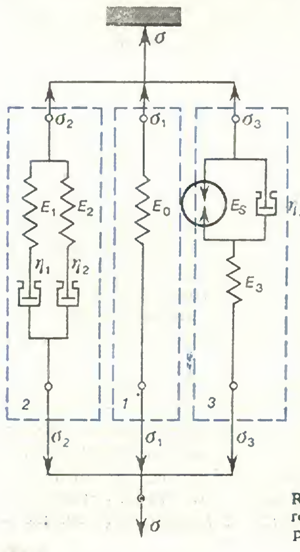
uogólniony model reologiczny

Zakłada się, że parametry E_0–E_3, E_s, η_1–η_3 modelu mają następujące cechy: są nieliniowymi funkcjami długości mięśnia l, nie są zależne od czasu (nie uwzględnia się wpływu takich parametrów, jak np. zmęczenie), nie zależą od pobudzenia mięśnia (rozpatruje się przypadek stałego pobudzenia).

Taki układ połączeń modeli składowych został przyjęty z następujących przyczyn: Mięsień nie pobudzony i rozciągany biernie w warunkach statycznych zachowuje się (rys. 12), jak nieliniowa sprężyna o charakterystyce F_B (l), stan ten charakteryzuje zespół *1* (o parametrze E_0; rys. 17). Stan mięśnia niepobudzonego w warunkach dynamicznych opisać można zespołem *2*, w którym pierwszy element Maxwella o parametrach E_1, η_1 opisuje tzw. składową szybką, a element o parametrach E_2, η_2 — składową wolną.

model składowej czynnej

Zespół *3* modeluje składową czynną siły rozwijanej przez mięsień (rys. 14). Wyniki badań wielu autorów wskazują na fakt, że model składowej czynnej powinien zawierać tzw. element siłowy (kurczliwy), który w tej pracy jest modelowany elementem o module sprężystości E_s. Element ten charakteryzuje wyłącznie zdolność mięśnia do rozwijania siły. Stwierdzono także, iż mięsień pobudzony zmienia swą lepkość oraz zawiera tzw. szeregowy element sprężysty, co w tym modelu jest uwzględnione przez wprowadzenie elementów o parametrach η_3, E_3.



Rys 16. Uogólniony model reologiczny zaproponowany przez badaczy polskich

Składowe uogólnionego modelu mięśnia

Warunki pracy	Stan mięśnia \ Wydłużenie mięśnia	$\varepsilon_{min} \leqslant \varepsilon < 0$	$0 \leqslant \varepsilon \leqslant \varepsilon_{max}$
Statyczne	niepobudzony $U < U_{min}$	$\sigma = 0$	$\sigma = \sigma_1$
	pobudzony $U_{min} \leqslant U \leqslant 1$	$\sigma = \sigma_3$	$\sigma = \sigma_1 + \sigma_3$
Dynamiczne	niepobudzony $U < U_{min}$	$\sigma_1 = \sigma_2$	$\sigma = \sigma_1 + \sigma_2$
	pobudzony $U_{min} \leqslant U \leqslant 1$	$\sigma = \sigma_2 + \sigma_3$	$\sigma = \sigma_1 + \sigma_2 + \sigma_3$

Zależnie od długości mięśnia l oraz jego stanu do opisu zachowania się mięśnia można wykorzystać określone kombinacje modeli składowych podane w tabeli (ε oznacza wydłużenie mięśnia, $\varepsilon = \Delta l/l$, U oznacza bezwymiarowe pobudzenie, $U_{max} = 1$).

modele mechaniczno-energetyczne

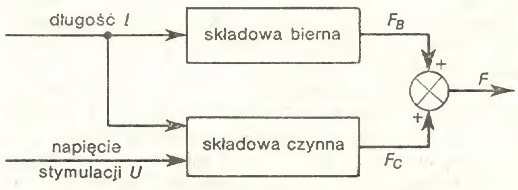
Inny rodzaj modelu mięśni izolowanych (model matematyczny) wynika z tzw. równania charakterystycznego Hilla podawanego zwykle w postaci

$$(F+a)(v+b) = (F_0+a)b,$$

gdzie: F jest siłą rozwijaną izotonicznie przez mięsień, F_0 — siłą rozwijaną izometrycznie przez mięsień przy długości l_0, v — prędkością skracania mięśnia, a, b — współczynnikami. Zależność tę A. V. Hill uzyskał w 1938 r. na podstawie analizy wyników pomiarów mechaniczno-energetycznych (pomiary ciepła wydzielanego przez mięsień w różnych fazach skurczu) i początkowo przypuszczał, że współczynniki a i b mają dla danego mięśnia wartości stałe. Od tamtego czasu teoria Hilla została znacznie rozszerzona i udoskonalona. Obecnie bardzo intensywnie rozwijanym kierunkiem są próby opisania charakterystyk mechanicznych mięśnia $F = F(l)$ i $F = F(v)$ w powiązaniu z przemianami biochemicznymi zachodzącymi podczas jego pracy i z występującymi wówczas zmianami w jego mikrostrukturze.

modele cybernetyczne

W ostatnich latach, w związku z rozpowszechnianiem się metod cybernetycznych, są podejmowane próby modelowania mięśni izolowanych, w których — zgodnie z teorią automatycznego sterowania — mięsień traktuje się jako obiekt wieloparametrowy, mający różne wejścia i wyjścia. Wprowadzając na wejścia takiego obiektu określone sygnały wymuszające pochodzenia mechanicznego i elektrycznego bada się odpowiedź obiektu i na tej podstawie dobiera się równania matematyczne opisujące model. Opisana metoda była zastosowana w badaniach przeprowadzonych w Polsce. Przyjęto, że mięsień izolowany jest wieloparametrowym, nieliniowym obiektem re-



Rys. 17. Model cybernetyczny mięśnia izolowanego

gulacji o dwóch wejściach (długość l i napięcie stymulacji U) i jednym wyjściu (siła $F = F_C + F_B$). Przyjęto także, iż obiekt ten można podzielić na dwa człony, odpowiadające składowym F_B i F_C siły rozwijanej przez mięsień (rys. 17). Na podstawie wyników badań doświadczalnych przeprowadzonych na mięśniu izolowanym brzuchatym łydki (*gastrocnemius*) żaby wyznaczono układ równań różniczkowych opisujących ten model mięśnia.

Wyniki uzyskane przy modelowaniu mięśni opisanymi metodami mają dwa główne kierunki zasto-

sowań: ułatwiają teoretyczne rozpatrywanie skomplikowanych zagadnień dynamicznych biomechaniki ruchu (np. zagadnienia współudziałów mięśni) oraz są pomocne przy syntezie mięśni sztucznych o właściwościach zbliżonych do mięśni naturalnych. Oba wymienione kierunki badań są obecnie intensywnie rozwijane ze względu na podejmowane próby skonstruowania antropomorficznych (człowiekopodobnych) robotów i manipulatorów.

Możliwości rozwoju siły mięśniowej przez trening

siła rozwijana przez jednostkę motoryczną

Siłę F_m rozwijaną przy stałej długości mięśnia przez jednostkę motoryczną należącą do danego mięśnia można wyrazić zależnością

$$F_m = S\sigma,$$

gdzie S jest przekrojem poprzecznym (tzw. przekrojem fizjologicznym) włókien mięśniowych, σ — naprężeniem rozwijanym przez te włókna. Siła rozwijana w danym momencie przez mięsień zależy zarówno od liczby synchronicznie (równocześnie) pobudzanych jednostek motorycznych (ich siły sumują się z odpowiednimi wagami) jak i od częstości ich pobudzania (sumują się skurcze pojedyncze poszczególnych jednostek). Ponieważ działalności mechanicznej włókien mięśniowych towarzyszy aktywność elektryczna, to stan pobudzenia całego mięśnia jest zakodowany zarówno w amplitudzie jak i w częstości elektromiogramu. W celu łącznego ujęcia obu tych parametrów przyjmuje się, że miarą pobudzenia mięśnia jest amplituda U tzw. scałkowanego elektromiogramu, otrzymanego przez uśrednienie (średnia ważona) elektromiogramu wyjściowego w kolejnych krótkich przedziałach czasu.

Przyjmując, że wartość maksymalna scałkowanego elektromiogramu U_{max} jest rejestrowana wówczas, gdy mięsień rozwija siłę maksymalną, tzn. gdy wszystkie jednostki motoryczne są pobudzone, to bezwymiarowy stosunek $\bar{u} = U/U_{max}$ jest proporcjonalny do liczby jednostek motorycznych (czyli do części przekroju fizjologicznego) mięśnia w danej chwili zaangażowanych w rozwijaniu siły.

Siła F_i rozwijana przez i-ty mięsień w warunkach statycznych jest więc funkcją jego przekroju fizjologicznego S_i, średniego naprężenia rozwijanego przez włókna σ_i, długości mięśnia l_i i wielkości $\bar{u}_i$. Długość l_i mięśni w organizmie zależy jednoznacznie od kąta obrotu w stawie α. Dla zespołu mięśni obsługujących staw o jednym stopniu swobody można napisać tzw. równanie udziałów w postaci:

równanie udziałów

$$M_z = \sum_{i=1}^{n} F_i(S_i, \sigma_i, \alpha, \bar{u}_i) \cdot r_i(\alpha),$$

gdzie: M_z jest momentem wypadkowej sił zewnętrznych względem osi stawu, r_i — ramieniem siły F_i względem osi stawu, n — liczbą mięśni (rys. 8). Jeżeli staw ma większą liczbę stopni swobody (np. staw ramienny o trzech stopniach swobody — il. 127a (tabl. 32)) należy rozpatrzyć równanie udziałów dla każdego ze stopni swobody oddzielnie.

Równanie udziałów w kilku odmianach opracowano i zbadano w Polsce. Obszerne badania mięśni obsługujących różne stawy kończyny górnej człowieka: łokciowy, promieniowo-nadgarstkowy, łokciowo-promieniowy i ramienny przeprowadzone w latach 1963–1979 pozwoliły na opracowanie metodyki ustalania udziału poszczególnych mięśni w realizacji określonego aktu ruchowego w warunkach statycznych i dynamicznych.

W sporcie i medycynie jest możliwe wywieranie wpływu na wartość siły mięśni przez odpowiedni trening. Trening mięśni następuje podczas każdej pracy fizycznej, ale jego skuteczność zależy głównie od intensywności wykonywanej pracy. Praca o dużej intensywności oznacza, że organizm rozwija podczas tej pracy dużą moc użyteczną. Pracę intensywną wykonują robotnicy fizyczni, ale nie jest to praca o maksymalnej mocy, ponieważ taka praca nie może być wykonywana przez kilka godzin. Zatem trening mięśni u robotników fizycznych występuje tylko w niewielkim stopniu. Intensywny trening uprawiają tylko sportowcy, dlatego wpływ treningu na zmianę siły mięśni zostanie zilustrowany przykładami zaczerpniętymi ze sportu.

Przekrój fizjologiczny włókien mięśniowych i osiągane przez nie naprężenie zmieniają swoje wartości zarówno pod wpływem pracy (treningu) jak i bezczynności. Zmiany te zależą od rodzaju treningu.

trening szybkościowy

Trening szybkościowy, polegający na krótkotrwałych ćwiczeniach od kilku do kilkunastu sekund, wykonywanych z maksymalną prędkością, powoduje głównie wzrost naprężenia mięśni.

trening siłowy

Trening siłowy i siłowo-wytrzymałościowy powoduje głównie wzrost przekroju fizjologicznego włókien mięśniowych. Trening siłowy polega na kilkusekundowym maksymalnym pobudzeniu mięśni, a siłowo-dowytrzymałościowy na długotrwałej (od kilku do kilkunastu minut) pracy przy średnim obciążeniu mięśni.

Zagadnienie wpływu treningu na wzrost siły mięśniowej można prześledzić posługując się równaniem udziałów. W tym równaniu parametry α i r_i nie podlegają treningowi. Natomiast takie parametry jak przekrój fizjologiczny S_i, naprężenie σ_i i pobudzenie $\bar{u}_i$ istotnie zależą od wytrenowania mięśnia. Ponadto, w warunkach dynamicznych siła mięśniowa zależy także od prędkości skracania mięśnia zgodnie z charakterystycznym równaniem Hilla. Zarówno prędkość skracania mięśnia, jak i wartość siły przy danej prędkości podlegają również wytrenowaniu.

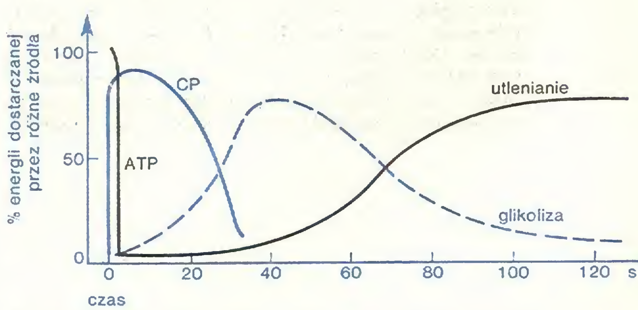
Wartość parametrów siłowych mięśni zależy głównie od stanu układu nerwowego oraz rodzaju źródeł energetycznych, które decydują zarówno o czasie trwania pracy, jak i rozwijanej przez mięśnie mocy

Zasób i moc źródeł energii

Źródło energii	Zasób energii J na 1 kg masy ciała	Moc W na 1 kg masy ciała	Maksymalny czas pracy s
Fosforylacja	420	54,4	4–6
Glikoliza beztlenowa	960	29,3	40–90
Procesy tlenowe	nieograniczony	15*	nieograniczony

* Przy długotrwałych wysiłkach moc zmierza do zera.

maksymalnej. Przy założeniu, że praca jest wykonywana z maksymalną intensywnością, charakterystyki czasowe wykorzystania źródeł energetycznych (→ Molekularne podstawy skurczu mięśnia) mają postać jak na rys. 18, a ich zasób i moc przedstawia tabela.



Rys. 18. Udział różnych źródeł energii w zależności od czasu pracy mięśnia

Zdolność do wykonywania pracy i rozwijania mocy przez człowieka o masie 75 kg

Źródło energii	Praca włożona J	Praca uzyskana J	Moc włożona W	Moc uzyskana W	Sprawność mechaniczna %
Fosforylacja	31 500	4 750	4 070	610	15
Glikoliza beztlenowa	71 200	14 250	2 200	440	20
Procesy tlenowe	nieograniczona	nieograniczona	1 125	290	26

W tabeli u góry podano z kolei jak są wykorzystywane poszczególne źródła energii przez człowieka na przykładzie człowieka o masie 75 kg.

Charakterystyki te decydują o doborze intensywności i czasu trwania wysiłku w treningu parametrów podlegających wytrenowaniu. Wzrost naprężenia mięśni σ_t jest trenowany głównie przez przedstawicieli szybkościowych i szybkościowo-siłowych dyscyplin sportu. Wzrost σ_t powoduje zwiększenie siły i prędkości ruchu rozwijanych przez mięśnie bez wyraźnego zwiększenia masy mięśniowej, czyli bez istotnego zwiększania przekroju fizjologicznego mięśni. Wzrost naprężenia mięśni σ_t uzyskuje się w wyniku ćwiczeń polegających na maksymalnym obciążeniu mięśni przez kilka lub kilkanaście sekund, co zwiększa zasób źródeł wysokoenergetycznych w mięśniu — głównie fosfokreatyny i częściowo glikogenu (rys. 18). Są to ćwiczenia w rodzaju krótkich biegów, skoków, serii uderzeń.

zwiększenie przekroju mięśni

Przekrój fizjologiczny S_t mięśni zwiększa się przede wszystkim przez wzrost plazmy w komórkach mięśniowych (tzw. mioplazmy), przy nie zmienionej liczbie włókien mięśniowych. Zwiększa się ona wówczas, gdy rośnie w niej niezbędny zapas substancji energetycznych — głównie glikogenu. Takie warunki występują wtedy, gdy praca jest wykonywana prawie z maksymalną intensywnością przez kilkanaście do kilkudziesięciu sekund (tabela z poprzedniej strony: Zasób i moc źródeł energii, rys. 18). Wzrost przekroju fizjologicznego jest czynnikiem pożądanym u kulturystów, zapaśników, miotaczy, a także ciężarowców, u których duża masa ciała nie przeszkadza, a na odwrót — pomaga w osiąganiu dobrych wyników sportowych.

zwiększenie liczby pobudzanych włókien

Umiejętność synchronicznego pobudzania maksymalnej liczby włókien mięśniowych (wzrost $\overline{u}_t$) jest głównie związana z działalnością układu nerwowego. Gwałtowne napinanie mięśni jest możliwe tylko przy wykorzystaniu bezpośrednio ATP (kwas adenozynotrójfosforowy). Jednakże zapas ATP starcza zaledwie na 2–3 maksymalne napięcia mięśni (rys. 18) i większa liczba powtórzeń (np. 3–7) musi odbyć się na koszt fosfokreatyny wchodzącej w reakcję odtwarzania ATP z ADP (kwas adenozynodwufosforowy) (CP na rys. 18). Metody treningu tej właściwości układu mięśniowego i nerwowego polegają na dynamicznym wymuszaniu pobudzenia dużej liczby włókien mięśniowych. Dzieje się to np. podczas zeskoków, gdy mięśnie muszą gwałtownie hamować rozpędzone ciało. Duża prędkość skracania mięśni zachodzi również wówczas, gdy przemieszcza się niewielkie masy — np. piłkę, oszczep, rakietę tenisową lub wykonuje ćwiczenia cykliczne bez dodatkowego obciążenia, np. bieg. W tym drugim wypadku prędkość skracania mięśni decyduje o dużej amplitudzie częstości ruchów wykonywanych przy dużej amplitudzie ruchów w stawach. Maksymalną częstość ruchów cyklicznych uzyskuje się po ok. 4–6 s, co jest głównie związane z wykorzystaniem fosfokreatyny jako źródła energii.

Ponadto trening mięśni jest stosowany w procesie rehabilitacji inwalidów. Zasady tego treningu są zbliżone do zasad treningu sportowego.

Problemy wymagające dalszych badań

właściwości a budowa mięśni

Wiele zagadnień dotyczących mięśni zostało jeszcze nie zbadanych, dlatego na pewno pozostaną one nadal w centrum zainteresowania fizjologów, biomechaników, inżynierów-bioników, biochemików i biofizyków. Dalszych badań wymaga m.in. ustalenie zależności właściwości mięśni od ich budowy (szczególnie mięśni gładkich). Nie jest także całkowicie wyjaśniony związek między zjawiskami chemicznymi, mechanicznymi i elektrycznymi zachodzącymi podczas pracy mięśnia. Jak dotąd w modelowaniu matematycznym i przy symulacji własności mięśni w różnych odpowiednikach technicznych nie udało się uwzględnić ich właściwości biochemicznych. Dokładniejsze zbadanie miopotencjałów rejestrowanych na powierzchni ciała lub za pomocą elektrod wkłuwanych i implantowanych powinno wyjaśnić udział włókien i jednostek motorycznych w przebiegu globalnego elektromiogramu. Bliższe poznanie mechanizmu zmian miopotencjału pozwoli na lepsze wykorzystanie elektromiogramów do celów diagnostyki i do sterowania urządzeń technicznych, np. bioprotez.

zmiany miopotencjału

Interesująca i mało zbadana jest kwestia męczenia się mięśnia podczas kolejnych skurczów. Wiadomo w tej chwili, że przy odpowiednim stosunku czasów skurczów do czasów odpoczynku można znacznie dłużej utrzymywać mięsień w stanie zdolności do pracy.

współdziałanie mięśni

Kolejnym zagadnieniem, które wymaga dalszych badań jest współdziałanie mięśni podczas wykonywania ruchów w stawach. W każdym ruchu bierze udział wiele mięśni, które z różnym wkładem sił uczestniczą w pokonywaniu zewnętrznego obciążenia kończyny. Dotąd podano kilka hipotez mechanizmu współdziałania mięśni obsługujących np. staw łokciowy lub promieniowo-nadgarstkowy w warunkach statycznych. Zbyt mało jest natomiast zbadane współdziałanie mięśni w warunkach dynamicznych, szczególnie w stawach o większej liczbie ruchów, jak np. ramiennym lub biodrowym. Poznanie mechanizmu współdziałania mięśni pozwoli na budowanie zaawansowanych manipulatorów antropomorficznych o lepszej sprawności energetycznej, dokładniejszym pozycjonowaniu i krótszym czasie działania przy wykonywaniu określonych zadań.

J. R. Bendall *Muscles, Molecules and Movement, An essay in the contribution of Muscles*, London 1970; E.V. Hill *First and Last Experiments in Muscle Mechanics*, London 1970; A. Morecki *Manipulatory bioniczne*, Warszawa 1976; A. Morecki i in. *Bionika ruchu*, Warszawa 1971; A. Morecki i in. *Badanie własności mechanicznych mięśni. Wykłady z biofizyki*, Łódź 1975; A. Morecki i in. *Cybernetyczne systemy ruchu kończyn zwierząt i robotów*, Warszawa 1979; D. R. Wilkie *Mięsień*, Warszawa 1974.

Biocybernetyka

Ryszard Gawroński

Biocybernetyka jako metoda badań procesów w złożonych układach biologicznych

Dzięki powstałym około dwudziestu lat temu nowym metodom badań złożonych układów biologicznych, stosunkowo szybko zmieniły się poglądy na podstawowe zjawiska zachodzące w organizmach żywych i obecnie trudno sobie wyobrazić nowoczesne badania biologiczne bez udziału specjalistów w zakresie teorii sterowania i przetwarzania informacji oraz ich metod. Już z definicji, jaką podał w 1948 r. Norbert Wiener w swojej klasycznej pracy *Cybernetics, or Control and Communication in the Animal and the Machine* ('Cybernetyka, komunikacja i sterowanie w zwierzętach i maszynach'), wynika, że cybernetyka w równym stopniu może być stosowana w badaniach biologicznych, jak i w technicznych. Biologowie z dość dużą rezerwą przyjmowali tak rewolucyjne zmiany w metodach badań przyrodniczych. Większość znanych naukowców pracujących w dziedzinie biocybernetyki — to technicy, oprócz nich raczej lekarze (np. psychiatrzy W. Mc Culloch lub W. Ross Ashby) niż biologowie. Obecnie na świecie biocybernetyką zajmują się setki instytutów naukowych lub laboratoriów w szkołach wyższych, a jej wpływ, mniej lub więcej uświadomiony, obserwuje się w większości prac biologicznych. Nie ogranicza się on, jak niektórzy sądzą, do wykrycia pętli sprzężeń zwrotnych lub wyrysowania tzw. schematów blokowych, tzn. rysunków, w których poszczególne elementy (prostokąty, koła) oznaczają etapy procesu lub podzespoły układu, a linie łączące te elementy obrazują wzajemne wpływy podzespołów. Do tego celu nie trzeba by powoływać nowej gałęzi nauki ani tworzyć nowych zespołów badawczych, wystarczyłyby zwyczaje lub wyniki uzyskane w naukach technicznych. Warto tu podkreślić, że potrzeba nowego podejścia wynika przede wszystkim z trudności analizy złożonych procesów biologicznych metodami klasycznymi, które mają w zasadzie charakter opisowy i nie wykorzystują w dostatecznym stopniu nowych osiągnięć innych działów wiedzy (oprócz techniki pomiarowej).

cybernetyczne metody w biologii

Cybernetyczne metody w biologii polegają na:

możliwie precyzyjnym opisie jakościowym i ilościowym wszystkich mierzalnych wielkości, które występują w procesie;

ustaleniu bądź założeniu (w postaci odpowiednich hipotez) związków między interesującymi nas wielkościami;

wykorzystaniu metod matematycznych do opisu badanych zjawisk i formowaniu tzw. modelu matematycznego procesu;

badaniu procesu nie tylko w stanie ustalonym, ale również badaniu jego dynamiki, tzn. zmian zachodzących w czasie;

wykorzystaniu teorii sterowania i powstałej niedawno teorii systemów do analizy procesów w złożonych układach, a zwłaszcza w układach z pętlami sprzężeń zwrotnych, w układach adaptacyjnych i optymalnych lub układach hierarchicznych;

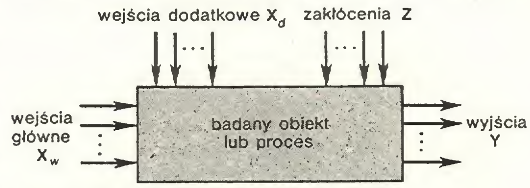
wykorzystaniu teorii informacji i metod przetwarzania informacji do oceny różnych procesów, w których dawniej uwzględniano głównie aspekty energetyczne;

wykorzystaniu metod technicznego modelowania do weryfikacji hipotez i analizy przebiegów procesów biologicznych, zwłaszcza wówczas, gdy analiza matematyczna jest zbyt trudna, a wyniki skomplikowane i nieczytelne.

Pierwsze trzy cechy stosowania metod cybernetycznych w biologii wynikają ze związków biocybernetyki z matematyką, która w coraz większym stopniu staje się podstawowym narzędziem badań (jak to wcześniej nastąpiło w fizyce i w różnych dziedzinach techniki).

metoda czarnej skrzynki

Pierwszy etap analizy układu lub procesu biologicznego polega zwykle na zastosowaniu tzw. metody czarnej skrzynki (rys. 1). Przypuśćmy, że o wewnętrznej strukturze badanego obiektu nie wiemy nic lub tak mało, że symbolicznie możemy go przedstawić np. w postaci prostokąta o nieznanej stukturze wewnętrznej. W tym wypadku wygodnie będzie posłużyć



Rys. 1. Metoda „czarnej skrzynki"

układ względnie odosobniony

się pojęciem układu względnie odosobnionego, wprowadzonym przez niedawno zmarłego prof. Henryka Greniewskiego.
W odróżnieniu od stosowanego w fizyce, a także w cybernetyce, pojęcia układu całkowicie odosobnionego, w układzie względnie odosobnionym istnieją wyróżnione punkty powierzchni brzegowej, przez które możliwy jest wpływ otoczenia na procesy wewnętrzne układu. Ponadto wyróżnia się inne punkty, przez które następuje oddziaływanie układu na otoczenie. Pierwszego rodzaju punkty nazywamy wejściami układu, drugiego zaś — jego wyjściami. W układach biologicznych ustalenie wejść i wyjść napotyka zwykle duże trudności i wymaga wstępnej znajomości badanego procesu.

pszczoła jako układ cybernetyczny

Przypuśćmy np., że pragniemy opisać zachowanie się pszczoły w czasie jej pierwszego lotu od ula do miejsca zbierania pokarmu. Pszczoła korzysta z informacji zewnętrznych i z pewnych danych wewnętrznych (pamięci), wynikających z dotychczasowych doświadczeń. Jako informacje zewnętrzne należy traktować dane o zachowaniu się (parametry ruchu) innych zbieraczek przylatujących do ula, a zwłaszcza tzw. taniec pszczół po przylocie w obszar ula. Liczba i zakres zmian tych parametrów mogą być dość duże i w naszym przykładzie nie wiemy dokładnie, które parametry są istotne, a które można pominąć; dopiero seria badań prowadzi do wybrania istotnych sygnałów wejściowych. Ponadto pszczoła przypuszczalnie do ustalenia kierunku lotu uwzględnia dane o zapachu pokarmu zebranego przez inne zbieraczki, o położeniu Słońca, kierunku wiatru itp. Na pszczołę oddziałują również pewne czynniki zakłócające, np. zjawienie się owadów będących jej naturalnymi wrogami, ruchomych obiektów, które trzeba ominąć itp. Za wyjścia można przyjąć zmieniony kierunek lotu i ewentualnie jego średnią szybkość. Jak widać, można rozróżnić trzy rodzaje wejść (będziemy je oznaczać czcionką półgrubą — jak wektory wielowymiarowe — rozumiejąc np. przez $\mathbf{X}_w$ zbiór wszystkich wejść głównych):
— wejścia główne $\mathbf{X}_w$, mające zasadniczy wpływ na przebieg zjawiska (w eksperymencie biologicznym są to często wejścia sterowane przez eksperymentatora);
— wejścia pomocnicze $\mathbf{X}_d$, które można zmierzyć i należy uwzględnić przy ustalaniu warunków pomiaru;
— wejścia zakłócające $\mathbf{Z}$, których na ogół nie można zmierzyć, ale można niekiedy ustalić pewne jego parametry statystyczne, np. średnią częstość zjawiania się danego zjawiska.

Nietrudno zauważyć, że podane tu rozumowanie zawsze charakteryzowało badania fizyczne oraz biologiczne i nie widać tu jeszcze potrzeby odwoływania się do cybernetyki. Ale współczesny eksperymentator zaczyna stawiać pytania ilościowe:

pytania eksperymentatora

Jakie są związki między sygnałami wejściowymi X_w i wyjściowymi Y?

Jaki jest wpływ na te związki wejść dodatkowych X_d?

Jaki jest wpływ stanu obiektu (dotychczasowy rozwój i dojrzałość osobnicza, przeprowadzone doświadczenia i in.) na związki między X_w a Y?

Jaki jest wpływ zakłóceń Z na wyjścia obiektu? Wymienione wyżej zależności można w ogólnej postaci przedstawić za pomocą pewnej, na razie nieznanej funkcji F:

$$Y = F(X_w, X_d, Z, s),$$

gdzie przez wektor s oznaczono zespół parametrów charakteryzujących stan obiektu.

Celem badań jest oczywiście ustalenie postaci funkcji F, a ściślej — zespołu funkcji, ponieważ F jest również wektorem o liczbie k składowych równej liczbie wyjść Y_i, $i = 1, 2, ..., k$. Powstają natychmiast zasadnicze pytania:

1. Czy możliwe jest ustalenie postaci funkcji F, czyli ilościowego opisu obiektu przedstawionego w postaci czarnej skrzynki, a więc czy istnieje ciąg pomiarów wielkości X_w, X_d, i Y, który umożliwi taki opis z żądaną dokładnością?

2. Czy istnieje tylko jedna postać funkcji F, tzn. czy istnieje tylko jeden sposób opisu obiektu?

3. Jak należy zaplanować eksperyment, by ewentualne ustalenie opisu obiektu (postaci funkcji F) nastąpiło w możliwie najkrótszym czasie i możliwie najmniejszym kosztem?

identyfikacja obiektu

Sformułowane w tych trzech punktach zadanie nazywa się w cybernetyce identyfikacją obiektu i przy jego rozwiązaniu korzystamy często — ze względu na występowanie zakłóceń Z — z bardzo subtelnych metod matematycznych, najczęściej statystycznych.

W każdym razie wiadomo, że na drugie pytanie odpowiedź jest negatywna i bez dodatkowych ograniczeń nie można jednoznacznie opisać obiektu. Można np. żądać, aby funkcja F miała postać szeregu potęgowego o minimalnej liczbie składników. Identyfikacja polega wtedy na określeniu liczby i wartości współczynników przy kolejnych potęgach zmiennych wejściowych. Procedura obliczenia tego typu współczynników w bardziej złożonych wypadkach wymaga stosowania komputera.

model układu oddechowego

W odróżnieniu od wyżej podanego przykładu coraz częściej spotykamy się z taką sytuacją, że posiadamy pewne wstępne jakościowe i ilościowe dane o strukturze wewnętrznej badanego obiektu i możemy je wykorzystać przy ustalaniu jego opisu. Konstruując np. model układu oddechowego, należy uwzględnić nie tylko wiedzę dotyczącą budowy i własności płuc i układu mięśni klatki piersiowej, ale i dane o ośrodku nerwowym sterującym procesem oddychania. Wiadomo, że w rdzeniu przedłużonym, tworzącym najniższą funkcjonalnie część mózgowia, znajdują się ośrodki wdechu i wydechu, których praca zależy od zawartości CO_2 we krwi. Uwzględniając tego typu dane, można narysować schemat blokowy obrazujący wewnętrzną strukturę obiektu oraz niektóre własności jego elementów (zob. rys. 10). W takiej sytuacji mówimy w przenośni, że mamy do czynienia nie z czarną, lecz z szarą skrzynką, ponieważ posiadamy pewne informacje o jej wnętrzu.

szara skrzynka

Rola sprzężenia zwrotnego

W pierwszym etapie rozwoju biocybernetyki podkreślano, że istotną cechą metod biocybernetycznych jest wyodrębnienie pętli sprzężeń zwrotnych. Sprzężenie zwrotne jest jednym z najważniejszych pojęć cybernetyki, umożliwia wyjaśnienie szeregu zjawisk występujących w przyrodzie. Sprzężenie zwrotne jest obecnie wykorzystywane powszechnie w urządzeniach technicznych, zwłaszcza w złożonych systemach sterowania. Najbardziej ogólnie można stwierdzić, że ze sprzężeniem zwrotnym mamy do czynienia, gdy jedna lub kilka wielkości wyjściowych z rozpatrywanego układu zostaje przekazana z powrotem na wybrane wejścia tego układu. Żartobliwie powołujemy się tu niekiedy na analogię do węża, który zaczyna zjadać własny ogon — ale jest to analogia jedynie strukturalna, bo oczywiście zjawiska fizyczne, jakie występują w przypadku sprzężenia zwrotnego, są innego rodzaju.

sprzężenie zwrotne

Sprzężenie zwrotne dzielimy na ujemne (rys. 2b) i dodatnie (rys. 3). W wypadku ujemnego sprzężenia zwrotnego sygnał wyjściowy y jest przykładany na wejście z przeciwnym znakiem, przeciwdziała jakby zmianom zachodzącym w układzie. Jeżeli natomiast mamy do czynienia z dodatnim sprzężeniem zwrotnym, sygnał wyjściowy dodawany jest do odpowiednich sygnałów wejściowych, co powoduje często podtrzymanie bądź narastanie procesów.

ujemne i dodatnie sprzężenie zwrotne

Ujemne sprzężenie zwrotne wyjaśnia dobrze stabilizację różnych wielkości fizycznych i biochemicznych występujących w układzie biologicznym. Tak np. wzrost temperatury otoczenia hamuje odpływ ciepła wytwarzanego stale przez organizm człowieka. Jeden ze sposobów kompensacji tego wpływu polega na poceniu się (a np. u psów — na szybszym oddechu), które ułatwia odpływ ciepła i stabilizuje temperaturę ciała.



element sumujący sygnały

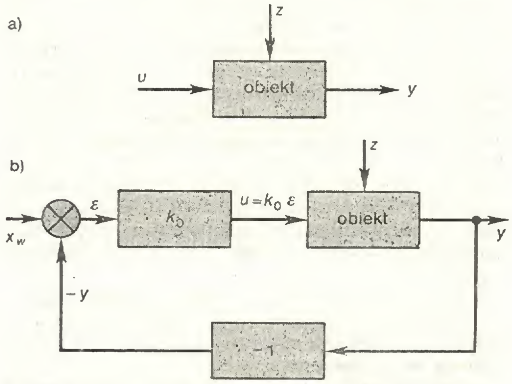


element zmieniający znak



wzmacniacz o wzmocnieniu k_0

Na rys. 2a przedstawiono dowolny obiekt, którego jedna z wielkości wyjściowych y powinna być stabilizowana bądź zmieniana zgodnie z pewną regułą. Na obiekt można wpływać za pomocą pewnego sygnału sterującego u. Ponadto na obiekt oddziałują zakłócenia z zwykle o charakterze przypadkowym i trudnym do bezpośredniego pomiaru. Schemat na rys. 2b ilustruje rolę ujemnego sprzężenia zwrotnego, zaznaczono na nim: x_w — wartość zadaną wielkości stabilizowanej (wartość sterującą); y — stabilizowaną wielkość wyjściową, której wartość powinna być możliwie zbliżona do wartości x_w; z — zakłócenia; $\varepsilon = x_w - y$, tzw. sygnał błędu, charakteryzujący odchylenie wielkości wartości sterującej (zadanej) od wyjściowej. Przypuśćmy, że zachowanie się obiektu (rys. 2a) można opisać prostym równaniem:



Rys. 2. Działanie pętli ujemnego sprzężenia zwrotnego: a) obiekt sterowany sygnałem u, b) tenże obiekt objęty pętlą

$$y = k_u u + k_z z, \quad (2)$$

gdzie u — sygnał sterujący obiektem, $k_u = \Delta y/\Delta u$ — wzmocnienie w torze sygnału sterującego, $k_z = \Delta y/\Delta z$ — wzmocnienie w torze zakłóceń, Δu i Δz przyrosty wielkości u i z wywołane w obiekcie dla wyznaczenia współczynników k_u i k_z. Rozpatrzmy zachowanie się takiego obiektu po objęciu go pętlą sprzężenia zwrot-

nego. Ponadto przypuśćmy, że ustawiono dodatkowy wzmacniacz umożliwiający dość duże wzmocnienie sygnału błędu:

$$u = k_0\varepsilon = k_0(x_w - y). \tag{3}$$

pętla ujemnego sprzężenia zwrotnego

Oprócz obiektu i wzmacniacza w skład pętli sprzężenia zwrotnego wchodzi komparator, sprawdzający zgodność sygnału wyjściowego z sygnałem wejściowym i przekazujący do obiektu sygnał sterujący u, oraz element zmieniający znak sygnału wyjściowego y. Oczywiście do przekazania informacji potrzebna jest zawsze jakaś (na ogół niewielka) ilość energii, będącej nośnikiem informacji. Ilość tej energii określają pewne prawa badane w teorii informacji.

Na podstawie rysunku 2 można napisać następującą prostą zależność:

$$y = k_0 k_u (x_w - y) + k_z z. \tag{4}$$

Rozwiązując względem y otrzymamy:

$$y = \frac{k_0 k_u x_w + k_z z}{k_0 k_u + 1}. \tag{5}$$

Analizując wzór (5) nietrudno zauważyć, że w miarę wzrostu iloczynu $k_0 k_u$ (czyli w miarę wzrostu wzmocnienia w torze sygnału błędu i torze sygnału sterującego) maleje rola członu $k_z z$, a więc maleje wpływ zakłóceń na działanie układu. Jeśli $k_0 k_u \gg 1$, wzór (5) można uprościć:

$$y = x_w + \frac{k_z}{k_0 k_u} z. \tag{6}$$

Widać tu jeszcze wyraźniej, że dzięki pętli sprzężenia zwrotnego wpływ zakłóceń został zmniejszony proporcjonalnie do $k_0 k_u$.

układ śledzący

Inny przykład umożliwi nam wyjaśnienie jeszcze jednego pojęcia używanego w teorii sterowania (a wcześniej — w podstawach automatyki), a mianowicie pojęcia układu śledzącego. Wartość wejściowa x_w nie zawsze musi być stała; może być zmieniana, gdy tego wymaga zadanie układu. Przypuśćmy dla uproszczenia, że jednokierunkowy ruch kończyny jest sterowany jednym mięśniem. Zatem wielkość skurczu musi być tak sterowana, aby przebieg wyjściowy „śledził" zmiany sygnału sterującego, a wynik sterowania powinien możliwie mało zależeć od zakłóceń. Jeżeli sterowanie skurczem odbywa się przez zmianę $x_w(t)$ w pętli sprzężenia zwrotnego, to zgodnie z wzorem (6) otrzymamy:

$$y(t) = x_w(t) + \frac{k_z}{k_0 k_u} z(t), \tag{7}$$

gdzie $y(t)$ może obecnie oznaczać np. położenie końca kończyny względem tułowia.

biologiczne procesy dynamiczne

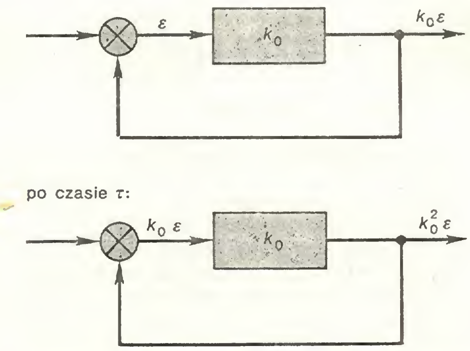
Znacznie bardziej skomplikowane zjawiska występują wówczas, gdy zachowanie się obiektu wymaga uwzględnienia procesów dynamicznych, tzn. procesów związanych ze zmianami energetycznymi w układzie, lub gdy obiekt może magazynować energię, np. w postaci energii kinetycznej czy potencjalnej. (W mechanice proces dynamiczny oznacza ruch cząstek pod wpływem sił, w teorii sterowania o dynamice mówimy wówczas, gdy w układzie zachodzą zmiany wielkości charakteryzujących różne rodzaje energii). Obiekt zawierający kilka wejść i kilka wyjść oraz elementy magazynujące energię może być opisany za pomocą układu równań różniczkowych, a badanie zjawisk występujących przy sterowaniu układami biologicznymi wymaga dobrej znajomości współczesnej teorii sterowania.

Stabilność układu

Jedno z zagadnień, które się rozwiązuje metodami teorii sterowania, polega na ustaleniu, czy opracowany przez biocybernetyka model matematyczny procesu jest modelem stabilnym. Stan układu lub modelu nazywamy niestabilnym, jeżeli — mówiąc w uproszczeniu — bardzo małe zaburzenia powodują powstanie w układzie procesów (np. wychyleń lub drgań), które wyprowadzają układ poza stany dopuszczalne w normalnej pracy układu. Ołówek postawiony pionowo na ostrzu nie znajduje się w stanie stabilnym, ponieważ najmniejsze nawet zaburzenie powoduje szybką zmianę jego położenia na poziome.

stan niestabilny

Można wykazać, że występowanie w układzie biologicznym lub technicznym dodatniego sprzężenia zwrotnego prowadzi do niestabilności. Ilustruje to rysunek 3, na którym przyjęto w celu uproszczenia, że



Rys. 3. Działanie pętli dodatniego sprzężenia zwrotnego

zamiast obiektu znajduje się prosty element opóźniający każdy przebieg o stałą wartość τ. Jeżeli w pętli sprzężenia zwrotnego wzmocnienie jest większe od jedności, to drobne zaburzenie np. na wejściu układu powoduje po czasie τ zjawienie się na wyjściu sygnału $k_0\varepsilon$, a to po kolejnych przejściach przez pętlę wywoła sygnały o wartości $k_0^2\varepsilon$, $k_0^3\varepsilon$ itd.; jeżeli przy tym $k > 1$, to sygnał będzie narastał aż do zjawienia się dodatkowych ograniczeń amplitudy procesu. Występujące w niektórych schorzeniach układu nerwowego silne drżenie kończyn czy napady padaczkowe można wyjaśnić zjawianiem się pętli dodatniego sprzężenia zwrotnego. Na ogół jednak organizm uruchamia w takich wypadkach dodatkowe pętle, tzw. hamujące, które wygaszają nadmierne oscylacje (np. przez zmniejszenie wzmocnienia). Mamy tu do czynienia z nowym zjawiskiem, które należy do tzw. zjawisk adaptacyjnych. Powstanie nowej sytuacji lub zmiany warunków pracy układu mogłyby spowodować niekorzystne dla organizmu procesy. Aby temu zapobiec, organizm uruchamia nowe procesy, np. zmiany wzmocnienia w poszczególnych pętlach, które kompensują zaburzenia i sprowadzają warunki pracy układu do warunków dopuszczalnych.

pętla hamująca

układy adaptacyjne

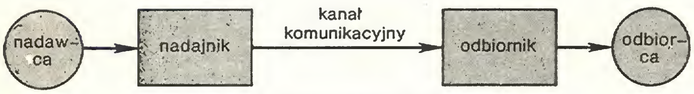
Układy adaptacyjne mają co najmniej dwie pętle sprzężenia zwrotnego. Jedna z nich np. stabilizuje proces w normalnych warunkach, druga — steruje akością procesu ub włącza się z chwilą przekroczenia typowych warunków pracy. Organizmy żywe zawierają tysiące pętli sprzężenia zwrotnego, a ich rolę można będzie stopniowo wyjaśniać prowadząc badania biocybernetyczne.

Zastosowanie teorii informacji

Chociaż nasza wiedza o procesach występujących w organizmach żywych, zwłaszcza w układzie nerwowym, znajduje się w początkowym stadium, rozwój teorii informacji przyniósł rewelacyjne zmiany w poglądach na te procesy i w metodach ich badania. W 1948 r. C. E. Shannon, badając systemy komunikacyjne, wprowadził miarę ilości informacji dla oceny zjawisk zachodzących między nadawcą a odbiorcą. Najprostszy system przekazywania wiadomości przedstawiono na rys. 4. Celem systemu jest przekazanie

sygnał — nośnik informacji

wiadomości od nadawcy do odbiorcy, czyli — inaczej mówiąc — u odbiorcy musi nastąpić jakiś proces, który będzie pewnym (na ogół niejednoznacznym) odwzorowaniem jakiegoś zjawiska zachodzącego u nadawcy. Aby to było możliwe, nadawca musi mieć do swej dyspozycji nadajnik (np. aparat telegraficzny, sygnalizator świetlny lub narząd mowy), który wysyła sygnał (akustyczny, elektryczny, świetlny itp.). Sygnał jest nośnikiem informacji. Oznacza to, że pewne parametry sygnału charakteryzują stan nadawcy. Parametry te muszą być przesłane kanałem komunikacyjnym i przyjęte w odbiorniku, a następnie przekazane odbiorcy jako dane o pewnym stanie nadawcy.



Rys. 4. Najprostszy system informacyjny

Przy konstruowaniu technicznych urządzeń komunikacyjnych, jak i przy badaniu wzajemnego przekazywania informacji przez środowisko i organizm żywy (a także wewnątrz organizmu) dążymy do wyodrębnienia źródła sygnałów i ich znaczenia dla badanego układu, charakteru i własności procesu fizycznego będącego nośnikiem informacji (np. rozchodzenia się fal akustycznych czy elektromagnetycznych, przepływu prądów elektrycznych lub procesów elektrochemicznych), ewentualnego wpływu zakłóceń oraz sposobu reagowania odbiorcy na ten proces.

Najprostsza sytuacja występuje wówczas, gdy u nadawcy możliwe są dwa stany, np. głodny lub nasycony. Pisklę ptaka w gnieździe sygnalizuje milczeniem stan nasycenia, a krzykiem lub piskiem stan głodu. Pierwszy stan oznaczamy np. 0, drugi 1. Ilość informacji zawartej w wiadomości o wyborze jednego z dwu możliwych i jednakowo prawdopodobnych stanów została nazwana 1 bitem.

ilość informacji

bit

Jeżeli u nadawcy możliwe są 4 stany, to wbrew pierwszemu wrażeniu liczba informacji zawartej w wiadomości o tym stanie jest równa 2 bity. Najpierw dzielimy 4 stany na dwie pary i ustalamy, o którą z tych dwóch par chodzi (1 bit), a następnie ustalamy stan w danej parze (1 bit). Podobnie, jeżeli mamy $s = 2^n$ stanów u nadawcy, to wiadomość o wybranym stanie (spośród s jednakowo prawdopodobnych) zawiera

$$\log_2 s = \log_2 2^n = n$$

bitów informacji. Wzór ten jest słuszny, gdy chodzi o stany jednakowo prawdopodobne. Najprostszym i naturalnym uogólnieniem uwzględniającym możliwość różnych prawdopodobieństw występowania poszczególnych stanów nadawcy jest następujący wzór wyprowadzony przez Shannona:

$$I = -\sum_{i=1}^{K} p(i)\log_2 p(i) \text{ bitów,} \quad (8)$$

gdzie $p(i)$ — prawdopodobieństwo zjawienia się stanu oznaczonego numerem i u nadawcy.

Nie wnikając głębiej w problemy teorii informacji, trzeba podkreślić, że do istotnych zagadnień należy wybór sposobu kodowania informacji, tzn. wybór tych parametrów w sygnale, których zmiana przenosi informację. Od sposobu kodowania może zależeć, czy mimo zakłóceń informacja zostanie przekazana prawidłowo odbiorcy. Drugie ważne zagadnienie polega na odpowiedniej selekcji tych informacji, tak aby najważniejsze dotarły do odbiorcy, a nieistotne zostały pominięte. Wiąże się to z tzw. zagadnieniem selekcji danych o otoczeniu, której dokonuje stale każdy organizm żywy za pomocą układu nerwowego. Przede wszystkim selekcjonowane zostają sygnały o niebezpieczeństwie, potem o pokarmie, w pewnych sytuacjach o partnerze seksualnym, a następnie np. o możliwości komfortowego spędzenia czasu itp.

kodowanie informacji

selekcja informacji

Modelowanie

znaczenie modelowania zjawisk biologicznych

Wobec ogromnej złożoności zjawisk w organizmach żywych opisy matematyczne nawet bardzo uproszczonych układów lub funkcji organizmu żywego są również wyjątkowo skomplikowane. Opisy te mają często postać układów równań różniczkowych lub różniczkowo-różnicowych (tzn. niektóre składniki tych równań odpowiadają różnym czasom), a ich rozwiązanie analityczne jest z reguły niemożliwe. Niekiedy korzysta się w takiej sytuacji z maszyn cyfrowych (komputerów), znacznie przyspieszających obliczenia, ale otrzymane wyniki dotyczą tylko konkretnych wypadków o ustalonych parametrach, przy tym często nie wystarczają do poznania całego bogactwa zjawisk, jakie mogą wystąpić w badanym obiekcie. Dlatego właśnie w biocybernetyce dużego znaczenia nabierają metody modelowania fizycznego zjawisk biologicznych (→ modelowanie procesów biologicznych).

Modelowanie różnorodnych zjawisk fizycznych, technicznych, a ostatnio nawet ekonomiczno-społecznych, jest często jedyną możliwością choćby częściowego poznania procesów w złożonym układzie (systemie). W tym celu wykorzystuje się (bądź specjalnie konstruuje) urządzenia techniczne opisywane w przybliżeniu tymi samymi równaniami co badany układ. Urządzenia te umożliwiają łatwą zmianę sygnałów wejściowych, struktury i wielkości parametrów poszczególnych elementów modelu oraz pomiar dowolnych procesów zarówno na wyjściu, jak i wewnątrz modelu.

modele techniczne

Należy wyraźnie podkreślić, że model w postaci urządzenia mechanicznego, elektrycznego, a ostatnio najczęściej elektronicznego, jest z reguły bardzo uproszczonym odwzorowaniem rzeczywistych procesów biologicznych i dlatego ogromnego znaczenia nabiera wybór tych własności i parametrów badanego zjawiska biologicznego, które mają być odwzorowane w modelu. W rozdziale dotyczącym układu nerwowego podane zostaną przykłady modelowania fizycznego z wykorzystaniem zarówno typowych maszyn matematycznych, jak i wyspecjalizowanych urządzeń technicznych.

modele procesów biologicznych

Biocybernetyka a układy regulacji

W układach biologicznych zwykle wydziela się kilka powiązanych hierarchicznie poziomów regulacji: poziom regulacji procesów wewnątrzkomórkowych, regulacja ustrojowa przemian metabolicznych w tkankach i podstawowych procesów wegetatywnych, regulacja zachowania się organizmu przez wyższe poziomy układu nerwowego.

Dotychczas udział biocybernetyki w analizie procesów pierwszego i drugiego poziomu regulacji był stosunkowo niewielki, jednakże próby tworzenia modeli matematycznych i programów komputerowych do badania poszczególnych układów regulacji są coraz częstsze. Na przykład przy modelowaniu procesu syntezy białka z uwzględnieniem roli kwasów nukleinowych i enzymów w reprodukcji określonych białek korzysta się ze schematów blokowych, chociaż jest to dopiero wstęp do badań cybernetycznych. Więcej jest prac polegających na modelowaniu takich procesów w organizmie jak sterowanie oddechem, regulacja obiegu krwi, stabilizacja poziomu temperatury, ilości wody i innych płynów w organizmie, regulacja hormonalna różnych przemian metabolicznych i wiele innych.

biologiczne układy regulacji

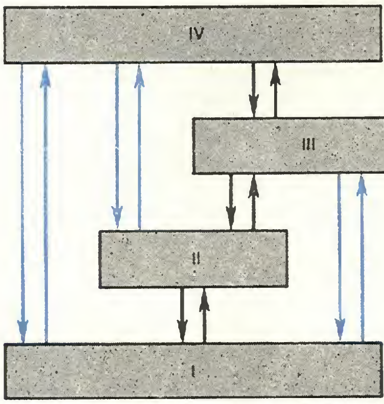
Wszystkie te przykładowo wymienione procesy są ze sobą w różnym stopniu powiązane i układy regulacji muszą w odpowiednim stopniu uwzględniać te powiązania. Odpowiedzialna za pracę i koordynację układów regulacji w organizmach żywych jest część

układ wegetatywny

układu nerwowego zwana układem wegetatywnym. Analizując liczne dane o układzie wegetatywnym pod kątem widzenia ogólnej struktury i charakteru procesów, można stwierdzić następujące jego własności:

struktura hierarchiczno-równoległa

1. Hierarchiczno-równoległa organizacja współzależności poszczególnych ośrodków i obwodów (rys. 5). Warto podkreślić, że wzajemne powiązania różnych obwodów i wpływy nadrzędne między nimi mogą być różne, a nawet mogą się zmieniać w zależności od warunków otoczenia. Tak np. człowiek nie ma (i nie ma potrzeby) bezpośredniego dowolnego wpływu ani na bicie serca lub skurcz naczyń krwionośnych, ani na pracę wielu ośrodków hormonalnych. Wpływ

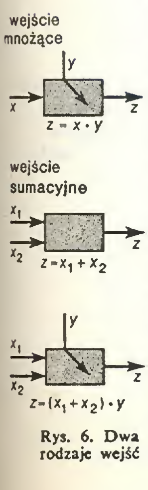


Rys. 5. Struktura hierarchiczno-równoległa. Strzałkami czarnymi oznaczono połączenia między sąsiednimi szczeblami hierarchii, niebieskimi — połączenia bezpośrednie łączące szczeble bardziej odległe, umożliwiające szybkie przekazywanie bodźców do tych szczebli

ten pojawia się w sposób pośredni, np. przez stany emocjonalne, lub występuje u niektórych specjalnie wytrenowanych osób. Z drugiej strony wiele procesów jest ściślej ze sobą powiązanych w porównaniu z innymi. Dotyczy to np. regulacji bicia serca, ciśnienia krwi, skurczu naczyń krwionośnych i w ogóle procesów w układzie krążenia. Układ sterowania biciem serca jest związany z regulacją ukrwienia różnych narządów, np. układu trawiennego lub mięśni. Ustalenie ogólnej struktury powiązań stanowi pierwszy etap badań modelowych.

liczne pętle sprzężenia zwrotnego

2. Występowanie licznych obejmujących się wzajemnie pętli sprzężenia zwrotnego. Badanie dynamiki takich pętli, zwłaszcza przy uwzględnieniu własności nieliniowych i inercyjnych (a w tym i opóźnienia), ogranicza się na razie do konkretnych wypadków, ale można wysunąć ogólne przypuszczenie, że dodatkowe pętle zmniejszają wpływ wahań parametrów i zakłóceń o różnych szybkościach zmian. Przykład takiego układu z kilkoma pętlami podany jest na rys. 10 i zostanie opisany nieco dalej.



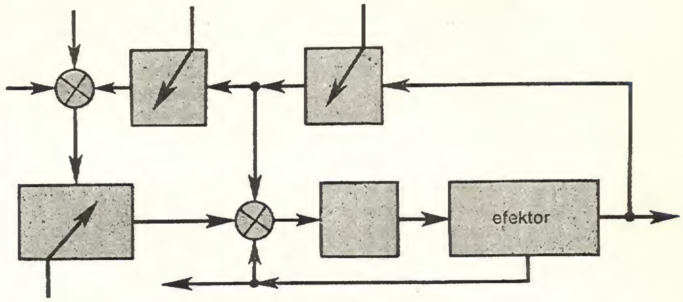
Rys. 6. Dwa rodzaje wejść

Układ nazywamy nieliniowym, jeżeli nie można do niego zastosować zasady superpozycji (reakcja na sumę dwóch bodźców nie jest równa sumie reakcji na bodźce składowe); układ nazywamy inercyjnym, jeżeli zawiera elementy gromadzące energię pod jakąkolwiek postacią (kinetyczną, potencjalną, chemiczną itp.).

3. Występowanie dwóch rodzajów wejść do poszczególnych pętli regulacji. Oprócz opisanych w poprzednim rozdziale wejść typu sumacyjnego w układach biologicznych występują bardzo często wejścia mnożące (rys. 6). Zmieniają one wzmocnienie w poszczególnych elementach pętli sprzężenia zwrotnego, co w zasadniczy sposób zmienia własności, a zwłaszcza dynamikę procesu regulacji. Przy omawianiu układu sterowania mięśniami (rys. 23) zostanie podany przykład wpływu zmiany wzmocnienia na szybkość skurczu mięśni. W pierwszej bowiem fazie skurczu zwiększa się czułość układu na przychodzące sygnały sterujące, wskutek czego proces rozpoczyna się prędzej niżby to nastąpiło przy stałym wzmocnieniu.

kilka sygnałów wejściowych

4. Występowanie kilku sygnałów wejściowych (rys. 7). Element wyjściowy, zwany często efektorem (mięsień lub gruczoł), otrzymuje sygnały sterujące, do-



Rys. 7. Hipotetyczny schemat typowej pętli sterowania w organizmie żywym z kilkoma wejściami (sumującymi i mnożącymi)

cierające do niego różnymi drogami. Wyższe szczeble hierarchicznego układu regulacji otrzymują informację nie tylko o wartości regulowanej wielkości, ale i o wartościach sygnałów w poszczególnych punktach układów sterowania, w tym również o wartości sygnałów błędu, stanu czujników itp.

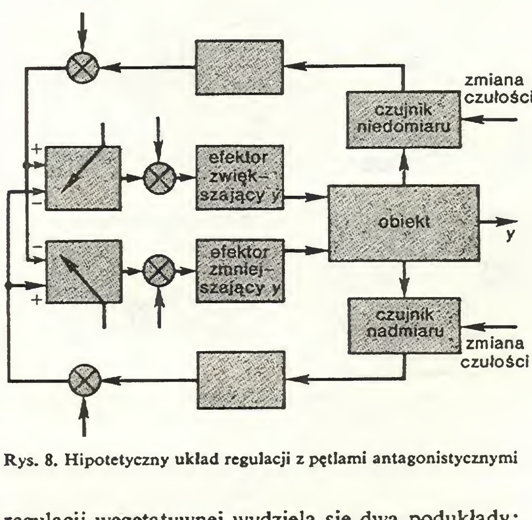
hierarchia czasowa

5. Hierarchia czasowa zjawisk. W układzie nerwowym, a zatem i w układach sterowania, występują procesy o różnych stałych czasowych. Na ogół można wydzielić dwa typy pętli: pętle szybkie i pętle wolno działające. Pętle szybkie są związane z receptorami o małej stałej czasowej i umożliwiają szybką reakcję na określony bodziec, natomiast pętle powolne, najczęściej o charakterze adaptacyjnym, umożliwiają dopasowanie procesu do nowych warunków otoczenia.

pętle szybkie, pętle powolne

ośrodki antagonistyczne

6. Występowanie ośrodków antagonistycznych. Sygnały w ośrodkach biologicznych mają z reguły stały znak i dlatego w typowych układach regulacji procesów biologicznych występują dwie pętle (rys. 8). Jedna z nich służy do kompensacji niedomiaru lub zbyt małej wartości jakiejś wielkości, druga zaś umożliwia zmniejszenie nadmiaru lub zbyt dużej wartości regulowanej wielkości. Zresztą w całym układzie



Rys. 8. Hipotetyczny układ regulacji z pętlami antagonistycznymi

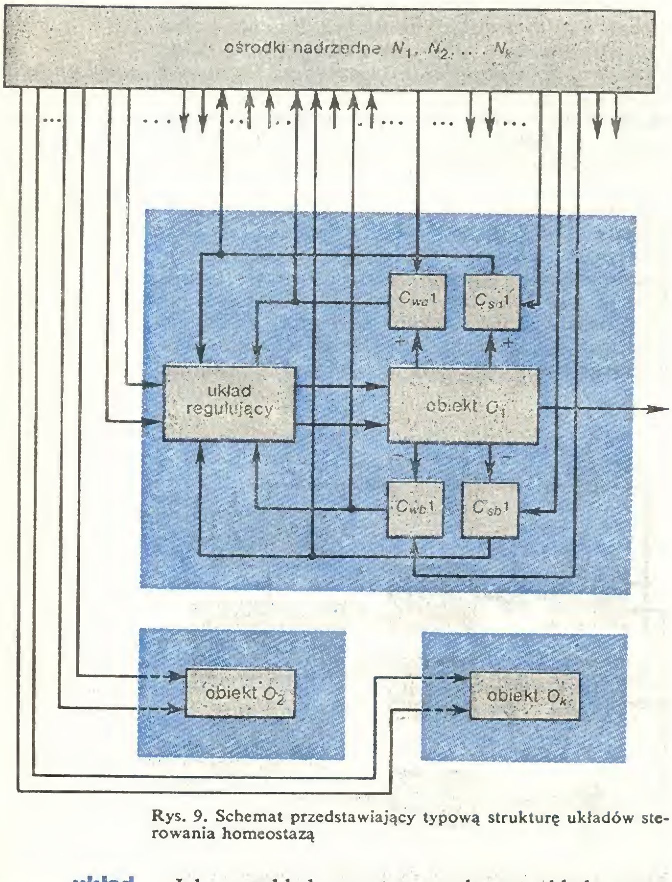
regulacji wegetatywnej wydziela się dwa podukłady: tzw. współczulny (sympatyczny) i przywspółczulny (parasympatyczny). Na ogół ośrodki układu współczulnego działają pobudzająco na różne funkcje organizmu, ośrodki układu przywspółczulnego działają na nie hamująco. Wprawdzie od tej zasady są wyjątki, ale zawsze są to działania antagonistyczne.

Należy podkreślić, że przytaczane tu schematy są bardzo uproszczone, a modele, z których korzystamy przy analizie zjawisk, zawierają wiele wejść i wyjść i pętli sprzężeń zwrotnych, co zostało jedynie zasygnalizowane dodatkowymi strzałkami na rys. 7, 8 itd.

Podsumowując powyższe uwagi na temat struktur układów sterowania organizmów żywych, można przedstawić ogólny schemat układu sterowania mającego zapewnić organizmowi takie warunki, w których by wszystkie procesy życiowe przebiegały w optymalny sposób. Na rys. 9 przedstawiono wycinek najniższych szczebli hierarchii takiego układu; nazwiemy go układem homeostazy. Poszczególne obiekty regulacji O_1, O_2, ... O_k itd. są oczywiście powiązane

układ homeostazy

zależnościami fizyczno-chemicznymi. Zgodnie z poprzednio podanymi zasadami czujniki nadmiaru (oznaczone indeksem a) i niedomiaru (oznaczone indeksem b) c_{wa1}, c_{sa1}, c_{wb1}, c_{sb1}, c_{wa2}, c_{sa2} itd. są źródłami sygnałów korekcyjnych dla odpowiednich ośrodków regulacji. Zarówno wzmocnienia czujników, jak i pobudliwości ośrodków mogą być zmieniane przez ośrodki nadrzędne. Ponadto przewidziano w układzie występowanie pętli regulacji o małej i dużej stałej czasowej, czyli pętle działające szybko (oznaczone indeksem s) i pętle działające wolno (oznaczone indeksem w). Ośrodki nadrzędne N_1, N_2, ..., N_k również tworzą pary antagonistyczne i sterują zarówno pobudzeniem poszczególnych obwodów, jak i ich wzmocnieniem.



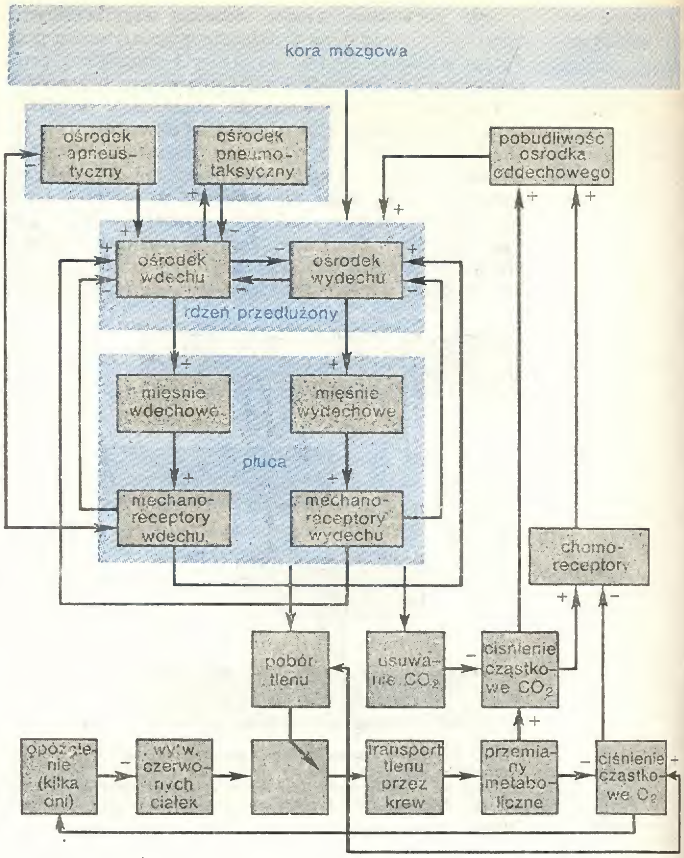
Rys. 9. Schemat przedstawiający typową strukturę układów sterowania homeostazą

układ sterowania oddechem

Jako przykład rozpatrzmy schemat układu sterowania oddechem. Chociaż badaniom tego układu poświęcono wiele prac zarówno fizjologicznych, jak i cybernetycznych, istnienie i rola wielu połączeń i własności tego układu nie są jeszcze wyjaśnione. Ogólna jego struktura została przedstawiona na rys. 10. Intensywność oddychania zależy od składu gazowego krwi, co można przedstawić wzorem Graya:

wzór Graya

$$v_r = K_0 + K_1 P_{CO_2} + K_2 \mathrm{pH} - K_3 P_{O_2}, \qquad (9)$$



Rys. 10. Schemat blokowy układu sterowania oddechem (wg A. Nechaya)

w którym v_r — intensywność oddychania, K_0, K_1, K_2, K_3 — współczynniki, P_{CO_2} — cząsteczkowe ciśnienie dwutlenku węgla we krwi, P_{O_2} — cząsteczkowe ciśnienie tlenu, pH — stężenie jonów wodorowych.

Odpowiednie chemoreceptory, znajdujące się zarówno w centralnym układzie nerwowym, jak i na jego części peryferyjnej, umożliwiają pomiar stężeń CO_2 i O_2. Maksimum czułości czujnika CO_2 przypada nieco powyżej cząstkowego ciśnienia normalnego dla CO_2 (tj. P_{CO_2} = 40 mm Hg), natomiast maksimum czułości na zmiany ciśnienia tlenu występuje przy ciśnieniu cząstkowym znacznie niższym niż normalne dla tlenu, a więc gdy P_{O_2} = 55 mm Hg. W układzie sterowania występują zatem dwie nadrzędne pętle sterujące intensywnością oddechu, z których jedna reaguje na wzrost CO_2, a druga na spadek O_2 we krwi.

mechanoreceptory wdechu i wydechu

Najniższy szczebel sterowania jest układem typowo samowzbudnym i o regulowanym wzmocnieniu. W płucach znajdują się dwa typy mechanoreceptorów: wdechu i wydechu. Pobudzenie ich pobudza odpowiednio ośrodki wdechu i wydechu w rdzeniu przedłużonym i łącznie z mięśniami wdechu i wydechu tworzy zasadniczą część układu wydechowego. Natomiast ośrodki znajdujące się nieco wyżej, tzw. pneumotaksyczny i apneustyczny, wpływają zarówno na częstotliwość, jak i na amplitudę procesów przez dwie pętle ujemnego sprzężenia zwrotnego. Pierwszą pętlę stanowi połączenie ośrodka pneumotaksycznego z ośrodkiem wdechu. Ośrodek pneumotaksyczny hamuje ośrodek wdechu z siłą zależną od pobudzenia, jakie od niego otrzymuje. Druga pętla jest dłuższa i obejmuje ośrodek wdechu, mięśnie wdechowe, mechanoreceptory i ośrodek apneustyczny. Ośrodek apneustyczny pobudza ośrodek wdechu, a receptory sygnalizujące wdech hamują ośrodek apneustyczny.

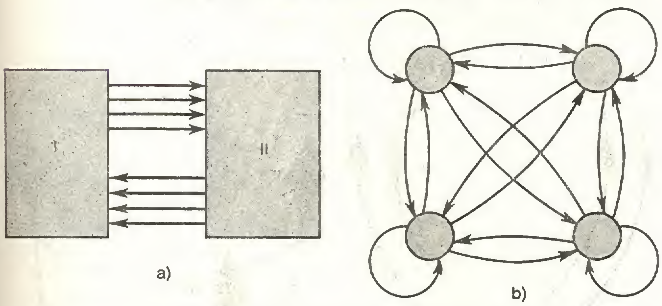
badanie układu oddechowego

W badaniach układu oddechowego przeprowadzonych na zwierzętach stwierdzono oscylacyjne procesy w ośrodkach wdechu i wydechu nawet po ich izolowaniu od mechanoreceptorów i od ośrodków nadrzędnych. Przypuszcza się, że w ośrodkach wdechu i wydechu istnieją sieci komórek złożone z wielu zamkniętych pętli. Komórki w tych pętlach pobudzają się kolejno, a potencjały ich pobudzenia rozprzestrzeniają się dość wolno w pewnym obszarze, tak jak się to dzieje w mięśniu sercowym. Potencjał progowy tych komórek narasta w funkcji czasu, a pobudzenie dochodzi do nasycenia. Wskutek ujemnego sprzężenia zwrotnego między ośrodkami pobudzenie jednego ośrodka hamuje pracę drugiego i odwrotnie, co prowadzi do rytmicznej pracy układu.

Występowanie w organizmach żywych wielu sprzężeń zwrotnych nasuwa pewne zasadnicze pytanie. W jaki sposób zapewniona jest stabilność tak złożonego układu, w którym istnieje bardzo duże prawdopodobieństwo zjawienia się dodatniego sprzężenia zwrotnego pod wpływem zmieniających się warunków zewnętrznych i wewnętrznych? W. R. Ashby wykazał doświadczalnie możliwość istnienia takiej elastycznej struktury sterowania, która zapewnia stabilność złożonego układu niezależnie od zakłóceń. Skonstruował on urządzenie nazwane homeostatem. W sposób najbardziej ogólny można go nazwać układem znajdującym stan równowagi we wszelkich stwarzanych mu warunkach. Ogólna struktura homeostatu przedstawiona została na rys. 11a. Homeostat składa się z dwóch zasadniczych części. Część pierwsza (I) zawiera cztery układy połączone ze sobą w sposób pokazany na rys. 11b. Jak widać, połączenia te tworzą sieć licznych pętli sprzężeń zwrotnych. Między innymi uwzględniono tu również sprzężenia zwrotne działające na własny element. Druga część (II) homeostatu składa się z szeregu przełączników krokowych (działających wtedy, gdy któryś z parametrów osiągnie graniczną wartość dopuszczalną) powodujących zmianę parametrów części pierwszej.

homeostat

Istota pracy układu polega na tym, że zmienne opisujące przebiegi w części I mogą przybierać wartości tylko w pewnym dopuszczalnym przedziale. Osiągnięcie granic przedziału przez którąkolwiek ze zmiennych (np. prądu w elemencie części I) powoduje wysłanie odpowiedniego sygnału do części II i zadziałanie przełączników krokowych, zmieniających parametry części I. Jeśli część I osiągnie stan równowagi wewnątrz obszaru dopuszczalnych wartości, to przełączniki części II będą nieruchome — i całość znajdzie się w stanie równowagi. Bardziej prawdopodobna jest jednak taka sytuacja, że któraś ze zmiennych w części I osiągnie wartość graniczną — wtedy zadziała część II, zmieniająca parametry części I, i rozpocznie się nowy cykl pracy zmian przebiegów w części I, który zakończy się bądź osiągnięciem stanu równowagi, bądź osiągnieciem przez którąś ze zmiennych wartości granicznej, co poprowadzi do następnego cyklu pracy. Cykle będą się powtarzać dopóty, dopóki układ nie znajdzie się w stanie równowagi.

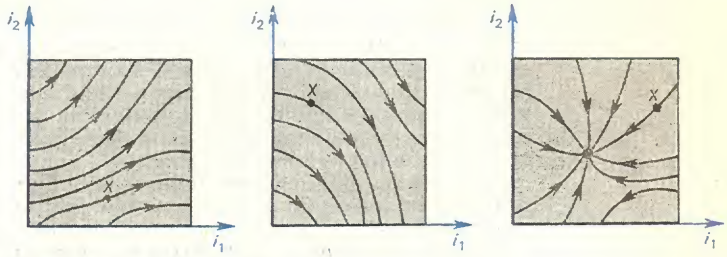


Rys. 11. Homeostat: a) struktura ogólna, b) pętle sprzężenia zwrotnego między układami homeostatu

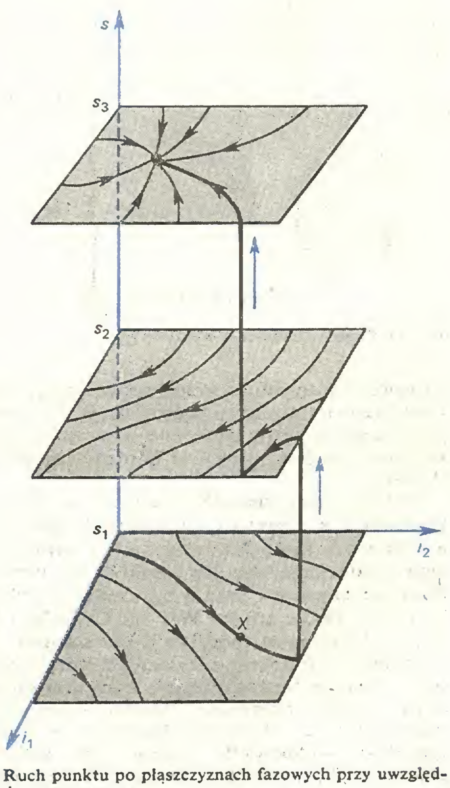
Przypuśćmy na chwilę, że w układzie zmieniają się tylko dwie zmienne, czyli prądy i_1 i i_2. Na rys. 12 podano kilka przykładów tzw. płaszczyzn fazowych (w technice i matematyce stosuje się termin obrazy fazowe), tzn. zestawu krzywych obrazujących wzajemne powiązania wielkości i_1 i i_2. Strzałki oznaczają kierunek zmian zachodzących w miarę upływu czasu. Chwilową wartość prądu charakteryzuje pewien punkt x poruszający się po krzywych. Ponadto na rysunkach podano linie ograniczające brzegi dopuszczalnych zmian wartości zmiennych. Gdy punkt x, poruszając się w miarę upływu czasu, osiągnie brzeg obszaru, nastąpi opisane wyżej zadziałanie układu krokowego i zmiana płaszczyzny fazowej na inną. Uwidacznia to rys. 13. Wprowadzono tu trzecią zmienną s, ilustrującą stan przełączników. Zmienna ta może przybierać dyskretne wartości, tzn. że zmienia się skokowo od wartości s_1 do s_2, dalej do s_3 itd. Na rys. 13 podano przykładowy ruch punktu x obrazujący zmiany prądów i_1 i i_2 w miarę upływu czasu i poszukiwania stabilnej płaszczyzny fazowej, czyli takiej, przy której punkt równowagi (węzeł trwały) znajduje się wewnątrz obszaru. Jak widać z rysunku, płaszczyzną tą okazała się płaszczyzna przy wartości s_3. W praktyce jednak poszukiwanie punktu stabilnego może trwać bardzo długo. Interesujące własności

płaszczyzny fazowe

punkt równowagi



Rys. 12. Przykłady płaszczyzn fazowych



Rys. 13. Ruch punktu po płaszczyznach fazowych przy uwzględnieniu zmian parametru s

opisanego tu urządzenia ujawniają się, gdy eksperymentator wprowadza do układu nowe zmiany, np. zmienia kierunki lub wartości prądów w elementach części I. Okazuje się, że nawet bardzo duże i złośliwe zmiany parametrów homeostatu (tzn. takie, które powodują powstanie pętli dodatniego sprzężenia zwrotnego) nie mogą uniemożliwić znalezienia stanu stabilnego, wydłużają tylko czas jego poszukiwania.

sprzężenie zwrotne w homeostacie

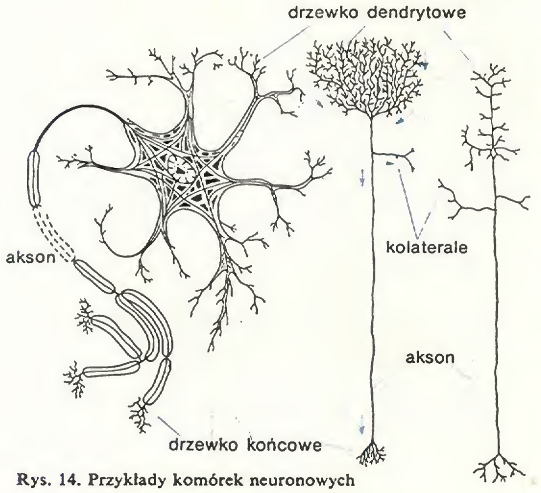
Warto zwrócić uwagę na to, że homeostat posiada dwa typy sprzężeń zwrotnych. Jedne z nich (przedstawione na rys. 11) są pętlami lokalnymi realizującymi oddziaływania między poszczególnymi zmiennymi, a sprzężenia zwrotne drugiego typu działają nadrzędnie i włączając się okresowo zmieniają związki między elementami powiązanymi przez pętle lokalne. Pętla nadrzędna działa dopóty, dopóki układ nie osiągnie zamierzonego celu, jakim jest w tym wypadku osiągnięcie stanu równowagi stabilnej.

Takie właśnie hierarchiczne struktury występują w organizmach żywych, a homeostat jest jednym z przykładów ich ogromnych możliwości. Przykłady innych złożonych struktur będą podane w następnym rozdziale.

Neurocybernetyka

modelowanie i analiza procesów układu nerwowego

Najbardziej rozwinięty dział biocybernetyki, neurocybernetyka, zajmuje się modelowaniem i analizą procesów w układzie nerwowym. Jej rozwój wiąże się oczywiście z rozwojem badań neurofizjologicznych, a zwłaszcza z rozwojem teorii działania pod-



Rys. 14. Przykłady komórek neuronowych

neuron

stawowego elementu systemu nerwowego, jakim jest komórka nerwowa, czyli neuron. Z rys. 14, na którym przedstawiono różne typy neuronów, widać, że nawet anatomicznie jest to zwykle bardzo skomplikowany element.

W latach trzydziestych panowała opinia, że dokładniejszy, a zwłaszcza matematyczny opis zarówno neuronu, jak i układów złożonych z większej liczby tego typu elementów jest praktycznie niemożliwy. Punktem zwrotnym prac w tym zakresie był opublikowany w 1943 r. artykuł W.S. Mc Cullocha i W. Pittsa, w którym na podstawie ówczesnych danych o własnościach neuronu stworzono teoretyczne podstawy analizy sieci złożonych z bardzo uproszczonych modeli neuronów. Dzięki odpowiedniemu sformułowaniu własności elementów i charakteru sygnałów wejściowych można było wykorzystać aparat logiki matematycznej do analizy ogolnych własności tego typu sieci.

neuron formalny

Wprowadzony przez Mc Cullocha model elementu, nazywany zwykle neuronem formalnym, może się znajdować w jednym z stanów, dwu a mianowicie w stanie pobudzenia (aktywności) — oznacza się go zwykle umownie 1 — lub w stanie niepobudzenia (pasywności), który oznaczamy umownie 0. Zmiana stanu jest możliwa jedynie w ustalonych momentach czasu t, $t+1$, $t+2$, ..., $t+k$, które można określić za pomocą liczb całkowitych. Każdy neuron formalny posiada wejścia pobudzające i hamujące (rys. 15). Zostaje on pobudzony wówczas, gdy liczba wejść, do których przyłożono sygnały pobudzające, jest większa od zadanej wartości, zwanej wartością progową, i gdy jednocześnie nie działa żadne wejście hamujące. Tak np. w układzie podanym na rys. 15, jeżeli próg ma wartość 2, muszą zostać pobudzone dowolne trzy wejścia pobudzające spośród czterech narysowanych x_1, x_2, x_3, x_4, a jednocześnie na wejściach hamujących x_5 i x_6 pobudzenia muszą być równe zeru.



Rys. 15. Schemat połączeń neuronu

Późniejsze badania neurofizjologiczne wykazały, że warunkiem pobudzenia jest, aby różnica między liczbą wejść pobudzających a liczbą wejść hamujących była równa wartości progowej lub większa od niej. Rozszerza to możliwości sieci. Ponadto uważa się, że każdy element wprowadza jednostkowe opóźnienie (np. 1 milisekundę), równe czasowi trwania jednego taktu pracy sieci.

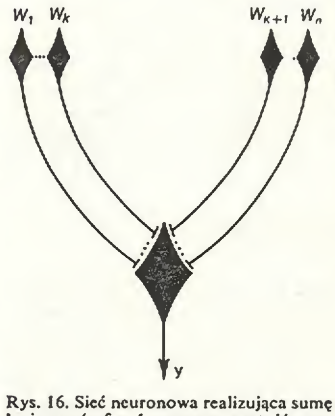
Z tak zdefiniowanych elementów można zestawić proste sieci realizujące podstawowe działania logiczne. Łącząc np. elementy w sposób podany na rys. 16, na wyjściu neuronu formalnego o progu 1 otrzymamy pobudzenie wtedy, i tylko wtedy, kiedy zostanie pobudzony co najmniej jeden neuron wejściowy. Zapisując to w sposób formalny, otrzymamy:

$$N(t+1) = W_1(t) \vee W_2(t) \vee W_3(t) \vee \dots \vee W_n(t),$$

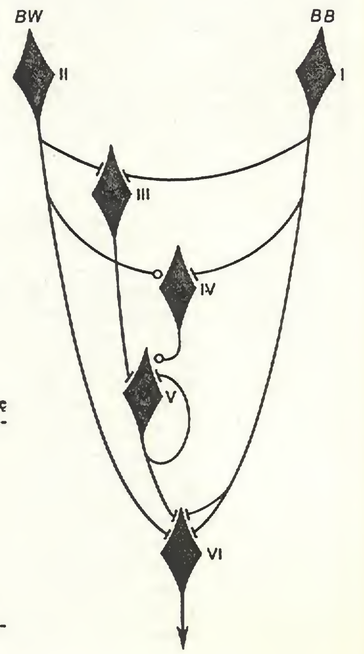
gdzie $N(t+1)$ — stan pobudzenia neuronu wejściowego w chwili $t+1$; $W(t)_i$ — stan pobudzenia neuronu i w chwili t; a zatem sieć realizuje sumę logiczną zdarzeń, którymi są sygnały wejściowe.

Jeżeli w tej samej sieci dokonamy pewnej zmiany, a mianowicie próg będzie równy liczbie neuronów wejściowych n, to wyjście będzie pobudzane jedynie wówczas, gdy pobudzone zostaną wszystkie wejścia. Oznacza to, że sieć realizuje iloczyn logiczny

$$N(t+1) = W_1(t) \wedge W_2(t) \wedge \dots \wedge W_n(t).$$



Rys. 16. Sieć neuronowa realizująca sumę logiczną (cyfra 1 oznacza wartość progową)



Rys. 17. Sieć realizująca uczenie się odruchu warunkowego

model uczenia się odruchu warunkowego

Liczne prace dotyczące tego typu sieci wykazały, że mają one ogromne możliwości i mogą realizować bardzo szeroką klasę procesów, o których dawniej myślano, że są możliwe jedynie w organizmach żywych. W szczególności sieci te mogą modelować odruchy warunkowe, to znaczy mogą się uczyć reagowania na bodziec warunkowy, który przez pewien czas towarzyszył bodźcowi bezwarunkowemu. Na rys. 17 przedstawiono sieć, która w uproszczony sposób realizuje powstanie odruchu warunkowego. W układzie są dwa wejścia: wejście bezwarunkowe *BB* i wejście warunkowe *BW*. Bodziec bezwarunkowy pobudza neuron I, a ten sygnałem podwójnej intensywności (dwa wejścia) pobudza z kolei neuron wyjściowy VI. Pobudzenie jedynie wejścia warunkowego *BW* nie wystarcza do pobudzenia komórki VI, która ma próg 2. Łączne pobudzenie obu wejść uruchamia komórkę iloczynową III, ta zaś z kolei komórkę V. Wskutek zastosowania sprzężenia zwrotnego impuls zjawiający się na wyjściu neuronu V jest z powrotem przykładany na jego wejście. Otrzymujemy układ samowzbudny, w którym stale krążą impulsy (jeśli nie zostaną przerwane przez przyłożenie sygnału hamującego z neuronu IV). Po takim łącznym zjawieniu się sygnałów na obu wejściach stan modelu jest więc inny i przy następnym pobudzeniu wystarczy już przyłożenie bodźca jedynie do wejścia *BW*, aby (łącznie ze stale docierającymi sygnałami z komórki V) pobudzić komórkę wyjściową VI.

Jak widać, komórka V została tu wykorzystana jako element pamięciowy. Ponadto przewidziano tu wygaszanie odruchu za pomocą komórki IV. Jak wspomniano, samo przyłożenie bodźca warunkowego pobudza komórkę IV i wygasza komórkę V.

Układy tego typu mogą być bardzo rozbudowane, tak aby tworzenie się odruchu nastąpiło dopiero po wielokrotnym przyłożeniu pary bodźców: warunkowego i bezwarunkowego.

Nowe informacje o własnościach komórek nerwowych wykazały, że są to elementy znacznie bardziej skomplikowane. Wiele prac neurocybernetycznych w latach sześćdziesiątych poświęcono modelowaniu własności zarówno pojedynczych neuronów, jak i prostych sieci złożonych z takich modeli. Na rys. 18 przedstawiono przykładowy schemat elektronicznego modelu neuronu, a na rys. 19 — zdjęcie tego typu modeli wykonanych w Zakładzie Bioniki byłego Instytutu Cybernetyki Stosowanej PAN. Nadal nie ma jednak jednolitych poglądów na temat tych własności neuronu, które są najbardziej istotne z punktu widzenia przekazywania i przetwarzania informacji w systemie nerwowym.

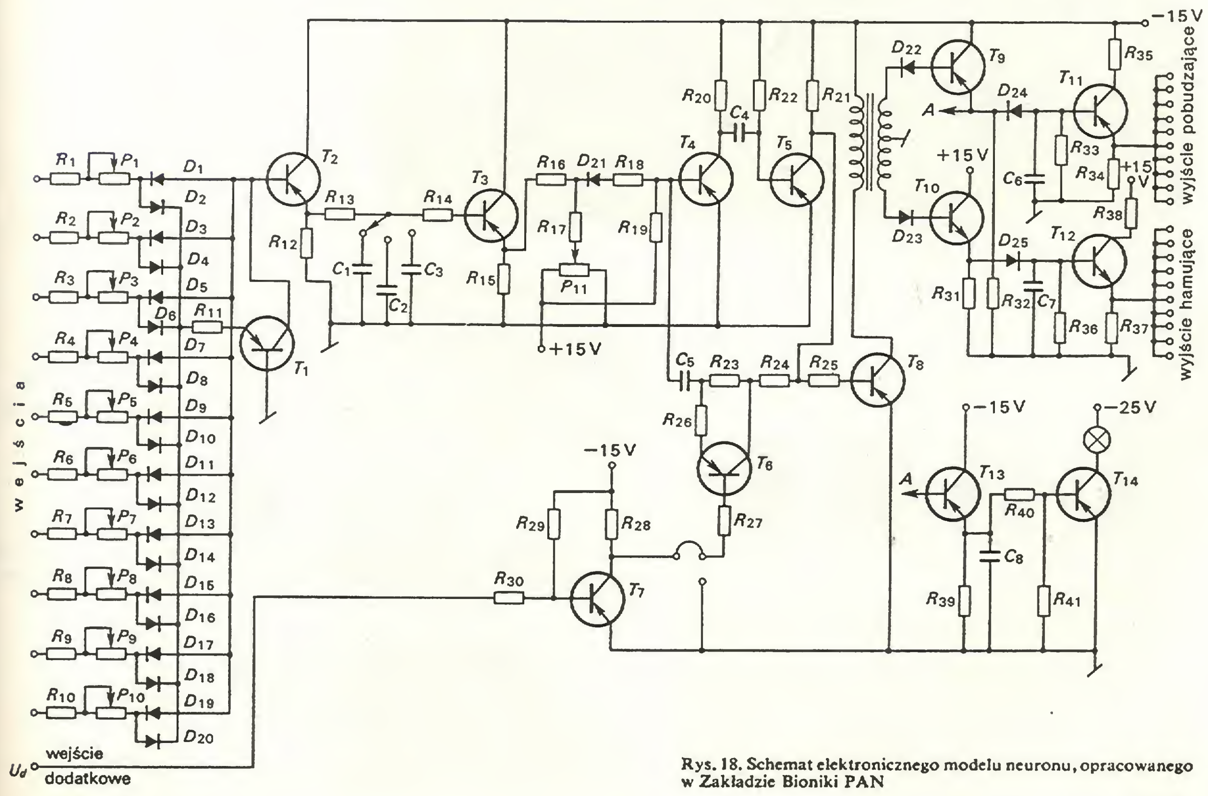
modele neuronu

Rys. 20 przedstawia schemat funkcjonalny modelu, neuronu, w którym wykorzystano obecne poglądy na przebiegi zjawiska w neuronie. Układ ma szereg wejść, z których jedne przenoszą pobudzenia, inne — hamowania. Sygnały wejściowe mnożymy przez pewien współczynnik w_i, zwany wagą wejścia, następnie sygnały z wszystkich wejść sumujemy i otrzymujemy wypadkowy sygnał pobudzający:

$$e_\Sigma = \sum_{i=1}^{k} x_i \, w_i(u_i),$$

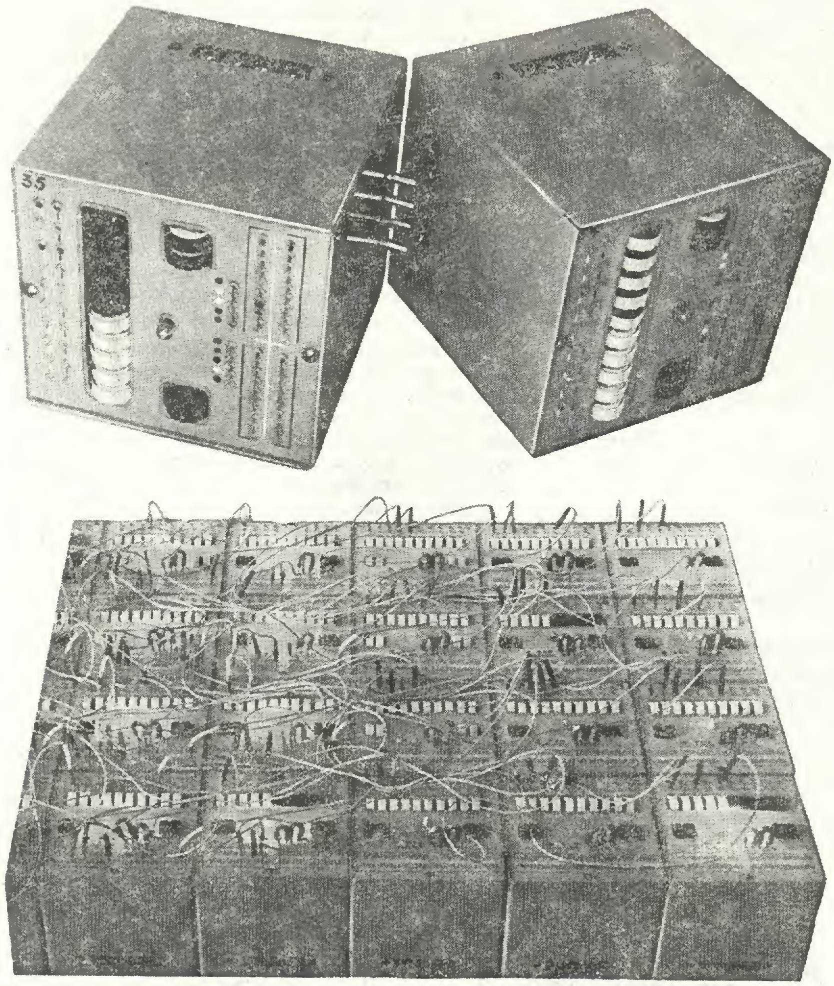
przy czym k — liczba wejść, $w_i(u_i)$ — waga i-tego wejścia, u_i — sygnał sterujący wielkością wagi.

W omawianym modelu wprowadzono drugi rodzaj wejść w postaci sygnałów u, zmieniających wartość wagi wejścia. Przypuszcza się, że z istnieniem tego rodzaju sygnałów związane są tzw. plastyczne własności neuronu, czyli możliwość uczenia się. Jeżeli występują sygały u, można modelować wytwarzanie odruchów warunkowych oraz inne zjawiska związane z procesami uczenia się w organizmach żywych. Neuron odznacza się pewną bezwładnością, co oznacza, że jego zachowanie się zależy od częstości impulsów przykładanych do wejść. Dlatego też model neuronu zawiera element inercyjny o charakterze układu całkującego *RC*, uśredniającego w pewnym zakresie przebieg e_Σ. Następnym blokiem jest tzw. generator progowy impulsów. W bloku tym wytwarza się krótki impuls o stałej amplitudzie za każdym razem, gdy wypadkowe pobudzenie e_Σ przekracza



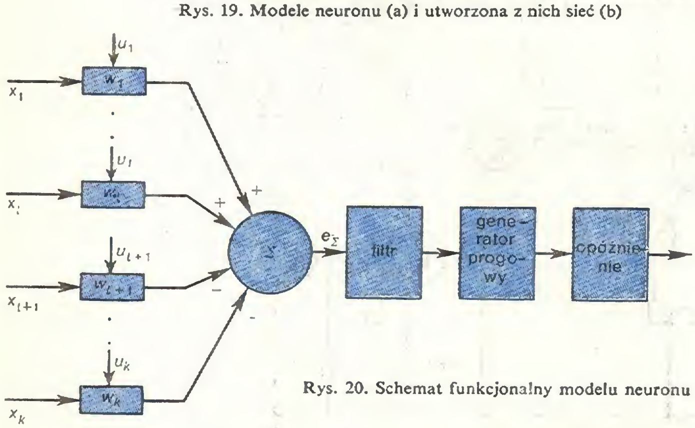
Rys. 18. Schemat elektronicznego modelu neuronu, opracowanego w Zakładzie Bioniki PAN

wartość progową. Im większa jest różnica pomiędzy wartością progową a e_{Σ}, tym większa jest częstotliwość generowanych impulsów. Ponadto w modelu neuronu uwzględnia się stałe opóźnienie, występujące z reguły w komórce nerwowej. Sygnał wyjściowy z modelu może być podany na wejścia większej liczby innych neuronów sieci modelujących pewne fragmenty tkanki nerwowej.



a) b)

Rys. 19. Modele neuronu (a) i utworzona z nich sieć (b)



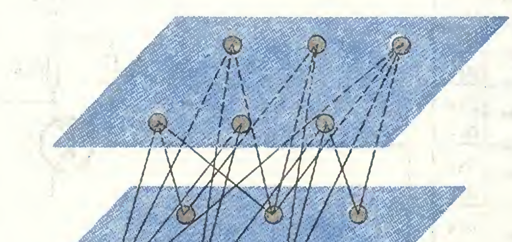
Rys. 20. Schemat funkcjonalny modelu neuronu

sieci neuronów

Podstawowe zagadnienie, które się pojawia na tym etapie badania procesów w układzie nerwowym, polega na ustaleniu sposobu połączeń między neuronami, czyli struktury sieci. Na podstawie niektórych danych anatomicznych oraz przesłanek wynikających z badań elektrofizjologicznych (czyli badań procesów elektrycznych w tkance) tylko w nielicznych wypadkach można ustalić ogólne zasady łączenia neuronów w sieci. Takim przykładem mogą być sieci warstwowe z połączeniami lokalnymi (rys. 21). Sieci podobnego typu odgrywają prawdopodobnie istotną rolę w procesie przetwarzania informacji wzrokowej docierającej do siatkówki oka. Wykazano m.in., że sieci te mogą służyć do wykrywania pewnych cech lokalnych obrazu, takich jak zakończenia, załamania i zagięcia linii, jak również skrzyżowania i rozgałęzienia.

modele percepcji wzrokowej

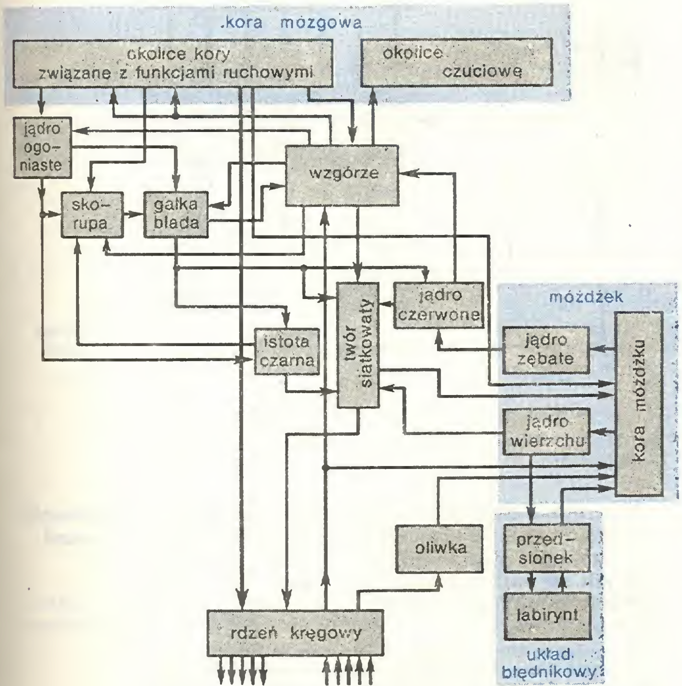
W wyniku badań nad procesami rozpoznawania obrazów zachodzącymi w tzw. analizatorze wzrokowym, czyli w części układu nerwowego zajmującego się analizą sygnałów docierających do siatkówki oka,



Rys. 21. Sieć warstwowa z połączeniami lokalnymi

zbudowano szereg modeli percepcji wzrokowej. Badaniom tym zawdzięczamy udoskonalenie urządzeń do automatycznego rozpoznawania obrazów, a zwłaszcza do skonstruowania automatycznych czytników tekstów drukowanych lub maszynopisowych (czytniki takie są już stosowane w polskich drukarniach przy automatycznym składaniu tekstów).

Oprócz prac nad modelowaniem procesów percepcji wzrokowej i słuchowej modelowano również sterowanie mięśniami. Z teoretycznego i praktycznego punktu widzenia jest rzeczą bardzo interesującą, w jaki sposób układ nerwowy steruje zespołem kilkudziesięciu czy kilkuset mięśni, uwzględniając przy tym liczne własności mechaniczne układu ruchowego człowieka i zwierzęcia oraz sytuację otoczenia. Okazało się, że część układu nerwowego sterującego ruchem ma również bardzo złożoną budowę hierarchiczną (rys. 22). Pomimo licznych badań nie mamy o niej wystarczających informacji, by zbudować jej zadowalający model. Można jednak w przybliżeniu zmodelować najniższy stopień w hierarchii układu ruchowego, zawierający elementy sterujące skurczem mięśni.

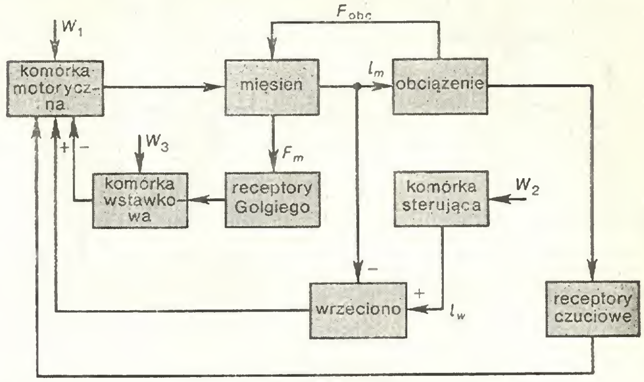


Rys. 22. Uproszczony schemat układu ruchowego ssaków

Przykład takiego układu sterowania mięśniami przedstawiono na rys. 23. Ma on jedynie zilustrować ogólny charakter tego typu schematów. Znajdujące się w rdzeniu kręgowym komórki motoryczne (zwane niekiedy komórkami alfa) powodują skurcz włókien mięśniowych przez wysłanie odpowiedniego pobudzenia włóknami nerwowymi. Skurcz mięśnia przenosi się oczywiście na obiekt, którym może być np. podnoszony przedmiot, a ciężar tego przedmiotu wpływa na wielkość skurczu mięśnia i rozwijaną przez niego siłę. W mięśniu oraz w ścięgnach istnieją specjalne elementy pomiarowe, tzw. wrzeciona i organy Golgiego, służące do pomiaru skurczu mięśnia. Zmianę długości mięśnia l_1 sygnalizuje wrzeciono komórkom motorycznym znajdującym się w rdzeniu kręgowym. Dzięki organom Golgiego komórki te otrzymują też informację o sile rozwijanej przez mięsień. Również receptory czucia powierzchniowego mogą przekazywać bodźce do komórek motorycznych alfa. Z powyższego widać, że komórka motoryczna jest pewnego rodzaju centrum skupiającym informacje o efekcie, jaki wywołała ona sama, powodując skurcz mięśnia.

układ sterowania mięśniami

Obwodowo-rdzeniowy układ sterowania mięśniami zawiera zatem kilka pętli (rys. 23), a własności elementów są do tego stopnia skomplikowane, że układ wymaga specjalnych badań modelowych. W odróżnieniu od klasycznych układów technicznych odznacza się on szerokim zakresem adaptacji do zmiennych warunków obciążenia, stosunkowo dużą precyzją i szybkością działania, budzi więc zainteresowanie nie tylko biologów, ale i techników. Przedstawiając w bardzo uproszczony sposób niektóre wyniki badań modelowych, można powiedzieć, że proces sterowania skurczem mięśnia dzięki jednoczesnemu wykorzystaniu kilku rodzajów sygnałów wejściowych i kilku obwodów sterowania umożliwia znacznie szybszą i precyzyjniejszą regulację położenia kończyny niżby to było możliwe w wypadku klasycznego układu ze sprzężeniem zwrotnym. Ponadto analiza tego układu umożliwia wyjaśnienie pewnych sytuacji patologicznych, jak np. sztywność mięśniową oraz spastyczność.



Rys. 23. Funkcjonalny schemat układu sterowania mięśniem: l_m długość mięśnia, F_{obc} siła obciążenia, F_m siła rozwijana przez mięsień, l_w sygnał ustalający długość wrzeciona, W_1, W_2 i W_3 wejścia sygnałów z innych ośrodków nerwowych

Modelowanie wyższych szczebli układu ruchowego napotyka znacznie większe trudności, chociaż np. ostatnie wyniki badań dotyczące roli móżdżku w procesie korekcji sterowania ruchem są bardzo interesujące z biocybernetycznego punktu widzenia.

Postęp badań w neurocybernetyce jest ściśle związany z rozwojem badań w dziedzinie neurofizjologii, a zwłaszcza elektrofizjologii, i z wykorzystaniem nowoczesnych metod techniki, zwłaszcza elektroniki.

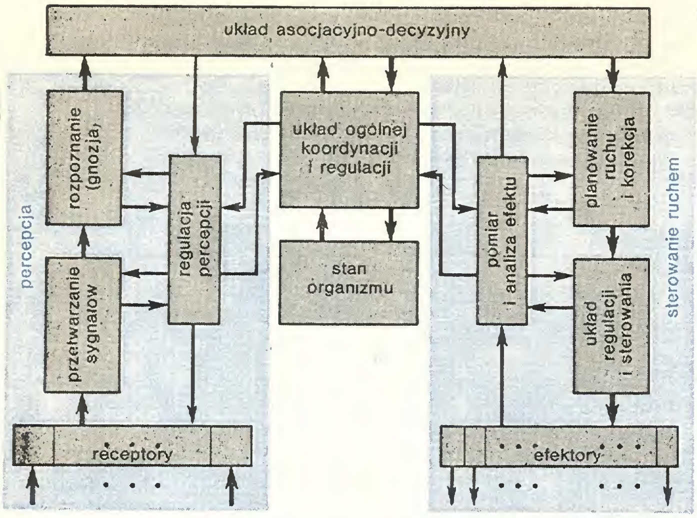
układ nerwowy — powiązania między ośrodkami

Na podstawie współczesnych poglądów neurofizjologicznych można przedstawić, bardzo zresztą ogólnikowo, schematy funkcjonalnych powiązań (które nie budzą większych wątpliwości) pomiędzy głównymi ośrodkami układu nerwowego. Przykład takiego schematu przedstawiono na rys. 24.
Schemat ten składa się z 3 części:

1. Układ dośrodkowy, w którym następuje przekazywanie i analiza sygnałów docierających z otoczenia. Sposób analizy jest regulowany przez nadrzędne ośrodki koordynująco-decyzyjne.
2. Centralny układ asocjacyjno-decyzyjny, w którym na podstawie stanu otoczenia, stanu organizmu, dotychczasowego doświadczenia i prognozy skutków zostaje wypracowana decyzja o reakcji organizmu.
3. Układ odśrodkowy, w którym na podstawie podjętej decyzji wypracowany jest plan ruchu (ogólnej reakcji organizmu), uwzględniający sytuację statyczną i dynamiczną oraz doświadczenie organizmu w zakresie optymalizacji ruchu.

Ponadto wydzielić można układ koordynacji i regulacji różnych ośrodków, umieszczony w środkowej części rysunku.

Szczególnie obiecujące są prace związane z mode-



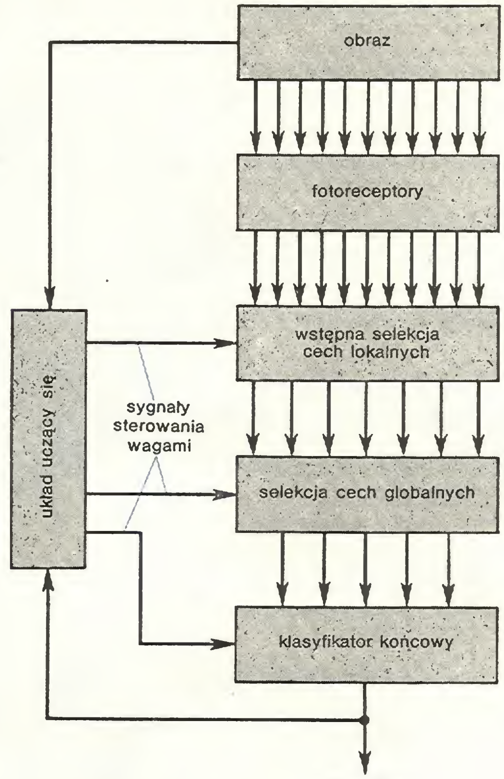
Rys. 24. Funkcjonalny schemat powiązań między głównymi ośrodkami układu nerwowego

lowaniem procesów uczenia się i badania sieci neuronowych o zmiennej strukturze. Umożliwią one wyjaśnienie niektórych podstawowych procesów psychicznych związanych z podejmowaniem decyzji i zachowaniem się zwierząt oraz staną się źródłem koncepcji przy konstruowaniu złożonych urządzeń technicznych do selekcji informacji i sterowania złożonymi systemami.

układ uczący się rozpoznający obrazy

Jako przykład przedstawiono na rys. 25 uproszczony funkcjonalny schemat układu uczącego się rozpoznawania obrazów. Zgodnie z obecnymi poglądami rozpoznawanie obrazu, np. litery, cyfry, figury geometrycznej, fotografii osoby czy znajomych przedmiotów, odbywa się w kilku etapach. Poszczególne etapy zostały przedstawione na rysunku. Ostatecznym rezultatem rozpoznania jest ustalenie nazwy oglądanego przedmiotu; kod nazwy jest przekazany w organizmach żywych dalszym częściom układu nerwowego, a w urządzeniach technicznych — urządzeniom informatyki wykorzystującym rozpoznanie.

Ostatnie dwa (a być może trzy) etapy rozpoznania zawierają sterowane elementy, np. w postaci wyżej opisanych modeli neuronów o sterowanej wadze. Zadaniem układu uczącego się jest wywołanie takich zmian w układzie selekcji cech obrazu oraz w klasyfikatorze, aby urządzenie rozpoznające popełniało jak najmniej błędów. Tak np. jedna z reguł uczenia się polega na tym, że przy błędnym rozpoznaniu zmniejszają się wagi cech, które uczestniczyły w rozpoznaniu, natomiast przy prawidłowym rozpoznaniu rosną wagi cech (wzmocnienia w odpowiednich kanałach), które się przyczyniły do prawidłowego rozpoznania, pozostałe ewentualnie maleją. Niekiedy można teoretycznie wykazać (niekiedy zaś praktycznie sprawdzić), że proces uczenia się doprowadza do rozpoznawania przez tego typu układ uczący się różnego typu obrazów, np. liter, cyfr lub prostych obrazów konturowych. Ostatnio pojawiły się również urządzenia do rozpoznawania bardziej złożonych obrazów, np. fotografii osób, lub dźwięków mowy.



Rys. 25. Funkcjonalny schemat układu uczącego się rozpoznawania obrazów

sterowanie robotami

Inne równie przydatne w przyszłości zastosowanie sieci uczących się polegać będzie najprawdopodobniej na sterowaniu ruchem różnego rodzaju robotów uczących się zastępowania człowieka w warunkach niebezpiecznych lub nużących i nieopłacalnych dla człowieka.

M. A. Arbib *Mózg, maszyna, matematyka*, Warszawa 1968; W. R. Ashby *Wstęp do cybernetyki*, Warszawa 1961; A. J. Berg *Informacja i cybernetyka*, Warszawa 1970; A. A. Bratko i in. *Modelowanie czynności psychicznych*, Warszawa 1973; E. A. Feigenbaum, J. Feldman *Maszyny matematyczne i myślenie*, Warszawa 1972; R. Gawroński *Rozpoznanie i decyzja*, Warszawa 1970; *Bionika — system nerwowy jako układ sterowania*, red. R. Gawroński, Warszawa 1970; W. Głuszkow *Wstęp do cybernetyki*, Warszawa 1967; W. Karczewski *Zjawiska elektryczne w organizmie*, Warszawa 1963; P. de Latil *Sztuczne myślenie — wstęp do cybernetyki*, Warszawa 1958; W. Słuckin *Mózg i maszyny*, Warszawa 1957; J. Z. Young *Model mózgu*, Warszawa 1968.

Fizyka medyczna

Oskar Chomicki

Fizykę medyczną można określić jako dział fizyki stosowanej, który traktuje o zastosowaniu zasad i metod fizycznych we wszystkich dziedzinach zapobiegania, rozpoznawania i leczenia chorób ludzkich, jak również w badaniach naukowych mających na celu zdrowie człowieka.

Jest to definicja niejednoznaczna, co wynika stąd, że fizyka medyczna należy do nauk z pogranicza fizyki, biologii, medycyny i techniki.

związek między a medycyną fizyką

Z historycznego punktu widzenia związek pomiędzy fizyką a medycyną można podzielić na dwa okresy: przed odkryciem promieniowania jonizującego (promieni X i radu) i po jego odkryciu. Pierwszy okres rozpoczął się w czasach Odrodzenia — do końca XIX w. (związek między fizyką a medycyną opierał się głównie na osiągnięciach w fizyce dokonywanych przez lekarzy i na zdobyczach medycyny będących dziełem fizyków; tabela). Już Galileusz, który prawdo-

Niektórzy słynni lekarze zajmujący się badaniami fizycznymi oraz słynni fizycy interesujący się medycyną

Wiek	Nazwisko	Zawód wyuczony	Osiągnięcia w dziedzinie
XVI	Glibert (Anglia)	lekarz	magnetyzm
	William Harvey (Anglia)	lekarz	dynamika krwiobiegu, zastosowanie metod naukowych w medycynie
XVII	Giovanni Alfonso Borelli (Włochy)	przyrodnik	praca mięśni, mechanika ruchu ssaków
	Robert Hooke (Anglia)	lekarz	sprężystość ciał, mikroskopia
XVIII	Luigi Galvani (Włochy)	anatom	elektryczność w medycynie
	Alessandro Volta (Włochy)	fizyk i fizjolog	elektryczność
	Joseph Black (Szkocja)	lekarz	ciepło
	Thomas Young (Anglia)	lekarz	teoria falowa światła, optyka fizjologiczna, widzenie barwne
XIX	Hermann Ludwig Helmholtz (Niemcy)	fizyk, anatom i fizjolog	oftalmoskop, teoria widzenia barwnego, akustyka i teoria słyszenia
	D'Arsonval (Francja)	fizyk	terapia elektrycznością, diagnostyka

podobnie początkowo studiował medycynę, do badań nad wahadłem używał podobno własnego pulsu. Dopiero jednak jego rodak Santorio stał się poniekąd pierwszym fizykiem medycznym; on to zastosował termometr do mierzenia temperatury ciała ludzkiego.

powstanie fizyki medycznej

Drugi okres rozwoju fizyki medycznej, zapoczątkowany w 1895 r. odkryciem promieni *X* przez Roentgena oraz radu przez małżonków Curie, trwa do dziś. Sami fizycy nie mogą już przeszczepiać swoich odkryć do medycyny; oba te działy stały się zbyt rozległe i wyspecjalizowane. Powstała więc nowa specjalizacja, fizyka medyczna, która się włączyła do ścisłej współpracy z medycyną, szczególnie radiologiczną (tabela).

W różnych krajach stan fizyki medycznej jest różny. Zależy on od wielu czynników (organizacja służby zdrowia, poziom ekonomiczny itp.). Liczba fizyków medycznych na milion ludności w 1975 r.: od 0,02 (Nigeria) do 17 (Wielka Brytania). Znaczne różnice występują w tym zakresie nawet w krajach uprzemysłowionych (np. Francja ma wskaźnik 0,9, Szwecja — 5,0, Polska zaś — 3,0).

praca fizyka medycznego

Większość fizyków medycznych pracuje w dziedzinie zastosowań promieniowania jonizującego, część — w dziedzinach, które nabierają coraz większego znaczenia, jak np. zastosowanie podczerwieni, dziedzina laserów, bioprądów czy mikrofal. Oblicza się, że np. w Wielkiej Brytanii działy te obejmują obecnie ok. 25% wszelkiego typu zastosowania fizyki do medycyny (tabela).

Zakres działalności fizyka medycznego

Dziedzina	Działalność fizyka
Współpraca z lekarzami w zakresie pracy klinicznej	pomiary ciśnienia i szybkości krążenia krwi oraz płynów tkankowych bioelektryczna czynność ciała elektroniczne metody diagnostyczne urządzenia techniki medycznej (serce-płuco, sztuczna nerka itp. — przy współpracy z inżynierami) aparatura optyczna ocena i planowanie dawek w radiodiagnostyce i radioterapii metody medycyny nuklearnej metody diagnostyki ultradźwiękowej
Współpraca z lekarzami w zakresie prac badawczych	radiobiologia i biofizyka nowe metody pomiaru promieniowania nowe urządzenia do napromieniowania zastosowanie maszyn liczących zastosowanie teorii informacji do diagnostyki analiza aktywacyjna modelowanie matematyczne
Ochrona przed promieniowaniem	kontrola narażania na promieniowanie w szpitalach i laboratoriach pomiary radioaktywności całego ciała
Nauczanie fizyki	studenci medycyny radiologowie i radioterapeuci przyrodnicy studenci radiobiologii

zastosowanie fizyki w medycynie

Typowe zastosowania fizyki w medycynie

Specjalność	Zastosowanie fizyki
Anestezjologia	zastosowanie fizyki gazów, analiza gazów
Anatomopatologia	optyka mikroskopowa, technika radioizotopowa
Chirurgia	urządzenia biomechaniczne, technika radioizotopowa
Dermatologia	leczenie chorób skóry przez naświetlanie, pomiary kolorymetryczne, technika radioizotopowa
Fizykoterapia	leczenie krótkofalowe i mikrofalowe, stymulatory elektryczne, telemetria
Ginekologia i położnictwo	technika radioizotopowa, ultradźwiękowa, promieniolecznictwo
Kardiologia	elektrokardiografia, fonokardiografia, balistokardiografia, pobudzenie mięśni, aparatura do badania ciśnienia i przepływu krwi, stymulatory serca
Laryngologia	aparatura audiometryczna
Medycyna ogólna	technika radioizotopowa, zastosowanie ultradźwięków, technika podczerwieni
Neurologia i neurochirurgia	zastosowanie ultradźwięków, elektroencefalografia
Ochrona przed promieniowaniem	ustalenie przepisów bezpieczeństwa pracy, pomiar skażeń promieniotwórczych
Okulistyka	zastosowanie laserów i ultradźwięków
Radiologia i radioterapia (onkologia)	scyntygrafia, wzmacniacze obrazu, pomiary zawartości minerałów w kościach, dozymetria, planowanie leczenia

Najważniejsze rodzaje zastosowania termografii medycznej

Specjalność	Zastosowanie
Onkologia	wykrywanie raka sutka i przerzutów nowotworowych
Interna	choroby układu krążenia
Ortopedia	lokalizacja uszkodzeń mięśni, stawów i kręgów
Ginekologia i położnictwo	lokalizacja łożyska
Dermatologia	badania stopnia oparzenia, odmrożeń; chirurgia plastyczna
Farmakologia	badanie skuteczności leków i niektórych metod leczenia

zadania fizyka medycznego

Główne zadania stojące przed fizykiem medycznym wiążą się z jednej strony z pracą w zakładach fizyki akademii medycznych, które są terenem intensywnej działalności dydaktycznej i badawczej związanej z ogólnymi zagadnieniami medycznymi, z drugiej zaś strony pozostają w związku z jego pracą w służbie zdrowia, przede wszystkim w szpitalu, tylko bowiem szpital — z uwagi na swoje możliwości materialne i organizacyjne — może stanowić bazę działalności fizyka, wymagającego skomplikowanej i kosztownej aparatury i odpowiedniego zrozumienia ze strony personelu lekarskiego. Odpowiedzialność lekarza i fizyka dzieli się w ten sposób, że lekarz zleca określone badanie lub metodę leczenia, do którego wymagane jest zastosowanie aparatury lub metody fizycznej, natomiast fizyk dba o to, aby postępowanie było poprawne metodycznie i pozwoliło na uzyskanie zamierzonego skutku.

zajęcia rutynowe

Zadania fizyka medycznego w służbie zdrowia dzielą się na rutynowe i badawcze. Do zajęć rutynowych należy troska o niezawodność aparatury i rzetelność otrzymywanych wyników, co zwykle sprowadza się do nadzoru nad konserwacją aparatury i nad personelem technicznym obsługującym urządzenia do standardowych badań (przeliczniki, kamery, elektrokardiografy, encefalografy itp.).

O wiele ważniejsza i wymagająca wyższych kwalifikacji jest działalność fizyka medycznego w prowa-

badania naukowe

dzeniu badań naukowych i opracowywaniu nowych rozwiązań technicznych problemów medycznych stawianych przez lekarza. Fizyk może się w tym zakresie wykazać całą swoją wiedzą fizyczną, gdyż dziedziny badań medycznych mogą być bardzo rozległe, od chirurgii kości, rehabilitacji (zastosowania mechaniki) począwszy, przez okulistykę (zastosowania optyki), fizykoterapię (zastosowanie prądu elektrycznego i elektrostatyki), aż do radiodiagnostyki, radioterapii i medycyny nuklearnej (zastosowanie fizyki atomowej). Wszechstronność wykształcenia fizyka pozwala na efektywną współpracę z lekarzami różnych specjalności, a równocześnie zapewnia ścisły, ujęty matematycznie (w miarę możliwości) pogląd na zagadnienia medyczne.

zadania w ochronie radiologicznej

Fizyk medyczny znajduje rozległe pole działania w zakresie ochrony przed promieniowaniem. Prowadzi badania radioaktywności naturalnej środowiska, sprawuje stały nadzór nad poziomem skażeń promieniotwórczych, które mogą wystąpić wskutek pracy takich urządzeń jak reaktory czy elektrownie atomowe, oraz współuczestniczy w pracach bardziej podstawowych, mających na celu badanie efektów promieniowania na tkankę czy komórkę. W Polsce wielu fizyków medycznych uzyskało tytuł inspektora ochrony radiologicznej, co podkreśla wagę pełnionych przez nich funkcji.

zadania w radiodiagnostyce

Rola fizyka medycznego w zakresie radiodiagnostyki polega na nadzorze nad bezpieczeństwem i higieną pracy, na czuwaniu nad typową aparaturą i przede wszystkim na ścisłej współpracy z lekarzem w opracowywaniu lub adaptowaniu nowych metod diagnostyki (nowe badania kontrastowe, tomografia, angiografia, kseroradiografia, zastosowanie wzmacniacza obrazu itp.). Należy podkreślić, że działalność fizyka podnosi w sposób zasadniczy jakość wykonywanych badań; przez ścisłą kontrolę aparatury osiąga się ponadto optymalizację dawki promieniowania użytej przy badaniu.

zadania w radioterapii

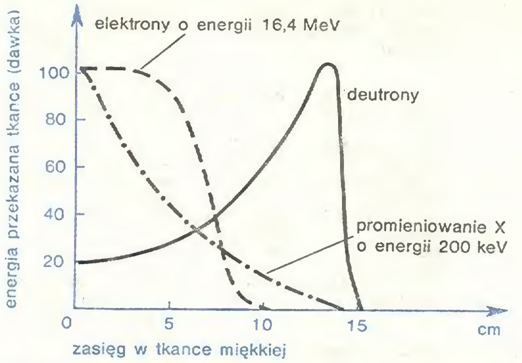
W radioterapii fizyk medyczny odgrywa jeszcze ważniejszą rolę: zajmuje się przede wszystkim planowaniem leczenia, a poza tym — opracowywaniem dodatkowych urządzeń do kierowania wiązką, kalibracją aparatury i rozwojem nowych metod radioterapeutycznych.

Promieniowanie jonizujące w medycynie

Promieniowanie jonizujące jest terminem, przez który będziemy przede wszystkim rozumieć promieniowanie składające się z cząstek naładowanych, takich jak elektrony, pozytony, protony, deuterony, cząstki alfa, piony, oraz cząstek nienaładowanych, jak fotony czy neutrony. Wszystkie te cząstki niosą ze sobą pewną energię i dopiero częściowe przekazanie tej energii materii, przez którą przechodzą, pozwala na ich wykrycie. Nas będą oczywiście interesować te rodzaje wzajemnego oddziaływania promieniowania i materii, które albo umożliwiają stosowanie urządzeń wykrywających promieniowanie (detektorów), albo powodują uszkodzenie lub zniszczenie tkanki.

cząstki naładowane — jonizacja

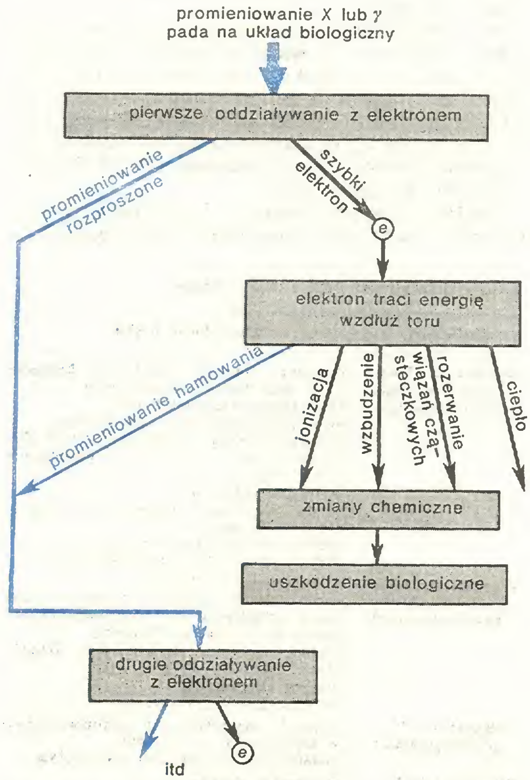
Przechodząc przez materię, naładowana cząstka może wywoływać jonizację, tj. wybijać elektrony z atomów, wzbudzać atomy, rozrywać wiązania cząsteczkowe, a także, jeżeli zostanie nagle zahamowana lub przyśpieszona, wysyłać promieniowanie elektromagnetyczne, zwane promieniowaniem hamowania. Z biologicznego punktu widzenia najważniejszą sprawą jest jonizacja wywołana przez cząstkę w tkance. Im cząstka jest cięższa, tym szybciej traci energię i tym wcześniej jest zatrzymana, a więc jej zasięg jest mniejszy. Na przykład zasięg cząstki alfa w tkance wynosi ok. 0,07 mm, cząstki beta (elektrony) — 4,2 mm (przy energii 1 MeV). Jeśli energia elektronów wynosi kilkanaście MeV, zasięg ten wzrasta do kilku cm. Krzywa zależności zasięgu od energii elektronów przekazanej tkance (rys. 1) ma kształt przypominający prostokąt o stosunkowo stromych zboczach.



Rys. 1. Zależność zasięgu promieniowania jonizującego od energii (energia promieniowania jest wyrażona w jednostkach względnych)

cząstki nienaładowane — różnorodne procesy

Promieniowanie nienaładowane, a więc przede wszystkim promieniowanie rentgenowskie lub gamma, zachowuje się zupełnie inaczej przy przejściu przez materię. We wzajemnym oddziaływaniu indywidualnego fotonu *X* lub gamma i elektronu przekazanie energii odbywa się wieloma sposobami. Cechą charakterystyczną wzajemnego oddziaływania wiązki fotonów *X* lub gamma i materii jest wykładnicza zależność energii przekazanej przez wiązkę promieniowania od grubości ciała pochłaniającego (krzywa na rys. 1; gdy cząstki są ciężkie, jak deuterony, krzywa ma charakterystyczne, ostre maksimum).



Rys. 2. Schemat oddziaływania promieniowania *X* lub γ z układem biologicznym. Całkowite pochłonięcie energii padającego fotonu *X* lub γ następuje po ok. 30 kolejnych aktach takiego oddziaływania

„historia życia" fotonu

By wyjaśnić złożoność procesów pochłaniania i rozpraszania promieniowania X padającego na układ biologiczny rozpatrzmy „historię życia" fotonu (rys. 2). Można przypuścić, że pierwotne wzajemne oddziaływanie fotonu i elektronu następuje w odległości średniej drogi przebytej w tkance. Z obliczeń wynika, że odległość ta wynosi 56 cm. Stosując teraz odpowiednie wzory opisujące rozproszenie i pochłanianie, otrzymamy, że ze 100 fotonów padających o energii 20 MeV powstaną w pierwotnym oddziaływaniu 44 pary elektron–pozyton, 56 szybkich elektronów comptonowskich oraz 56 rozproszonych fotonów o średniej energii 5,4 MeV. Pary elektron–pozyton ulegną anihilacji, dając początek 88 fotonom o energii 511 keV. Dalej śledzić będziemy los fotonów powstałych z anihilacji, choć podobne rozważania można by również zastosować do fotonów rozproszonych o energii 5,4 MeV.

Otóż po przebyciu w tkance drogi 10,3 cm owych 88 fotonów ulegnie rozproszeniu comptonowskiemu, tzn. powstanie 88 elektronów comptonowskich o energii 175 keV i 88 rozproszonych fotonów o energii 336 keV. Fotony te przebędą średnią drogę 9,1 cm do następnego miejsca oddziaływania. Proces ten będzie się powtarzał aż do chwili, gdy energia fotonów spadnie do kilkunastu keV, a wówczas zacznie przeważać proces fotoelektryczny, będzie wzrastała liczba fotoelektronów i będzie się zmniejszała liczba fotonów comptonowskich, aż pozostanie tylko jeden foton; a zatem proces rozproszenia i pochłaniania pierwotnych 100 fotonów o energii 20 MeV zostanie zakończony. Tak przedstawiona „historia życia" fotonu jest, oczywiście, bardzo uproszczona, ale przy zastosowaniu statystycznej metody Monte Carlo i obliczeń komputerowych, do np. 500 000 „historii", uzyskuje się poprawny ilościowy opis zjawisk pochłaniania i rozpraszania w tkance, co ma wielkie znaczenie przy określaniu dawki pochłoniętej.

dawka pochłonięta; grej

Jednostką dawki pochłoniętej jest grej — Gy; jest to dawka, przy której substancja o masie 1 kg pochłania energię 1 J (poprzednio stosowaną jednostką był rad; 1 rad = 10^{-2} Gy). Średnia dawka śmiertelna dla człowieka, którego całe ciało zostało napromieniowane promieniami X lub gamma, wynosi ok. 4 grejów, dopuszczalna zaś dawka roczna $\leq 5 \cdot 10^{-2}$ Gy. Ponieważ różne rodzaje promieniowania wywołują różny skutek biologiczny, wprowadza się tzw. współczynnik jakości promieniowania QF, który przy promieniowaniu X, gamma i elektronów równy jest 1, a np. przy szybkich neutronach sięga 10. Tak zwany równoważnik dawki jest określony jako iloczyn dawki pochłoniętej i współczynnika QF: stosowaną powszechnie jednostką równoważnika dawki jest rem (1 rem jest to równoważnik dawki 1 rad promieniowania o współczynniku QF = 1).

skutki działania promieniowania

Skutki działania promieniowania na ustrój zależą od wielu czynników: od wielkości dawki w określonej jednostce czasu, od masy ciała czy tkanki, wreszcie od biologicznych cech ustroju i wrażliwości jego poszczególnych tkanek. Pochłonięcie energii jonizacji przez komórkę powoduje zasadnicze zmiany w procesach wewnątrzkomórkowych, szczególnie w jądrze komórki: może nastąpić śmierć komórki, zahamowanie podziałów komórkowych, uszkodzenie chromosomów lub zmiany czynnościowe. Okazuje się, że wiele komórek może przeżyć dawki kilkuset rem, toteż, by całkowicie zniszczyć np. tkankę nowotworową, stosuje się dawki sięgające kilkunastu lub kilkuset tysięcy rem na określonym obszarze. Od dawna wiadomo, że promieniowanie jonizujące może pociągać za sobą szkodliwe skutki genetyczne; na ogół są one proporcjonalne do wielkości dawki, prawdopodobnie jednak nawet najmniejsze dawki, nawet pojedyncze akty jonizacji mogą prowadzić do mutacji genetycznych Szczególnie wrażliwy na promieniowanie jest organizm ludzki w czasie życia płodowego i to przede wszystkim w najwcześniejszym okresie 5–6 tygodni. Ponadto niektóre tkanki i narządy ciała ludzkiego, np. skóra, układ krwiotwórczy i limfatyczny, narządy rozrodcze, soczewka oka, błona śluzowa jelit, wykazują wysoką promienioczułość.

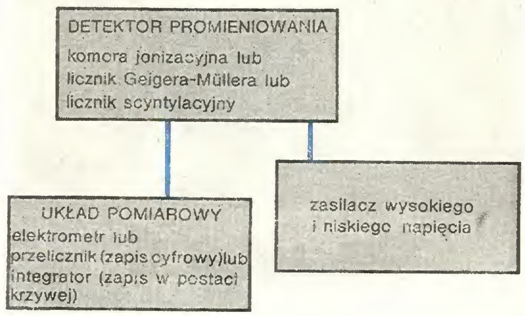
dawki dopuszczalne

Wszystkie te względy narzucają konieczność określenia maksymalnej dopuszczalnej dawki promieniowania jonizującego — i to różnej dla różnych grup ludności. Dla pracowników bezpośrednio narażonych na promieniowanie dawka nie powinna przekraczać 1,3 rema na kwartał, a w żadnym razie $5(N-18)$ remów w całym życiu (gdzie N jest wiekiem pracownika). Dla każdego człowieka dopuszczalna jest dawka 5 remów w ciągu pierwszych 30 lat życia.

Ludność narażona jest na promieniowanie jonizujące pochodzące z promieniowania kosmicznego, z promieniowania naturalnych pierwiastków promieniotwórczych oraz z promieniowania stosowanego w radiodiagnostyce i radioterapii, ponadto — w krajach uprzemysłowionych — na promieniowanie pochodzące z zakładów przemysłowych i instytutów badawczych stosujących radioizotopy, a także z pracujących elektrowni atomowych. Dokładna analiza wszystkich źródeł promieniowania jonizującego w Polsce prowadzi do wniosku, że przyjęta największa dopuszczalna dawka 5 rem na 30 lat uwzględnia z nadmiarem bezpieczeństwo całej ludności, nawet przy intensywnej rozbudowie do 2000 r. elektrowni atomowych.

mierniki promieniowania

Z drugiej strony — oddziaływanie promieniowania jonizującego na materię, o którym mówiliśmy wyżej z punktu widzenia jego szkodliwości dla organizmu, pozwala na jego wykrywanie przy zastosowaniu tzw. mierników promieniowania. Podstawowa zasada działania takich mierników ukazana jest na rys. 3. Zasadniczym elementem miernika jest detektor, tj. urządzenie służące do przetwarzania energii promieniowania na energię (sygnał) dogodną do pomiaru (najczęściej prąd elektryczny). W komorze jonizacyjnej jest to prąd ciągły a w licznikach Geigera–Müllera, w licznikach proporcjonalnych oraz scyntylacyjnych — prąd impulsowy, który następnie, po wzmocnieniu, zostaje zmierzony w układzie pomiarowym.



Rys. 3. Schemat typowych układów pomiarowych do mierzenia promieniowania jonizującego

Konstruuje się mierniki różnego rodzaju, zależnie od rodzaju promieniowania i od celu, któremu mają służyć. Dzisiaj jednak są to przeważnie tzw. liczniki scyntylacyjne, o których będzie mowa dalej w związku z omówieniem medycyny nuklearnej.

Radiodiagnostyka i radioterapia

radiologia

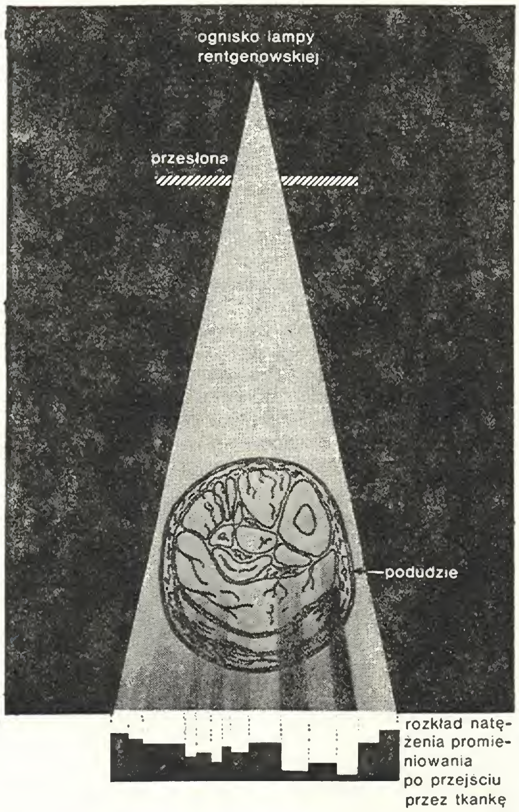
Radiodiagnostyka i radioterapia, w medycynie zwane łącznie radiologią, obejmują zastosowanie zarówno źródeł promieniowania rentgenowskiego (aparaty rentgenowskie) i promieni gamma czy strumieni cząstek, jak i zamkniętych lub otwartych źródeł promieniotwórczych (np. stosowane w radioterapii

źródła otwarte i zamknięte

igły radowe lub kobaltowe, które się umieszcza wewnątrz ciała chorego, są źródłami zamkniętymi, natomiast radioizotop jodu ^{131}J, stosowany w leczeniu m.in. przerzutów nowotworowych, uczestniczący w przemianie metabolicznej w organizmie, jest źródłem otwartym).

rentgenodiagnostyka

Radiodiagnostyka, zwana również rentgenodiagnostyką, jest najpowszechniej stosowaną metodą badań przy użyciu aparatów rentgenowskich. Ponieważ różne tkanki organizmu pochłaniają promieniowanie rentgenowskie w różnym stopniu, zależnie od ich składu chemicznego, na ekranie pokrytym substancją fluoryzującą pod wpływem promieni X powstaje obraz o różnym kontraście (rys. 4). Zamiast ekranu umieszcza się kliszę światłoczułą pomiędzy dwiema wzmacniającymi foliami, które świecą pod wpływem promieniowania X i dodatkowo naświetlają emulsję kliszy fotograficznej. Zdjęcia rentgenowskie są negatywami i są zawsze w tej postaci interpretowane przez lekarzy.



Rys. 4. Schemat otrzymywania obrazu rentgenowskiego na przykładzie prześwietlenia podudzia. Zmiana natężenia wiązki po przejściu przez tkanki zależy od ich rodzaju

seriografy, tomografy

Różne typy aparatów rentgenowskich przeznaczone są do różnych celów. I tak seriografy stosuje się do wykonywania wielu zdjęć w krótkim czasie, tomografy służą do wykonywania zdjęć warstwowych, wreszcie kamery do zdjęć małoobrazkowych pozwalają na szybkie przeprowadzenie masowych badań ludności. Obecnie coraz częściej stosuje się układy rentgenotelewizji złożone z aparatu rentgenowskiego, elektronowego wzmacniacza obrazu, kamery TV i monitora TV (rys. 5 i il. 25, tabl. 8). Dzięki ogromnej czułości wzmacniacza obrazu (kilka tys. razy większej niż czułość zwykłego ekranu rentgenowskiego) można znacznie obniżyć dawkę promieniowania pochłanianą przez pacjenta. Przez zastosowanie kserografii zwiększa się kontrast i rozdzielczość obrazów rentgenowskich, a ponadto unika się kłopotliwej obróbki klisz w ciemni fotograficznej.

rentgenotelewizja

tomografia rentgenowska

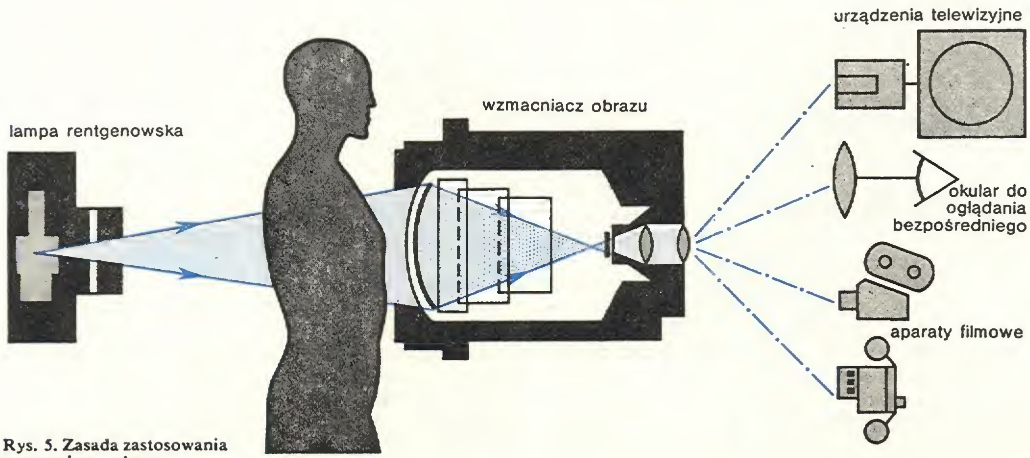
Najnowszym osiągnięciem w dziedzinie diagnostyki rentgenowskiej są tomografy komputerowe. Metoda konwencjonalnej tomografii rentgenowskiej (do zdjęć warstwowych) polega na tym, że lampa rentgenowska przesuwa się w trakcie naświetlania w jednym kierunku, a sprzężona z nią kaseta z kliszą fotograficzną przesuwa się w kierunku przeciwnym (rys. 8).

otrzymywanie obrazów

Jeżeli weźmiemy pod uwagę punkty A i B leżące w ciele pacjenta w płaszczyźnie obrotu układu lampa–kaseta PP_1, to rzutami tych punktów na kliszy w jej pierwszym położeniu będą punkty A_1 i B_1, w drugim zaś — A_2 i B_2. Wypadną one w tym samym miejscu kliszy i dlatego po jej wywołaniu będą widoczne jako ostre, natomiast rzuty każdego punktu leżącego poza płaszczyzną obrotu PP_1 (rys. 8) wypadną w różnych miejscach kliszy (punkty C_1 i C_2), nawet znajdą się poza kliszą (punkt C_3). Tak więc obraz tkanek leżących w płaszczyźnie obrotu układu lampa–kaseta jest ostry na zamazanym tle pochodzącym z nałożonych na siebie rzutów obrazów z innych płaszczyzn. Przez zmianę odległości lampa–kaseta zmienia się płaszczyznę ostrego obrazu, a zatem uzyskuje się możliwość oglądania kolejnych warstw tkanek. Jedna klisza rentgenowska odpowiada tylko jednej płaszczyźnie, a zatem uzyskanie pełnego, trójwymiarowego obrazu tkanek byłoby możliwe jedynie po złożeniu jedna na drugiej wszystkich klisz, na których leżą rzuty wszystkich płaszczyzn, co jest technicznie niewykonalne.

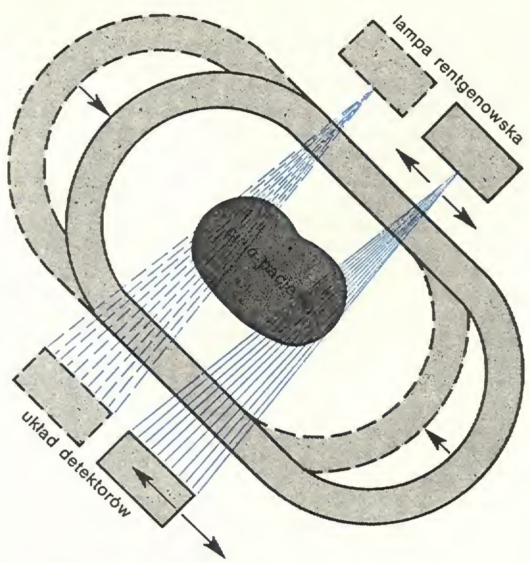
tomograf komputerowy

Wykorzystując minikomputer, udało się jednak zastosować samą ideę superpozycji rzutów tomograficznych do zbudowania nowego urządzenia, zwanego tomografem komputerowym. Nastąpił przewrót w metodach diagnostycznych medycyny, jak sądzą niektórzy — na miarę odkrycia samych promieni rentgenowskich. Twórcą pierwszego tomografu kom-



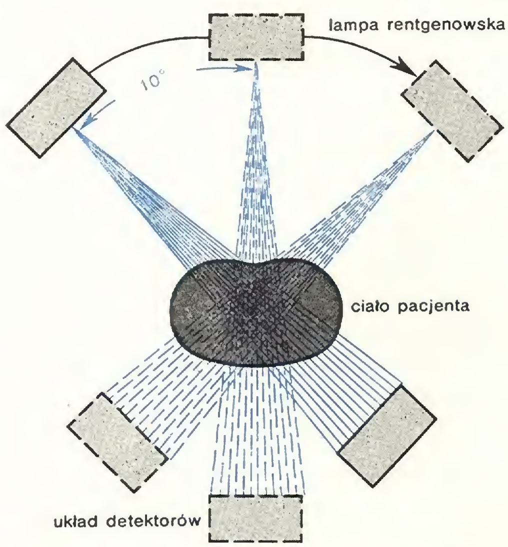
Rys. 5. Zasada zastosowania wzmacniacza obrazu

puterowego (EMI Brain Scanner) był Anglik, Godfrey N. Hounsfield, z firmy EMI (nagroda Nobla 1979 r.). Zastosowany w praktyce lekarskiej w 1973 r. tomograf służył przede wszystkim do badania mózgu.



Rys. 6. Przesunięcie liniowe aparatu tomograficznego

Tomograf komputerowy składa się z urządzenia przesuwającego (z lampą rentgenowską i układem detektorów; rys. 6 i 7), z konsoli rentgenowskiej



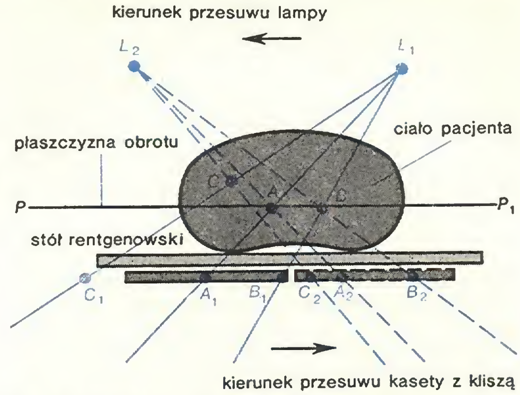
Rys. 7. Po fazie przesunięć liniowych aparat tomograficzny obraca się o 10°

budowa tomografu komputerowego

układu monitorowania i regulacji oraz z zespołu pamięci (dyskowej lub taśm magnetycznych) i ekranu TV, służących do rejestracji i oglądania otrzymanych obrazów. Lecz najważniejszą rolę w tomografie odgrywa szybki minikomputer, który stanowi serce całego urządzenia.

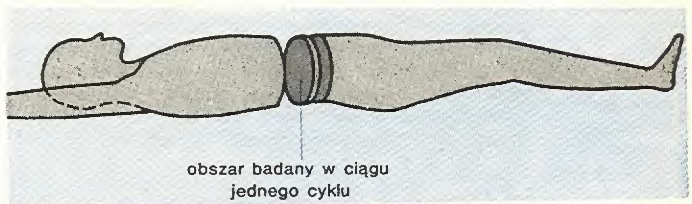
W takim tomografie bada się tkanki ciała w plastrach grubości 13 mm, wyciętych bezkrwawo, prostopadle do pionowej osi ciała (rys. 9). Dzięki zastosowaniu minikomputera można przeprowadzić analizę niewielkich nawet różnic w natężeniu promieniowania po przejściu przez tkanki, a więc i współczynników pochłaniania promieni rentgenowskich; to zaś z kolei daje możliwość odróżnienia tkanek niewiele różniących się gęstością.

Pacjenta umieszcza się na kozetce przechodzącej przez otwór tzw. bramki skaningowej urządzenia.



Rys. 8. Schemat otrzymywania rzutów tomograficznych

Kozetka może się przemieszczać wzdłuż osi poziomej przechodzącej przez bramkę. W bramce skaningowej znajduje się źródło promieni rentgenowskich i układ detektorów scyntylacyjnych. Źródło i detektory ustawione są naprzeciw siebie, a pomiędzy nimi znajduje się badana część ciała pacjenta. Dodatkowy detektor mierzy natężenie pierwotnej wiązki promieni rentgenowskich (il. 182, tabl. 47).



Rys. 9. Metodą tomograficzną bada się kolejne plastry ciała pacjenta

badanie za pomocą tomografu komputerowego

Badanie rozpoczyna się od fazy, w której bramka przesuwa się względem ciała pacjenta, przy czym wąska wiązka skolimowanych promieni rentgenowskich przechodzi przez ciało i pada na układ detektorów. W czasie jednego przesunięcia otrzymuje się jeden rzut tomograficzny z zarejestrowanych 18 000 pomiarów natężenia promieniowania. Po zakończeniu przesuwu bramka skaningowa obraca się względem ciała pacjenta o 10°, następnie powtarza się przesuw z jednoczesnym pomiarem i obrót bramki o 10°. I tak 18 razy, aż bramka zatoczy pół okręgu, tzn. wykona obrót o 180°. Wartości liczbowe dla każdego rzutu tomograficznego są w sposób ciągły wprowadzane do minikomputera, który oblicza łącznie 80 000 wartości współczynników pochłaniania dla badanego obszaru ciała i przekazuje do pamięci dyskowej, która może przechowywać wyniki ośmiu obrazów — ośmiu plastrów ciała.

Istota tomografii komputerowej polega na wykonywaniu dokładnych obliczeń — za pomocą minikomputera — współczynników pochłaniania promieni rentgenowskich dla każdego rzutu tomograficznego (co odpowiada na zdjęciu obszarom o różnym zaczernieniu), a następnie rekonstrukcji matematycznej całkowitego obrazu plastra z tych rzutów. Współczynniki pochłaniania obliczane są z dużą dokładnością. Pozwala to na uwidocznienie tkanek o małych różnicach gęstości, co tradycyjnymi metodami radiodiagnostyki było nieosiągalne. Układ wyjściowy z kolorowym monitorem telewizyjnym pozwala na oglądanie obrazów badanych plastrów ciała i rejestrację ich na kliszy (il. 22, tabl. 8 i il. 183, tabl. 47).

Czas badania jednego plastra wynosi 20 s (bada się 8 plastrów). Średni całkowity czas badania pacjenta wynosi ok. 30 min, w co wlicza się czas potrzebny na pracę minikomputera.

Ten nowy system przedstawiania obrazów rentgenowskich wymaga od lekarza radiologa innego sposobu interpretacji, gdyż większość zdjęć wykonywano dotychczas w płaszczyznach równoległych do pionowej osi ciała. Chociaż w tradycyjnej rentgenodiagnostyce (szczególnie w kseroradiografii) można odróżnić szczegóły budowy tkanki kostnej z dokładnością rzędu mikronów, to jednak dopiero w tomografii komputerowej uzyskuje się obrazy miękkich tkanek ciała, a więc większości wewnętrznych organów człowieka.

możliwości zastosowania tomografii komputerowej

Zakres medycznych zastosowań tomografii komputerowej jest ogromny i rozszerza się z każdym rokiem: począwszy od diagnostyki nowotworów czy zatorów w mózgu bez konieczności wprowadzania środka cieniującego lub powietrza (co się zawsze wiąże z niewygodą lub nawet niebezpieczeństwem dla pacjenta), poprzez badania profilaktyczne, sprawdzanie wyników leczenia — aż po badanie dynamiki organów (np. serca) — wszędzie ta metoda znajduje zastosowanie.

Obecnie prowadzi się również prace nad zastosowaniem tomografii komputerowej w dziedzinie krótkożyciowych radioizotopów pozytonowych, takich jak ^{11}C, ^{13}N czy ^{15}O, albo ciężkich jonów. Można sobie również wyobrazić zastosowanie komputerowej metody tomograficznej w technice sonografii ultradźwiękowej lub w technice pól magnetycznych.

terapia promieniami *X*

Naczelną troską konstruktorów diagnostycznych aparatów rentgenowskich jest stałe dążenie do zmniejszania dawki promieniowania związanej z badaniem. Natomiast w aparatach rentgenowskich stosowanych w terapii, przede wszystkim nowotworowej, istotne jest otrzymywanie możliwie silnych, jednorodnych i stabilnych wiązek promieni *X*, terapia bowiem polega na niszczeniu tkanki nowotworowej przy jak najmniejszym uszkodzeniu tkanki zdrowej. Tak rozumiane leczenie promieniami *X* wymaga opracowania szczegółowego programu napromieniowania.

Z fizycznego punktu widzenia terapię rentgenowską podzielić można na 4 działy: do ok. 100 kV (terapia powierzchniowa), do ok. 150 kV (terapia półgłęboka), do 500 kV (terapia głęboka) oraz od ok. 1 MV (terapia megawoltowa). Inne warunki muszą być spełnione, gdy chodzi o napromienianie skóry, a inne, gdy chodzi o tkanki nowotworowe znajdujące się głęboko w ciele chorego. Program napromieniania opracowuje lekarz radioterapeuta przy współpracy z fizykiem medycznym, o czym poniżej.

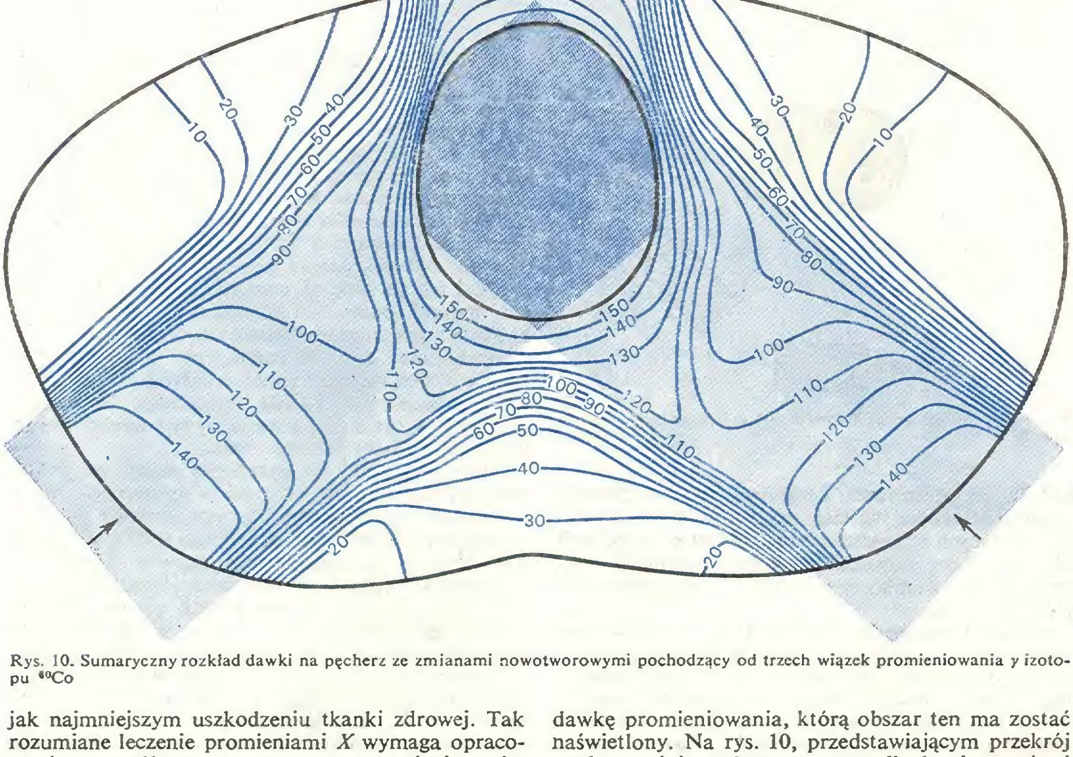
W rentgenoterapii stosuje się lampy rentgenowskie nieruchome (terapia stacjonarna) lub lampy, które wykonują ruch obrotowy (terapia obrotowa).

terapia promieniami γ

Oprócz terapii rentgenowskiej w leczeniu nowotworów stosuje się zewnętrzną i wewnętrzną terapię promieniami gamma. Dawka promieniowania gamma pochłonięta przez tkanki kostne jest mniejsza niż dawka promieniowania rentgenowskiego. Ponadto stosowanie promieni gamma pozwala na korzystniejsze napromieniowanie tkanek nowotworowych położonych głębiej w ciele przy jednoczesnym oszczędzaniu skóry. Wewnętrzna terapia gamma polega na stosowaniu tzw. aplikatorów w postaci igieł, walców itp. zawierających rad, kobalt-60, iryd-192 i cez-137, wprowadzanych śródtkankowo lub śródjamowo do organizmu. Zewnętrzna zaś terapia polega na stosowania aparatu do teleterapii Co-60, tzw. bomby kobaltowej. Jest to urządzenie zawierające ładunek aktywności rzędu 10^{14} Bq, umożliwiające krótkie i intensywne naświetlanie tkanki nowotworowej. Problemy planowania teleterapii są w tym wypadku podobne do planowania terapii rentgenowskiej. Bomby kobaltowe produkowane przez firmę Siemens nazywają się gammatronami (il. 52, tabl. 14).

zastosowanie bomby kobaltowej

Rozważmy tu przykładowo zagadnienie planowania kuracji — za pomocą bomby kobaltowej (Co-60) — pacjenta, u którego w pęcherzu wykryto zmiany nowotworowe. Po wykryciu tych zmian na zdjęciu rentgenowskim lekarz wskazuje fizykowi obszar odpowiadający anatomicznie pęcherzowi i całkowitą dawkę promieniowania, którą obszar ten ma zostać naświetlony. Na rys. 10, przedstawiającym przekrój w płaszczyźnie pęcherza prostopadły do pionowej osi



Rys. 10. Sumaryczny rozkład dawki na pęcherz ze zmianami nowotworowymi pochodzący od trzech wiązek promieniowania γ izotopu ^{60}Co

ciała pacjenta, obszar ten obwiedzony jest czarną linią. Rola fizyka medycznego polega na takim dobraniu warunków napromieniowania, aby maksimum dawki przypadło w tym właśnie obszarze, pozostałe zaś części ciała zostały jak najmniej napromieniowane. W wypadku przedstawionym na rysunku fizyk medyczny dobrał napromieniowanie nowotworu trzema wiązkami promieni gamma z trzech kierunków. Linie koloru niebieskiego są to tzw. izodozy, tj. krzywe łączące punkty o tej samej wartości dawki (podanej w procentach; dawkę odpowiadającą wiązce padającej od przodu przyjęto jako 100%). Izodozy te powstały z sumowania trzech rozkładów pierwotnych izodoz, które dotyczyły trzech kierunków oddzielnie, a które otrzymano z badań na fantomach — sztucznych układach własnościami imitujących ciało człowieka. Wiązki promieni gamma mają w tym przykładzie „średnicę" 8 cm i tego rzędu zasięg najefektywniejszego działania.

izodozy

fantom

Przygotowanie rozkładu izodoz, tzn. dobranie wszystkich warunków, takich jak odległości, wielkości dawek, liczba kierunków napromieniowania, zastosowanie specjalnych filtrów formujących wiązki itp., jest procedurą żmudną i pracochłonną, i dlatego powierza się je obecnie coraz częściej odpowiednio zaprogramowanym komputerom. Programy na komputer są oczywiście pisane również przez fizyków medycznych.

zastosowanie akceleratorów

W niewiele lat po skonstruowaniu pierwszych akceleratorów typu Van de Graaffa, betatronów i akceleratorów liniowych do celów badań fizycznych okazało się, że mogą one być z pożytkiem wykorzystane w radioterapii. Obecnie jest na świecie przeszło 400 betatronów i akceleratorów liniowych stosowanych w medycynie. Liczba pacjentów leczonych sięga miliona rocznie, a koszt leczenia — wieluset milionów dolarów.

Akceleratory służą do otrzymywania wiązek cząstek naładowanych, np. elektronów, protonów, lub pośrednio do otrzymywania wiązek cząstek nie naładowanych, jak neutrony i fotony. W radioterapii stawia się specjalne wymagania: energia cząstek musi być na tyle duża, aby dotarły one do obszaru nowotworu, natężenie wiązki musi być tak duże, żeby okres napromieniowania pacjenta był jak najkrótszy, sama zaś wiązka musi być jednorodna i łatwa do regulowania i kontroli. Wreszcie zależność wielkości dawki od grubości tkanki musi się charakteryzować szybkim narastaniem, możliwie płaskim maksimum i szybkim spadkiem (rys 1).

W radioterapii wykorzystuje się najczęściej elektrony o energiach do 40 MeV oraz promienie *X* od 8 do 20 MeV. Stosuje się również neutrony prędkie, produkowane najczęściej w cyklotronach izochronicznych. Wreszcie w celach leczniczych mogą być stosowane protony, deuterony, a nawet piony. Stosując protony o energii przekraczającej 100 MeV i technikę obrotową, otrzymuje się wyjątkowo korzystne warunki napromieniowania głęboko położonej tkanki nowotworowej przy małej dawce powierzchniowej. W radioterapii można również stosować wysokoenergetyczne ciężkie jony, dające dobry rozkład dawki, np. w wypadku nowotworów położonych głęboko.

Medycyna nuklearna

Uważa się powszechnie, że medycynę nuklearną można określić jako dziedzinę, która obejmuje wszystkie rodzaje zastosowania źródeł promieniowania jonizującego w postaci substancji radioaktywnych (promieniotwórczych) w diagnozie i leczeniu oraz w pracy badawczej z wyjątkiem zastosowania zamkniętych źródeł promieniowania, o których była mowa w części poświęconej radioterapii. Medycyna nuklearna szybko się rozwija. Wprowadza się bezustannie nowe radiofarmaceutyki, nowe techniki pomiaru i nowe testy diagnostyczne.

znaczniki

Metody diagnostyczne medycyny jądrowej — metody radioizotopowe, — opierają się na zastosowaniu tzw. znaczników. Atomy radioaktywnego izotopu danego pierwiastka wprowadzone do organizmu zachowują się jak atomy pierwiastka trwałego i uczestniczą we wszystkich procesach metabolicznych na podobieństwo agentów policji zmieszanych z tłumem demonstrantów i wysyłających radiowe raporty o sytuacji. Tak samo atomy radioaktywne, wysyłając promieniowanie gamma (lub elektrony), które mogą być wykryte za pomocą detektorów promieniowania, dają nam znać o przebiegu procesów w organizmie, tkance czy komórce. Ponieważ liczba znaczników radioaktywnych jest znikomo mała w porównaniu z odpowiednimi atomami trwałymi, efekt zakłócający procesy biologiczne, jaki mogłoby wywrzeć promieniowanie wysyłane przez znaczniki, jest w większości wypadków również znikomy. Nie można jednak pominąć efektu niszczącego tkankę i dlatego podawanie jakichkolwiek substancji radioaktywnych do organizmu musi być obwarowane odpowiednimi przepisami i zastrzeżeniami. Wszelkie substancje promieniotwórcze, tzw. radiofarmaceutyki, wprowadzane są do żywego organizmu pod kontrolą lekarską. Celowo nie używamy terminu „radiolek", gdyż substancje promieniotwórcze stosowane są w ogromnej większości do celów diagnostycznych, a nie leczniczych. Stosuje się je, aby zbadać funkcję określonego narządu w organizmie lub, aby uzyskać obraz struktury narządu. Muszą je charakteryzować cztery zasadnicze cechy: odpowiednie właściwości biologiczne, rodzaj i szybkość rozpadu radioizotopu, właściwości umożliwiające detekcję i łatwość produkcji.

radiofarmaceutyki

aktywność promieniotwórcza; bekerel

Ilość radioizotopu albo raczej proporcjonalną do niej aktywność promieniotwórczą mierzy się liczbą rozpadów na jednostkę czasu. Jednostką aktywności jest bekerel — Bq; 1 Bq jest to aktywność, przy której następuje 1 rozpad w czasie 1 sekundy (poprzednio stosowaną jednostką był kiur — Cj; $1\ Ci = 3{,}7 \cdot 10^{10}$ Bq). Można w zasadzie przyjąć, że ilości radioizotopu w radiofarmaceutykach stosowanych diagnostycznie są rzędu 10^4 Bq, stosowanych zaś w radioterapii — rzędu $(0{,}1–1) \cdot 10^9$ Bq.

Charakterystyczną cechą każdego radioizotopu jest jego fizyczny czas połowicznego zaniku $T_{1/2}$. Ze względu na szkodliwość radioizotopów dla organizmu oraz wymagania związane z detekcją radioizotopy medyczne podawane w postaci radiofarmaceutyków powinny mieć czas połowicznego zaniku $T_{1/2}$ nie krótszy niż 1 min i nie dłuższy niż 100 dni; energie emitowanych przez nie promieni beta powinny być jak najniższe, a energie promieni gamma nie większe niż 1 MeV.

biologiczny czas połowicznego zaniku

Każdą substancję podawaną żywemu organizmowi charakteryzuje czas, po którym organizm wydala połowę podanej ilości. Jest to tzw. biologiczny czas połowicznego zaniku $T^b_{1/2}$. W rezultacie obserwowany czas połowicznego zaniku, tzw. efektywny czas połowicznego zaniku podawanego radiofarmaceutyka jest mniejszy od $T_{1/2}$ w stopniu zależnym od $T^b_{1/2}$. Znajomość — z obserwacji — czasu efektywnego pozwala łącznie z danymi o zdolności jonizacyjnej radioizotopu na obliczenie dawki pochłoniętej przez określony narząd czy tkankę. Obliczenia takie są skomplikowane, wymagają użycia komputerów, ale są niezwykle ważne w medycynie nuklearnej, gdyż dzięki nim można wyznaczyć efekt szkodliwości podawania określonego radioizotopu. Ponadto radioizotopy znajdują zastosowanie w badaniach metabolicznych (tabela).

otrzymywanie radioizotopów

Większość radioizotopów stosowanych w medycynie otrzymuje się w reaktorze, niektóre w cyklotronie (^{123}J, ^{52}Fe, ^{18}F, ^{11}C, ^{13}N i ^{15}O). Obecnie coraz szerzej stosowane są radioizotopy krótkożyjące ($T_{1/2} = 6$ h, takie jak ^{99m}Tc czy ^{113}In). Otrzymuje się je bezpośrednio w pracowni medycznej z rozpadu dłużej żyjących radioizotopów macierzystych umieszczonych w tzw.

zastosowanie radioizotopów

Metody diagnostyczne i terapeutyczne w medycynie jądrowej

Diagnostyka	Terapia
Badanie czynności tarczycy Scyntygrafia tarczycy Badanie czynności wątroby Scyntygrafia wątroby Scyntygrafia mózgu Badanie czynności nerek Scyntygrafia nerek Scyntygrafia ognisk nowotworowych Scyntygrafia układu kostnego Scyntygrafia mięśnia sercowego Diagnostyka chorób układu krążenia Oznaczanie objętości krwi krążącej Diagnostyka zaburzeń ukrwienia Badanie metabolizmu, np. żelaza Badanie gospodarki wodnej	**Zastosowanie źródeł otwartych** Leczenie chorób tarczycy Leczenie chorób krwi Leczenie przerzutów nowotworowych **Zastosowanie źródeł zamkniętych** Leczenie kontaktowe i śródmiąższowe Wprowadzenie do przysadki mózgowej Leczenie nowotworów skóry za pomocą izotopów β^--promieniotwórczych

radioizotopy w medycynie

Radioizotopy najczęściej stosowane w medycynie nuklearnej i ich właściwości

Izotop	$T_{1/2}$	Energia β	Energia γ	Radiofarmaceutyka	Zastosowanie
^{14}C	5700 lat	bardzo niska	—	liczne związki organiczne	badania biomedyczne
^{60}Co	5,2 lat	średnia	wysoka	źródło zamknięte	leczenie nowotworów
^{85}Kr	10,4 lat (bardzo krótki $T_{1/2b}$)	—	średnia	gaz	przepływ krwi, zaburzenia oddychania
^{99}To	6 h	—	średnia	nadtechnecjan, związki znakowane	scyntygrafia tarczycy, nowotworów
^{113}I	1,7 h	—	wysoka	związki chelatowe	scyntygrafia nowotworow
^{125}J	60 d	niska	niska	jodek sodu, liczne związki organiczne (hormony)	badania gruczołów wydzielania dokrewnego (przeważnie *in vitro*)
^{131}J	8,1 d	średnia	średnia	jodek sodu, albumina, hipuran, hormony tarczycy	diagnostyka tarczycy, nerek, przepływ krwi (przeważnie *in vivo*)
^{198}Au	2,7 d	wysoka	średnia	zawiesina koloidalna	terapia nowotworowa
^{197}Hg	67 h	niska	niska	neohydryna	scyntygrafia nerek
^{203}Hg	47 d	średnia	średnia	neohydryna	scyntygrafia nerek

energia mała: < 100 keV
energia średnia: 100 keV $-$500 keV
energia duża: > 500 keV

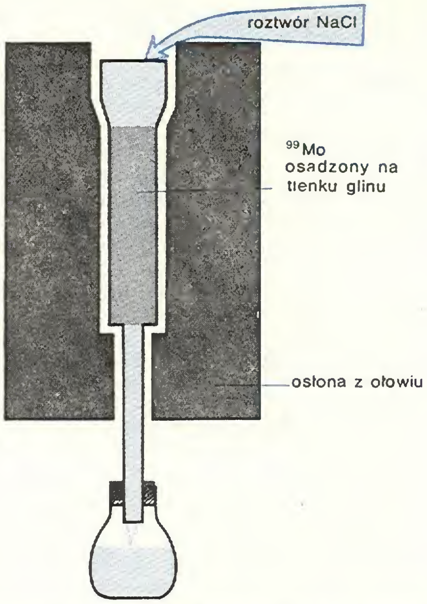
generatory

generatorach (rys. 11). Na przykład ^{99m}Tc otrzymuje się z generatora ^{99}Mo o czasie połowicznego zaniku 67 h. Radioizotopy te znajdują szczególnie szerokie zastosowanie w scyntygrafii.

Diagnostyczne metody radioizotopowe pozwalają na postawienie lepszej diagnozy niż metody nieradioizotopowe, gdyż umożliwiają otrzymywanie danych pewniejszych, bardziej szczegółowych, bardziej obiektywnych, bo ilościowych; ponadto umożliwiają wprowadzenie badań masowych (np. w diagnostyce nowotworowej) i zmniejszają koszt badań.

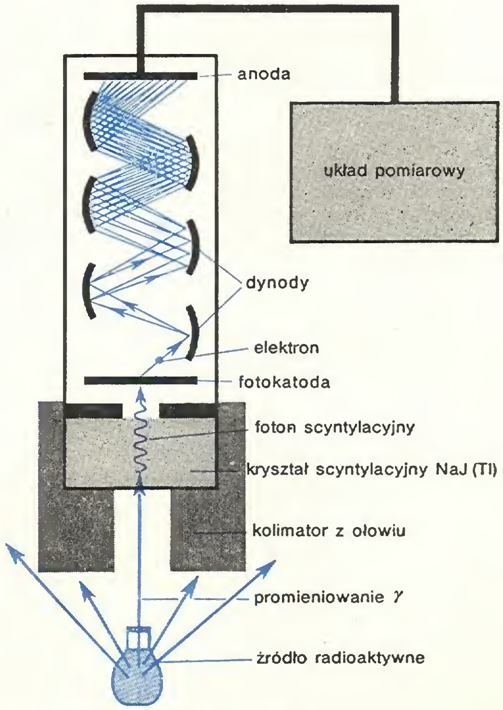
liczniki scyntylacyjne

Podstawowym przyrządem mierzącym ilość promieniowania (przeważnie gamma) ze znacznika krążącego w ustroju lub zawartego w próbce krwi pobranej od pacjenta jest licznik scyntylacyjny. Składa się on ze scyntylatora krystalicznego (NaJ aktywowany talem) i fotopowielacza połączonych ze sobą (rys. 12). Wiązka promieni gamma, ograniczona przez ołowiany kolimator, pada na scyntylator i powoduje powstanie w nim krótkotrwałych błysków świetlnych, tzw. scyntylacji. Światło tych scyntylacji, padając na światłoczułą katodę fotopowielacza, uwalnia z niej



Rys. 11. Kolumna do otrzymywania radioizotopów krótkożyjących. Izotop technetu ^{99m}Tc otrzymuje się np. wymywając roztworem chlorku sodu izotop molibdenu ^{99}Mo osadzony na kolumnie szklanej z tlenkiem glinu. Proces ten można powtarzać co ok. 24 h aż do spadku aktywności ^{99}Mo

strumień elektronów. Strumień ten, zwielokrotniony w fotopowielaczu, tworzy na wyjściu prąd elektryczny w postaci impulsów przekazywanych dalej do układów wzmacniających, kształtujących i liczących.



Rys. 12. Schemat sondy scyntylacyjnej

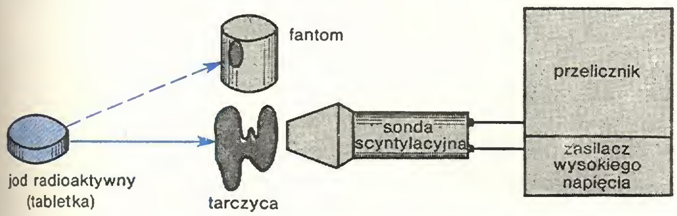
Liczniki scyntylacyjne mają największą ze wszystkich typów detektorów wydajność liczenia promieni gamma; ich kształty i rozmiary mogą być różne — od wąskich sond, średnicy ułamka centymetra, do wielkich detektorów średnicy kilkudziesięciu centymetrów.

metody *in vivo* i *in vitro*

Metody pomiarowe stosowane w medycynie dzielą się na metody *in vivo* (w żywym organizmie) i *in vitro* (poza organizmem, w probówce). Pomiary *in vivo* obejmują pomiary czynnościowe i lokalizacyjne (scyntygrafia). Obecną tendencją jest niewątpliwie przechodzenie od metod *in vivo* do metod *in vitro* (np. testy radioimmunologiczne, polegające na pomiarach pewnych parametrów biochemicznych we krwi z zastosowaniem znaczników „wbudowanych" w radiofarmaceutyk). Pomiary prowadzone *in vivo*, z uwagi na zmienną geometrię, zależną od pacjenta, jak i na wiele innych trudności związanych z charakterem promieniowania przechodzącego przez tkanki, są bardziej kłopotliwe i mniej dokładne. W pomiarach *in vitro* próbki krwi czy innych płynów ustrojowych, które zostały pobrane od pacjenta otrzymującego uprzednio radiofarmaceutyk lub do których radiofarmaceutyk został podany pozaustrojowo, mierzy się w scyntylatorze zwanym studniowym, w stałych warunkach geometrycznych.

badanie tarczycy

Typowym przykładem pomiarów czynnościowych *in vivo* może być test polegający na pomiarze jodochwytności tarczycy. Tarczyca jest gruczołem, który jod z pożywienia „wbudowuje" w hormon wydzielany do krwi (jod odgrywa bardzo ważną rolę w wielu procesach organizmu). Miarą czynności tarczycy jest m.in. ilość jodu, która zostaje wbudowana na jednostkę czasu. Otóż czynność tę można określić mierząc jaki procent podanego radiofarmaceutyka (np. jodek sodu $Na^{131}J$ w ilości ok. $4 \cdot 10^5$ Bq) znajduje się w tarczycy po upływie np. 24 godzin od podania (rys. 13). Stosując inne radiofarmaceutyki można określić działanie nerek, zmierzyć szybkość przepływu krwi krążącej, szybkość produkcji witaminy B_{12} w szpiku itp. (il. 184, 185, 187, 188 i tabl. 48).



Rys. 13. Zasada pomiaru jodochwytności tarczy. Po podaniu pacjentowi diagnostycznej dawki $Na^{131}J$ (w postaci tabletki) mierzy się liczbę impulsów nad tarczycą pacjenta (I_p) oraz — w takich samych warunkach geometrycznych — nad próbką $Na^{131}J$ o aktywności równej podanej, umieszczoną w sztucznej tarczycy, tzw. fantomie (I_f). Procent jodochwytności oblicza się z ilorazu I_p/I_f

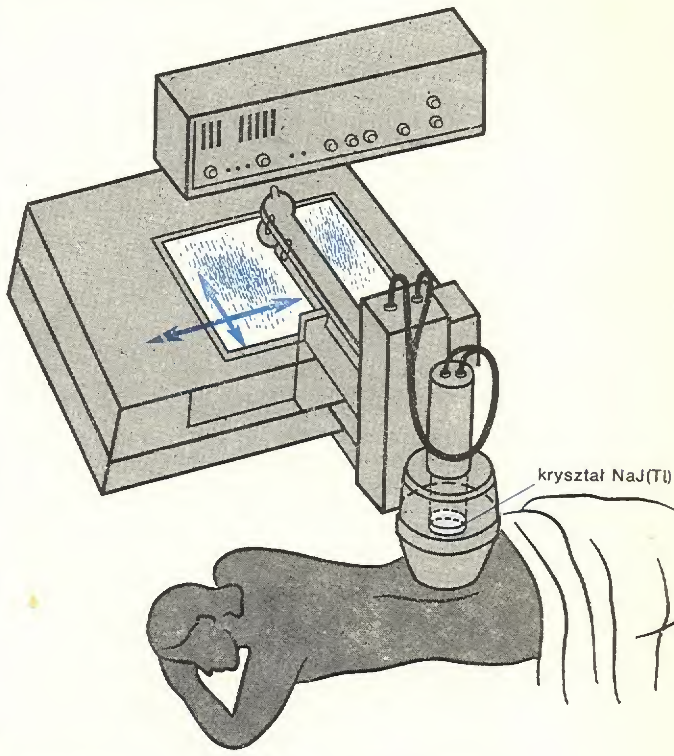
scyntygrafia radioizotopowa

Bardzo rozległą dziedziną metod określania radioaktywności *in vivo* jest scyntygrafia radioizotopowa. Umożliwia ona przedstawienie rozmieszczenia radioizotopów w tkankach ustroju w postaci dogodnej do interpretacji za pomocą wzroku lub komputerów.

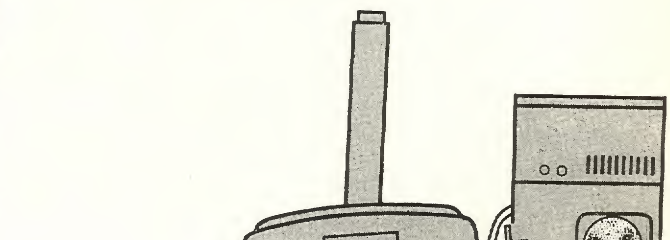
W diagnostyce zadaniem scyntygrafii jest wizualne przedstawienie badanego obszaru, tj. narządu lub tkanki patologicznej u pacjenta, w celu stwierdzenia zmian patologicznych lub określenia położenia, wielkości i kształtu tego obszaru oraz względnej ilości wychwytanego przez niego radiofarmaceutyka. Te raczej skomplikowane zadania spełniają aparaty scyntygraficzne; są to urządzenia detekcyjno-zapisujące, które przyjmują sygnały (w tym wypadku promieniowanie gamma), przetwarzają je, a następnie przedstawiają je jako obrazy płaskie, złożone z kresek, plamek lub liczb. Wszystkie aparaty scyntygraficzne można podzielić na dwa typy: scyntygrafy, zwane inaczej skenerami, oraz scyntykamery.

scyntygrafy

W scyntygrafie (rys. 14) głowica detektora przesuwa się ruchem meandrowym nad obszarem badanym. Głowica wyposażona jest w kolimator, będący soczewką zbierającą promienie gamma w danej chwili z niewielkiego obszaru. Na wyjściu otrzymuje się obraz o różnym zaczernieniu, który stanowi rzut płaski obszaru zawierającego radioizotop o różnej koncentracji (il. 187, tabl. 48).



Rys. 14. Scyntygraf — ruchomy aparat scyntygraficzny. Scyntygram kreskowy otrzymuje się za pomocą pisaka na papierze umieszczonym na stoliku scyntygrafu



Rys. 15. Scyntykamera — nieruchomy aparat scyntygraficzny. Scyntygramy otrzymuje się na ekranie oscyloskopu po opracowaniu przez komputer

scynty-kamery Scyntykamera (rys. 15) wyposażona jest w nieruchomy detektor (w postaci dużego i płaskiego scyntylatora), w którym powstaje scyntylacyjny obraz rozmieszczenia radioizotopu w tkance. Obraz ten zostaje elektronicznie zanalizowany i przedstawiony na ekranie oscyloskopu (il. 188, tabl. 48).

Przykładowo na il. 184 (tabl. 48) podane są dwa scyntygramy: jeden — fantomu tarczycy po podaniu Na ^{131}J — wykonany scyntygrafem, drugi — po podaniu ^{99m}Tc — wykonany scyntykamerą.

liczniki całego ciała Ze stosowaniem komputerów do ilościowej oceny scyntygramów i z dążeniem do rejestracji przebiegającego w czasie procesu rozchodzenia się znaczników w ciele wiąże się w pewnym stopniu inna ważna metoda badań radioizotopowych *in vivo*, która polega na stosowaniu liczników całego ciała. Są to niezwykle czułe mierniki wykrywające znikome ilości radioizotopów w ciele. Znajdują one zastosowanie w ochronie przed promieniowaniem (określanie wewnętrznych skażeń promieniotwórczych) oraz w badaniu niektórych długotrwałych przemian w organizmie.

źródła otwarte w terapii Zastosowanie terapeutyczne radioizotopów jako źródeł otwartych nie jest tak częste jak zastosowanie diagnostyczne; niektóre metody podane są w tabeli. Warto tu wspomnieć przykładowo o metodzie bezkrwawej operacji tarczycy za pomocą podawania doustnie jodku sodu Na ^{131}J w ilości (40–80) · 10^{6} Bq. Radiofarmaceutyk ten zostaje wychwycony przez tarczycę i w tak dużej ilości (1000 razy większej od ilości podawanej diagnostycznie) powoduje zniszczenie nadmiernie rozrośniętej tkanki tarczycy w chorobie Gravesa–Basedowa.

zadania fizyka medycznego Zakres pracy fizyka medycznego w medycynie nuklearnej jest szeroki. Zajmuje się on kontrolą bezpieczeństwa i higieny pracy, co ma szczególne znaczenie z uwagi na pracę z otwartymi źródłami promieniowania oraz stosowanie krótkożyjących radioizotopów. Otrzymywanie tych izotopów z generatorów wymaga pomiaru ich aktywności, dozowania i podawania ich w warunkach sterylnych i nie powodujących reakcji ubocznych. Poza tym konieczna jes współpraca fizyka medycznego z lekarzem przy opracowywaniu i prowadzeniu pomiarów *in vivo* i *in vitro*. Szczególnie pomiary *in vivo* wymagają starannego wzorcowania. Rozwój metod nowoczesnej scyntygrafii (kilka tysięcy prac publikowanych rocznie na całym świecie) zależy również w dużej mierze od współpracy lekarzy i fizyków medycznych.

Na zakończenie podkreślimy raz jeszcze, że obecna tendencja rozwojowa w dziedzinie zastosowania promieniowania jonizującego w medycynie polega na zmniejszaniu aktywności substancji podawanych pacjentowi w celach diagnostycznych, a nawet na przechodzeniu do badań stosujących radioizotopy pozaustrojowo, z drugiej zaś strony — na dalszym rozwoju potężnych urządzeń terapeutycznych wykorzystujących promieniowanie jonizujące, a także zastosowaniu innych niż wymienione w artykule źródeł promieniowania.

Promieniowanie niejonizujące w medycynie

Jak wspomniano na wstępie, dział fizyki medycznej obejmujący dziedziny zastosowania promieniowania niejonizującego staje się mniej więcej od 10 lat konkurentem działu promieniowania jonizującego. Z jednej strony jest to wynikiem nowych odkryć fizyki i nowych rozwiązań techniki, z drugiej zaś — tendencji (w diagnostyce medycznej) do stosowania metod jak najmniej szkodliwych dla pacjenta. Tak więc znaczenia nabierają metody stosujące promieniowanie laserowe, promieniowanie podczerwone (termografia) oraz fale ultradźwiękowe (→ Badanie ośrodków za pomocą ultradźwięków).

Promieniowanie laserowe

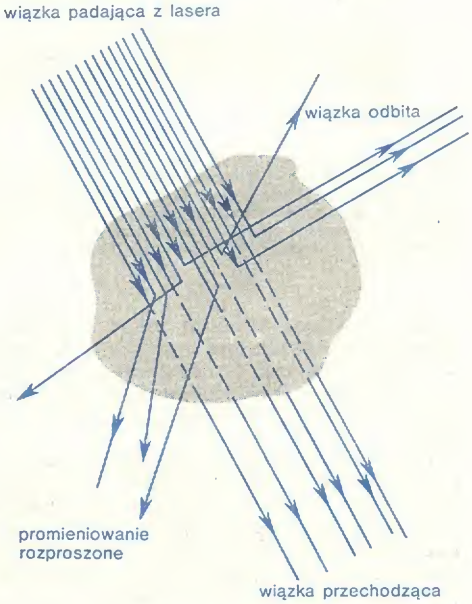
zalety promieniowania laserowego Promieniowanie laserowe charakteryzuje wysoki stopień spójności i monochromatyczności; można je wytwarzać w postaci silnie skolimowanych wiązek o średnicy nawet rzędu długości fali tego promieniowania. Ta druga cecha pozwala na ogromne zwiększenie gęstości mocy wiązki, co z kolei daje możliwość skupienia dużych ilości energii na minimalnym obszarze. Szczególne znaczenie mają w biomedycynie lasery molekularne ze względu na dużą wydajność, niski koszt wytwarzania i niewielkie rozmiary. Ponadto promieniowanie lasera molekularnego przypada na część podczerwoną widma, która jest silnie pochłaniana przez tkankę.

lasery w medycynie Lasery wykorzystuje się obecnie w medycynie do trzech zasadniczych celów: w terapii — do niszczenia tkanki metodą „cięcia" lub „odparowywania" (bezkrwawa chirurgia), w diagnostyce — jako źródło oświetlające oraz do celów badawczych w biomedycynie.

Terapeutyczne i diagnostyczne zastosowania laserów

Specjalność	Zastosowanie: terapia	Zastosowanie: diagnostyka
Okulistyka	koagulacja siatkówki mikrochirurgia	badanie zaćmy
Onkologia	niszczenie tkanki nowotworowej	holografia ultradźwiękowa, rentgenowska
Chirurgia	cięcie tkanek miękkich i twardych	oświetlanie narządów od wewnątrz (endoskopia)
Stomatologia	usuwanie próchnicy, plombowanie	—
Dermatologia	usuwanie tatuażu, procesy rozrostowe	—

Ogólnie biorąc, wiązka promieniowania laserowego oddziałuje na tkankę, jak to przedstawiono na rys. 16. Wiele efektów występujących przy odbiciu i rozproszeniu nie ma dla nas większego znaczenia, najważniejsze jest termiczne oddziaływanie promieniowania na tkankę, zasadniczo związane z pochłanianiem (nie



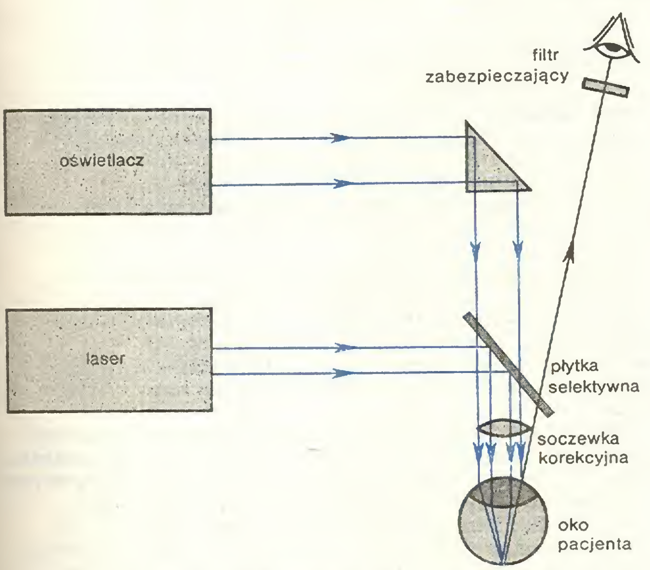
Rys. 16. Schemat oddziaływania wiązki laserowej z tkanką

należy jednak zapominać, że promieniowanie odbite lub rozproszone może powodować uszkodzenie tkanki). Oddziaływanie termiczne zależy przede wszystkim

od gęstości mocy promieniowania. Po przekroczeniu pewnej wartości gęstości występuje zjawisko odparowywania tkanki na obszarze, na który pada wiązka promieniowania. Warto podać, że do odparowania 1 g tkanki, której gęstość jest w przybliżeniu równa gęstości wody, potrzeba energii ok. 2000 J. Najtrudniejszą sprawą z praktycznego punktu widzenia jest dobranie odpowiednich ilości energii. Przy zbyt dużej gęstości mocy mogą powstawać małe eksplozje, utrudniające stosowanie lasera do cięcia tkanek, przy zbyt małej — mogą wystąpić dodatkowe komplikacje.

laser jako skalpel

Ostra wiązka laserowa stosowana jako skalpel chirurgiczny umożliwia przeprowadzanie czystych cięć w tkankach, a przez przypalanie rany — zmniejsza krwawienie. Takich bezkrwawych zabiegów można dokonywać na narządach silnie ukrwionych, jak wątroba, płuca czy mózg. Czyni się również próby użycia promieni laserowych do usuwania próchnicy zębów, a nawet do plombowania zębów. Najbardziej jednak rozpowszechnione jest zastosowanie wiązki laserowej w okulistyce, a mianowicie w mikrochirurgii ocznej do łączenia (koagulacji) odklejonej siatkówki z naczyniówką w oku ludzkim. Urządzenie



Rys. 17. Schemat działania głowicy koagulatora laserowego

koagulator laserowy

służące do tego zabiegu zwie się koagulatorem laserowym, zabieg zaś polega na tym, że wiązkę laserową kieruje się przez źrenicę tak, aby soczewka skupiła ją w miejscu, w którym ma powstać koagulacja; silny impuls świetlny wywołuje odczyn zapalny, w następstwie czego powstaje zrost, który „przykleja" siatkówkę do naczyniówki. Koagulator laserowy (rys. 17 i il. 16, tabl. 5) góruje nad uprzednio stosowanymi fotokoagulatorami krótkim czasem naświetlania i małą średnicą wiązki (o energii 10^{-3}–10^{-2} J).

technika „odparowywania"

Technika odparowywania rozleglejszych tkanek, wymagająca znacznie większych gęstości mocy niż cięcie czy koagulacja, znajduje zastosowanie przy niszczeniu tkanki nowotworowej przede wszystkim w miejscach na ciele pacjenta dostępnych do bezpośredniego naświetlania. Nie można jednakże ustalić jednej wartości dopuszczalnej dawki promieniowania laserowego. Istnieją zalecenia podające różne wartości, zależne od parametrów samych urządzeń laserowych, ale uważa się zawsze, iż najdelikatniejszym narządem jest oko ludzkie. Uznano, że dozwolone dawki promieniowania bezpośrednio padającego na źrenicę oka wahają się od $5\cdot10^{-8}$ do 10^{-6} J/cm², natomiast dawki promieniowania „odbitego" w 100% od powierzchni rozpraszającej wahają się od 0,07 do 0,9 J/cm². Jeśli chodzi o inne organy (skóra itp.), wartości dawek mogą być znacznie większe. Do napromieniowania narośli o średnicy kilku milimetrów wymagana jest energia kilku dżuli w impulsie; zniszczenie większych obszarów tkanki zdegenerowanej wymaga stosowania laserów dających wiązki o energii do 10 000 J. Zastosowanie światłowodów umożliwia niszczenie tkanki wewnątrz organizmu.

oświetlanie narządów

W niedalekiej przyszłości wiązka promieniowania laserowego może znaleźć zastosowanie w diagnostyce medycznej, a mianowicie do oświetlania narządów wewnętrznych. Wiąże się to również z zastosowaniem techniki przesyłania światła za pomocą światłowodów. Nawet holografia powinna znaleźć zastosowanie w medycynie, szczególnie z chwilą uzyskania spójnych wiązek promieni *X*. Już obecnie filtracja holograficzna w rentgenodiagnostyce poprawia czytelność zdjęć rentgenowskich. Czyni się próby wykorzystania w diagnostyce holografii ultradźwiękowej.

holografia w medycynie

Jeszcze inną dziedziną zastosowania promieniowania laserowego jest biomedycyna. Promieniowanie to pozwala na przeprowadzanie mikrooperacji wewnątrz ... pojedynczej komórki.

Nie należy jednakże zapominać o niebezpieczeństwach związanych z użyciem promieniowania laserowego. Nowe rodzaje jego zastosowania będą zatem wynikiem nowych rozwiązań w dziedzinie techniki laserowej i prac nad ustaleniem wpływu tego promieniowania na obiekty biologiczne.

zadania fizyka medycznego

Rola fizyka medycznego polega w tej dziedzinie przede wszystkim na udziale w projektowaniu urządzeń, które w diagnostyce lub terapeutyce mają częstokroć charakter prototypowy, a ponadto na obsłudze aparatury, określaniu warunków bezpieczeństwa pracy itp.

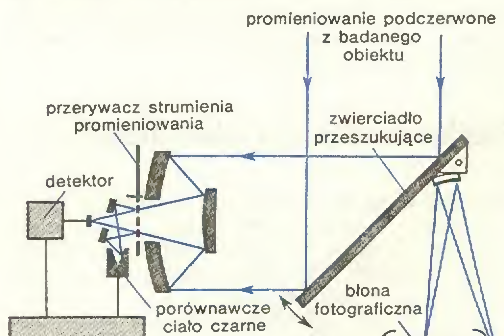
Promieniowanie podczerwone

Źródłem promieniowania podczerwonego są wszystkie ciała, a m.in. człowiek; wykorzystano to w diagnostyce. Zasada działania klasycznych detektorów promieniowania podczerwonego polega na wzroście temperatury odbiornika absorbującego promieniowanie, które nań pada. W nowoczesnych detektorach fotoprzewodzących lub fotoemisyjnych promieniowanie podczerwone przetwarzane jest bezpośrednio na prąd elektryczny.

Do otrzymywania obrazów przedmiotów wysyłających fale podczerwone stosuje się złożone układy optyczne: przetworniki obrazu, które przetwarzają obraz widzialny na ekranie fluorescencyjnym, oraz termografy, stanowiące detektory przeszukujące.

termograf

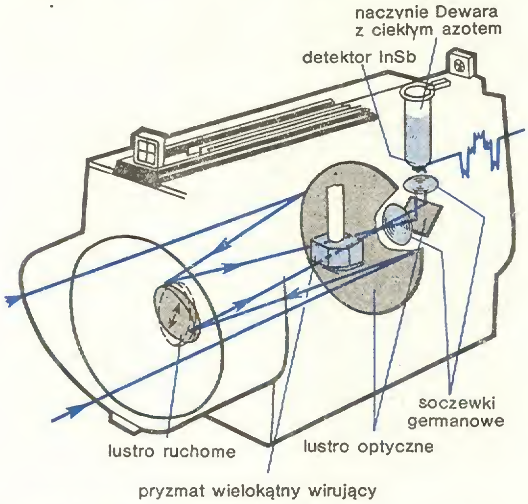
Zasada działania klasycznego termografu podana jest na rys. 18. Promieniowanie podczerwone (cieplne) badanego obiektu, którym jest np. ciało ludzkie,



Rys. 18. Schemat działania termografu

pada na detektor po odbiciu od zwierciadła nachylonego pod kątem 45° do kierunku wiązki promieniowania. Zwierciadło, wprawione w ruch, „przeszukuje" pole widzenia. Sygnały z detektora modulują natężenie światła lampki, która oświetla kliszę fotograficzną. W ten sposób otrzymuje się obraz rozkładu temperatury na obszarze ciała w postaci rozkładu zaczernienia kliszy.

AGA kamera Jednym z najnowszych termografów jest kamera AGA Thermovision (rys. 19). Charakteryzuje ją duża czułość (0,2°C), gdyż jako detektora użyto tu antymonku indu chłodzonego ciekłym azotem; „przeszukiwania" dokonuje się za pomocą drgającego lustra oraz obracającego się kwarcowego pryzmatu; otrzymuje się 16 obrazów termograficznych na sekundę, obserwowanych na ekranie monitora (lampa oscyloskopowa), fotografowanych lub filmowanych.



Rys. 19. Kamera termowizyjna firmy AGA

Za pomocą termowizyjnej kamery można również uzyskać obraz złożony z izoterm, tj. krzywych łączących punkty o tej samej temperaturze; ułatwia to diagnozę kliniczną.

zastosowanie termografii Termograf znajduje szerokie zastosowanie w praktyce medycznej. Interpretacja obrazu otrzymanego w termografie wymaga ustalenia związku między temperaturą normalną, podwyższoną lub obniżoną danej części ciała a stanem klinicznym. Ogólnie biorąc, podwyższenie temperatury łączy się ze zwiększoną przemianą metaboliczną, zwiększonym ukrwieniem, natomiast obniżenie temperatury wiąże się z ogólnym oziębieniem ciała, napięciem psychicznym lub ... paleniem papierosów (il. 186 i 189, tabl. 48).

Termografia jest metodą całkowicie bezpieczną dla pacjentów i choć dostarcza informacji o temperaturze tylko powierzchni ciała, znajduje coraz szersze zastosowanie w medycynie.

Fizyka medyczna nieradiacyjna

Ten dział fizyki, najskromniejszy w niniejszym artykule, dotyczy z jednej strony badań napięć bioelektrycznych i prądów czynnościowych w organizmie żywym, z drugiej zaś — zastosowań bodźców elektrycznych w celach terapeutycznych. Ta ostatnia dziedzina wchodzi jednak tradycyjnie w zakres tzw. fizykoterapii, która choć z nazwy przypomina fizykę medyczną, niewiele ma z nią wspólnego. Fizykoterapia polega bowiem na zastosowaniu w celach leczniczych lub zapobiegawczych naturalnych czynników fizycznych (np. działanie energii mechanicznej lub cieplnej, pewnych cech powietrza i wody) oraz czynników fizycznych sztucznie wytwarzanych przez różnego rodzaju przyrządy elektryczne. Natomiast badanie napięć bioelektrycznych stanowi — również tradycyjnie — domenę tzw. bioelektroniki. Dlatego też ograniczymy się tu do zasygnalizowania niektórych problemów, nie wdając się w szersze rozważania.

fizykoterapia

Z fizycznego punktu widzenia badanie napięć bioelektrycznych w organizmie jest fascynujące zarówno w skali makro (elektrokardiografia, elektroencefalografia), jak i mikro (zjawiska elektryczne w komórce).

napięcia bioelektryczne Zdolność wytwarzania napięć elektrycznych jest cechą wszystkich żywych komórek i tkanek. Związana jest z faktem występowania różnicy stężenia jonów potasu i sodu między wnętrzem komórki a obszarem pozakomórkowym. Ta różnica stężenia jonów prowadzi do utworzenia się potencjału zwanego spoczynkowym. Po przyłożeniu bodźca zewnętrznego potencjał spoczynkowy przechodzi w potencjał czynnościowy, przenoszony wzdłuż włókien nerwowych w myśl zasady „wszystko albo nic". Takim złożonym obrazem potencjałów czynnościowych jest zapis dobrze znanych krzywych, zwanych elektrokardiogramami lub encefalogramami. Krzywe te pozwalają lekarzowi na przeprowadzenie diagnozy chorób serca (EKG) czy anomalnej czynności mózgu (EEG). Ponadto znaczenie kliniczne posiada badanie potencjałów elektrycznych mięśni (elektromiografia) lub siatkówki oka (elektroretinografia).

Ogólnie biorąc, odbierane napięcia bioelektryczne są bardzo małe, często na poziomie szumów aparatury i zakłóceń sieci. Tak np. napięcia występujące przy rejestracji prądów mózgu wahają się od 0,005 do 0,1 mV. Dlatego też wymagania, jakie się stawia bioelektrykom konstruującym wzmacniacze wielokanałowe, są bardzo wysokie.

Obecnie urządzenia elektrokardiograficzne pracują w połączeniu z aparatami do ciągłego lub okresowego rejestrowania danych, szczególnie na tzw. oddziałach intensywnego nadzoru. Z kolei do analizy danych elektroencefalograficznych, których interpretacja jest zazwyczaj trudna, wprowadza się coraz szerzej komputery.

działanie pól elektromagnetycznych Ciekawą dziedziną, w której uczestniczą fizycy medyczni, jest coraz bardziej rozwijająca się dziedzina badań skutków biologicznego oddziaływania pól elektromagnetycznych, ze szczególnym uwzględnieniem mikrofal. Efekt oddziaływania pól elektromagnetycznych ma, jak się wydaje, charakter kumulujący. Szkodliwymi źródłami pól mogą być reklamy świetlne, telewizja itp. Źródeł tych nie można usunąć, ale efekt ich można ograniczyć przez odpowiednie warunki narzucone budownictwu mieszkaniowemu.

Jeżeli chodzi o bardziej wymierne i lepiej zbadane oddziaływanie mikrofal na tkankę, okazuje się, że głównym efektem jest działanie cieplne, choć występują również zjawiska, jak np. aberracje chromozomalne *in vitro* i *in vivo*, które wymagają innego wyjaśnienia. W związku z rozpowszechnianiem się stosowania mikrofal sprawy te nabierają coraz większego znaczenia.

Z przedstawionego wyżej krótkiego przeglądu fizyki medycznej widzimy, że ta nowa dziedzina zastosowań fizyki wyodrębnia się w pewnych działach w samodzielną specjalność (np. promieniowanie jonizujące), w innych zaś współpracuje z bioelektroniką, bioinżynierią czy biofizyką. Trudno jest przeprowadzić jakiś zdecydowany podział, a może nawet nie trzeba, gdyż nauka współczesna przestaje być zbiorem nie powiązanych ze sobą dziedzin, zaczyna coraz bardziej przypominać jedną wspólną budowlę opartą na tych samych fundamentach.

O. A. Chomicki, T. Górowski, W. Jasiński *Scyntygrafia kliniczna*, Warszawa 1971; R. Miller, R. Richwien *Podstawy elektroniki medycznej*, Warszawa 1973; A. Piątkowski, W. Scharf *Aparatura radiomedyczna w medycynie i biologii*, Warszawa 1972; A. Piątkowski, W. Scharf *Mierniki promieniowania jonizującego* 1980; A. Pilawski, *Podstawy biofizyki*, Warszawa 1977; *Radiologia*, red S. Z. Zgliczyński, Warszawa 1967.

Modelowanie matematyczne procesów biologicznych

Dymitr Czernawski i Ewa Skrzypczak

Modelowanie procesów biologicznych należy do ważnych i szeroko stosowanych metod w naukach biologicznych. Modelowanie matematyczne odgrywa tu szczególną rolę, zwłaszcza w stosunkowo młodych dziedzinach: biofizyce i biologii teoretycznej. Jego znaczenie oraz związane z nim problemy opiszemy powołując się na przykład starszej dyscypliny naukowej, a mianowicie fizyki. Zadaniem fizyki teoretycznej jest formułowanie postulatów i praw niesprzecznych ze znanymi faktami eksperymentalnymi oraz wyciąganie na ich podstawie wniosków, które powinny być konfrontowane z wynikami doświadczeń nie znanymi dotychczas bądź nie należącymi do zespołu faktów leżących u podstaw teorii. Sformułowanie teorii, a zwłaszcza jej skuteczne wykorzystanie dla otrzymania wniosków, wymaga często wprowadzenia upraszczających założeń, idealizujących rzeczywistość fizyczną. Takie założenia stanowią punkt wyjścia dla konstrukcji modeli w fizyce. Typowym, prostym tego przykładem jest w fizyce zagadnienie ruchu wahadła matematycznego, które stanowi model rzeczywistego fizycznego wahadła. W modelu pomija się siłę tarcia i oporu ośrodka, rozciągliwość nici, rozmiary i kształt ciężarka, zastępowanego w modelu przez tzw. punkt materialny. Uproszczenia wprowadzone w założeniach modelu ograniczają oczywiście wnioski do określonego zespołu procesów dotyczących badanego obiektu czy zjawiska.

model: wahadło matematyczne

Mimo oczywistej niedoskonałości modelu w porównaniu ze ścisłą, kompletną teorią jego skonstruowanie i analiza wzbogacają z reguły wiedzę o pewnym zespole zjawisk i na ogół poprzedzają następny etap, którym jest sformułowanie teorii fizycznej.

W odróżnieniu od fizyki biologię uważano do niedawna za dyscypline eksperymentalną, jeżeli nie opisową. Biofizyka natomiast bywa często potocznie uznawana za tę dziedzinę biologii, w której podstawową rolę odgrywa zastosowanie aparatury i eksperymentalnych metod fizyki w badaniu obiektów biologicznych. W ostatnich latach zrodziły się nowe dyscypliny: biologia teoretyczna i biofizyka teoretyczna. Podstawową funkcję w tych dziedzinach pełni modelowanie matematyczne procesów zachodzących w żywych obiektach lub w ich zespołach. Podobnie jak w fizyce, podejście modelowe w biologii wymaga starannego sformułowania założeń, ograniczenia rozważań do określonego zespołu faktów i zjawisk najbardziej istotnych dla badanego procesu, wprowadzenia uproszczeń. Następnym krokiem jest analiza modelu i wyciągnięcie wniosków, ułatwiających zrozumienie mechanizmu procesu, oraz opis, choćby w ogólnym zarysie, zachowania się badanego układu i przebiegu zjawiska w czasie.

biologia teoretyczna i biofizyka teoretyczna

Jednym z podstawowych założeń w modelach i teoriach fizycznych jest izolacja rozważanego układu od wpływu czynników zewnętrznych. Mówimy wówczas o układzie zamkniętym i odosobnionym. W modelowym podejściu do badania układów biologicznych istotną trudność stanowi fakt, że dopóki rozważane obiekty są żywe, dopóty nie mogą stanowić układu izolowanego i zamkniętego.

Odrębnym zagadnieniem utrudniającym konstruowanie i analizę modeli biologicznych jest ogromna liczba i różnorodność procesów zachodzących w obiektach biologicznych. Wybór procesów czy zjawisk najbardziej istotnych stanowi podstawowy krok przy konstrukcji modelu. Jest to trudne i odpowiedzialne zadanie, wymagające znajomości ogólnych charakterystyk rozważanych zjawisk oraz wykorzystania odpowiedniego aparatu i metod matematycznych.

Wspomniane trudności sprawiają, że młode dyscypliny — biologia i biofizyka teoretyczna — rozwijają się wolniej niż fizyka teoretyczna, a osiągnięcia ich nie mogą na ogół być teoriami o takim stopniu doskonałości i precyzji, jakie charakteryzują teorie fizyczne. Istotny rozwój biologii i biofizyki teoretycznej w ostatnich latach wiąże się w znacznym stopniu z rozwojem eksperymentalnej biologii i biofizyki oraz z zainteresowaniem tymi dziedzinami matematyków i fizyków teoretyków, którzy wnoszą w tę dziedzinę znajomość aparatu matematycznego oraz ogólnego procesu przechodzenia od hipotez do modeli i od modeli do teorii.

Niniejszy artykuł przedstawia niektóre aspekty współczesnej biologii i biofizyki teoretycznej związane z modelowaniem matematycznym procesów biologicznych jako szczególną metodą modelowania.

Klasyfikacja modeli

Jak już wspomniano wyżej, modelowe podejście do opisu i analizy jakiegoś procesu lub zespołu procesów wymaga znajomości pewnych podstawowych faktów, wynikających z obserwacji i doświadczeń, oraz przyjęcia pewnych założeń, które stanowią z reguły jakąś idealizację, a zarazem uproszczenie problemu. Następnym krokiem jest skonstruowanie modelu w postaci rzeczywistego układu fizycznego, poddającego się bezpośredniej lub pośredniej obserwacji, bądź w postaci układu równań matematycznych. Analiza działania układu fizycznego lub właściwości układu równań prowadzi następnie do wniosków.

konstruowanie modelu

Modele stosowane obecnie w biologii i biofizyce możemy podzielić (pamiętając jednak, że żadna klasyfikacja nie jest ścisła i doskonała) na modele fizyczne (w szczególności analogowe), matematyczne (dynamiczne) i statystyczne.

Modele fizyczne

Modele fizyczne są to układy fizyczne zawierające elementy mechaniczne, hydrauliczne, elektryczne lub elektroniczne, akustyczne czy optyczne, umożliwiające imitowanie określonych procesów biologicznych przez procesy fizyczne, które jest łatwiej obserwować i analizować. Pozwala to na wyciąganie wniosków o właściwościach procesów i ich przebiegu.

Ilustrację modelu fizycznego mogą stanowić modele, w których przepływ cieczy w odpowiednio dobranym układzie naczyń i rurek imituje obieg krwi, powietrza czy innych substancji w żywym organizmie lub w jego części. Wprowadzenie do układu dodatkowych elementów, takich jak błony półprzewodzące, wymiana, dopływ, czy odpływ płynu z układu itp., umożliwia uwzględnienie szeregu szczegółów właściwych modelowanemu zjawisku.

model Lilliego

Innym przykładem modelu fizycznego jest klasyczny model Lilliego, imitujący rozchodzenie się impulsu wzdłuż włókna nerwowego (aksonu). Podstawowym elementem modelu jest żelazny drut umieszczony w roztworze kwasu azotowego. W tych warunkach powierzchnia drutu pokrywa się szybko warstwą tlenku żelaza. Do układu doprowadzamy z zewnętrznego źródła impuls elektryczny. Wzdłuż drutu zaczyna się rozchodzić sygnał elektryczny, któremu towarzyszy chwilowe pozbawienie kolejnych odcinków

drutu warstwy tlenku żelaza. Na podstawie obserwacji obszaru znikania tej warstwy i pomiarów przeprowadzonych przy pomocy odpowiednich układów elektrycznych można ocenić prędkość rozchodzenia się sygnałów wzdłuż drutu oraz kształt impulsu. Stwierdzono, że sygnał taki powstaje i rozchodzi się, jeżeli wielkość pierwotna impulsu przekracza pewną wartość progową, oraz że kształt powstałego sygnału nie zależy od wielkości pierwotnego impulsu. Model ten umożliwia jakościową imitację rzeczywistego przebiegu zjawiska rozchodzenia się sygnałów wzdłuż włókien nerwowych.

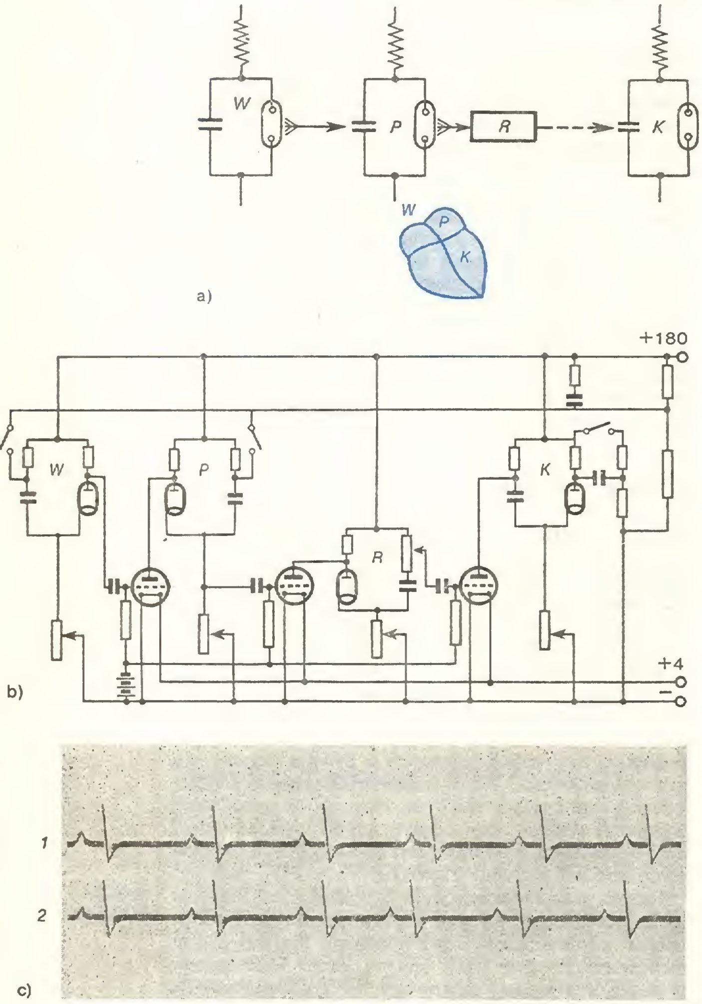
modele analogowe

Często do badania procesów biologicznych, a zwłaszcza ich dynamiki, stosuje się modele analogowe. Zasadą konstruowania modelu analogowego jest zastąpienie układu rzeczywistego przez układ innego rodzaju, opisywany takimi samymi zależnościami matematycznymi jak układ modelowany i mający parametry wzajemnie proporcjonalne — niezależnie od ich fizycznej natury. Modelami analogowymi mogą być układy elektroniczne, hydrauliczne.

model pracy serca

Dla stosowania modeli tego rodzaju wystarcza niekiedy tylko jakościowa znajomość analizowanego procesu; ścisłe formułowanie związków matematycznych nie jest wówczas niezbędne.

Przykładem takiego modelu analogowego jest klasyczny już model pracy serca zaproponowany przez Van der Pola (rys. 1). Składał się on z lampki neonowej, zapewniającej oscylacyjny przebieg zjawiska, oraz kondensatora i oporu elektrycznego, imitującego działanie węzła zatokowego. Prosty ten model pozwalał imitować procesy związane z biciem serca zarówno w przypadkach fizjologicznych, jak i patologicznych (arytmia) — w zależności od warunków pracy (w szczególności oporu elektrycznego w układzie analogowym). Opisanemu tu modelowi, opracowanemu ponad 40 lat temu, poświęcamy więcej miejsca, aby podkreślić, że bardzo proste założenia matematyczne oraz dość prymitywna aparatura mogą niekiedy odtwarzać złożone procesy biologiczne.



Rys. 1. Model pracy serca van der Pola: a) analogie między podukładami aparatury i modelowanego serca (*W* węzeł zatokowy, *P* przedsionek, *K* komora serca, *R* układ opóźniający), b) kompletny układ elektryczny, c) sztuczny elektrokardiogram otrzymany za pomocą układu elektrycznego dla przypadku fizjologicznego (*1*) i patologicznego (*2*)

maszyny analogowe

Odpowiednio skonstruowane maszyny analogowe umożliwiają bezpośrednią obserwację (na ekranie oscyloskopu) przebiegów czasowych interesujących nas zmiennych oraz zależności tych przebiegów od parametrów układu i od założeń stanowiących punkt wyjścia dla modelowania badanego procesu. Otwiera to szerokie możliwości zarówno w zakresie sprawdzania przyjętych założeń początkowych, jak i szczegółowego opisu przebiegu badanych zjawisk, a nawet umożliwia przewidywanie faktów trudnych lub wręcz niedostępnych bezpośredniej obserwacji w realnych procesach biologicznych.

Modele matematyczne

Modelem matematycznym badanego procesu bądź zespołu obserwacji dotyczących przejawów jakiegoś zjawiska nazywamy układ związków matematycznych, wiążących ze sobą te zmienne (wielkości mierzone czy — ogólniej — stanowiące przedmiot zainteresowania), wartości parametrów oraz czas — jeżeli rozważany model dotyczy dynamiki procesu, tzn. przebiegu analizowanych procesów w czasie. Zmiennymi, występującymi w modelu, mogą być np. stężenia substancji biorących udział w reakcjach biochemicznych, liczebności populacji, wielkości charakteryzujące badane obiekty, takie jak ich rozmiary, prędkości itp.

modele statyczne

Modele statyczne mają najczęściej postać jednego (lub więcej) związku algebraicznego. Kształt matematyczny takiego związku może wynikać z przesłanek *a priori*, wynikających z ogólnej znajomości badanych zjawisk, bądź z danych doświadczalnych, do których dopasowuje się postać związków matematycznych. Jeżeli — dla ilustracji — ograniczymy się do najprostszego przypadku jednego związku między dwiema zmiennymi, przy czym przypuszczamy, że związek ten jest liniowy, to modelem matematycznym procesu nazwiemy równanie o postaci $y = ax+b$, gdzie x i y są zmiennymi, zaś a i b — parametrami modelu. Zadaniem, jakie sobie wówczas stawiamy, jest: a) znalezienie odpowiedzi na pytanie, czy dane doświadczalne (zbiór par wartości x i y wyznaczonych ze znaną dokładnością) spełniają powyższy związek liniowy, oraz b) wyznaczenie wartości parametrów a i b i ich błędów. Rozwinięte są metody statystyczne umożliwiające wykonanie tych zadań oraz określenie tzw. poziomu ufności uzyskanych wyników. Jeżeli zgodność modelu z danymi doświadczalnymi okaże się niezadowalająca, należy testowany model (liniowe równanie w naszym przykładzie) odrzucić lub zmodyfikować. Taka modyfikacja polega zazwyczaj na wzbogaceniu postaci modelu o człony zawierające wyższe potęgi x lub y, bądź o człon zawierający inną, pierwotnie pominiętą, zmienną (z). Zastosowanie odpowiednich metod statystycznych pozwala następnie na znalezienie najlepiej pasującego do danego zespołu parametrów (a, b, ...) testowanej postaci modelu. Interpretacja tak otrzymanego i akceptowanego modelu prowadzi do lepszego opisu i zrozumienia badanych zjawisk. Ponadto znajomość związku (czy związków) między zmiennymi występującymi w modelu umożliwia następnie wykorzystywanie modelu do znajdowania wartości jednej zmiennej, gdy znamy pozostałe — bez przeprowadzania za każdym razem badań eksperymentalnych.

Modele dynamiczne, których przedmiotem jest analiza przebiegu badanego procesu w czasie, mają zazwyczaj postać układu złożonego z jednego lub więcej równań różniczkowych zawierających pochodne zmiennych względem czasu.

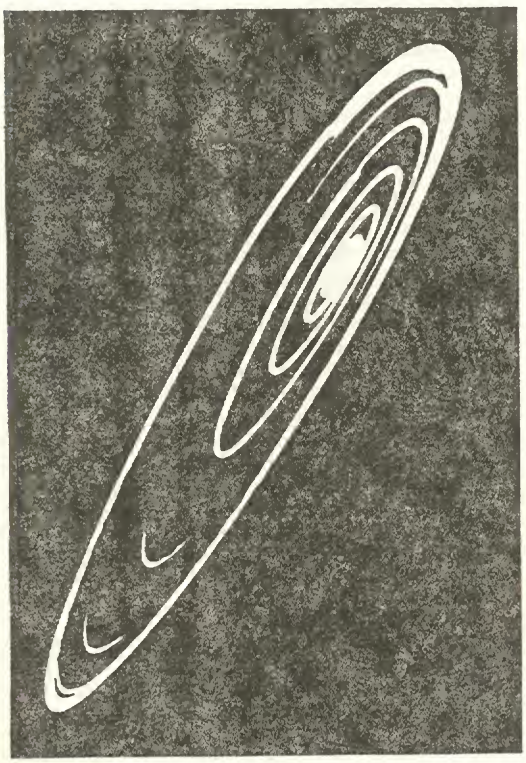
Analizy takiego modelu matematycznego można dokonać bądź stosując teorię równań różniczkowych (opis tego podejścia i przykłady jego zastosowania przedstawimy w rozdziałach: „Modele matematyczne" i „Wybrane przykłady modeli matematycznych w zagadnieniach biologicznych"), bądź za pomocą analogowych maszyn matematycznych. Rolą maszyny analogowej jest rozwiązanie układu równań stanowiących model i znalezienie związków między wielkościami dynamicznymi. Szczególnie ważna jest w tym wypadku możliwość badania reakcji modelowanego układu na zmiany parametrów układu, charakteryzujących warunki, w jakich działa rzeczywisty układ.

rola maszyny analogowej

Na rys. 2 widać zdjęcie ekranu oscyloskopowego maszyny analogowej, przy użyciu której analizowano model złożonego zjawiska — procesu fotosyntezy u roślin. Obraz otrzymany na ekranie przedstawia związek między zmiennymi dynamicznymi — w tym wypadku są to stężenia tzw. lekkich i ciężkich cukrów biorących udział w reakcjach chemicznych, które leżą u podstaw zjawiska fotosyntezy. Na osiach poziomej i pionowej odłożone są stężenia odpowiednich cukrów. Nawet pobieżne spojrzenie na zdjęcie prowadzi do ważnego wniosku o oscylacyjnym charakterze zmian w czasie stężeń lekkich i ciężkich cukrów, przy czym oscylacje stężeń tych cukrów są względem siebie przesunięte w fazie. Zdjęcie to przedstawia tzw. obraz fazowy badanego układu. Zastosowanie maszyny analogowej zastąpiło niełatwą w tym wypadku bezpośrednią analizę układu równań stanowiących model matematyczny zjawiska.

Ostatnio w biologii teoretycznej badane są modele matematyczne nie wymagające nawet formułowania konkretnych równań matematycznych; modele tego rodzaju prowadzą z reguły do wniosków jakościowych. Punktem ich wyjścia jest teoria katastrof R. Thoma. Przy takim podejściu do zagadnienia ustala się zespół zmiennych i parametrów istotnych dla badanego zjawiska. Zarówno zmienne, jak i parametry traktowane są na równych prawach — z tym, że parametrom przypisuje się powolniejszą zmienność w czasie niż zmiennym. Zbiór tych wszystkich wielkości, charakteryzujący chwilowy stan układu, stanowi zbiór współrzędnych punktu w przestrzeni wielowymiarowej. Zbiór stanów rozważanego układu przedstawiony jest przez pewną hiperpowierzchnię (wielowymiarową) w tej przestrzeni. Rozważania oparte na znajomości topologii pozwalają na wprowadzenie pewnej klasyfikacji takich powierzchni. Jakościowe przypisanie badanemu zjawisku określonego charakteru takiej powierzchni i jej osobliwości geometrycznych umożliwia dokonanie jakościowego opisu procesów odbywających się w układach biologicznych i wyciągnięcie pewnych wniosków o właściwościach tych procesów, jak np. gwałtowne skokowe zmiany jednych zmiennych przy jednoczesnej gładkiej zmienności innych. Tego rodzaju modele teoretyczne bywają stosowane np. do analizy takich zagadnień, jak praca serca, działanie neuronu, różnicowanie tkanek, oraz pewnych problemów z dziedziny embriologii i ekologii.

teoria katastrof Thoma



Rys. 2.

Bardziej szczegółowa dyskusja teorii katastrof w zastosowaniu do modelowania procesów biologicznych wykracza poza ramy niniejszego artykułu.

Modele statystyczne

W modelach matematycznych zmienne oraz parametry badanego układu są pewnymi wielkościami mającymi charakter wielkości średnich (np. stężenie substancji, liczebność komórek czy populacji, prędkość reakcji). Operowanie nimi jest uzasadnione wówczas, gdy przypadkowe fluktuacje w wartościach tych wielkości są małe w porównaniu z samymi wielkościami. Zdarza się jednak, że informacje dotyczące średnich wielkości charakteryzujących badane obiekty i procesy są niedostatecznie ściśle, a przypadkowe fluktuacje są porównywalne z samymi wielkościami. Odpowiada to sytuacjom, gdy liczba interesujących nas obiektów jest niewielka i pojęcie stężenia traci swój ścisły sens. Tak bywa w niektórych zagadnieniach ekologicznych, kiedy np. liczba osobników na danym terytorium wyraża się w dziesiątkach, a nawet w jednostkach, lub przy rozważaniu reakcji biochemicznych, gdy liczba cząstek biorących udział w reakcji jest w obrębie jednej komórki niewielka.

wartości średnie

W tego rodzaju zagadnieniach stosuje się modele statystyczne, w których rolę zmiennych odgrywają prawdopodobieństwa analizowanych zdarzeń. Związki między zmiennymi zapisuje się w postaci równań różniczkowych lub równań algebraicznych, albo też podaje się algorytmy, umożliwiające przeprowadzenie rachunków modelowych na komputerach metodą Monte Carlo. Modele wykorzystujące metodę Monte Carlo stosowane są np. przy badaniu i analizie zagadnienia rozwoju i podziału komórek wówczas, gdy procesy te w znacznym stopniu zależą od przypadkowych czynników otoczenia, w wyniku czego okres podziału i czas trwania kolejnych faz w cyklu komórkowym podlegają znacznym fluktuacjom.

modele wykorzystujące metodę Monte Carlo

Modele statystyczne, w których się stosuje równania różniczkowe, są w istocie — mimo różnic w określeniu zmiennych charakteryzujących badany układ — podobne do modeli dynamicznych, którymi się zajmiemy bardziej szczegółowo w następnym rozdziale. Analiza modelu statystycznego umożliwia nie tylko znalezienie związków między zmiennymi dynamicznymi modelowanego układu, lecz także ocenę roli fluktuacji i ich wpływu na przebieg zjawiska. Szczególne znaczenie takiej analizy przejawia się np. przy badaniu układów przełączeniowych (tryggerowych), w których niewielka przypadkowa fluktuacja wartości zmiennych czy parametrów może decydować o dalszym przebiegu dynamiki modelowanego procesu (dokładniej w rozdziale Modele przełączania).

Modele matematyczne (dynamiczne)

Jak już wspomniano wyżej, modelem matematycznym nazywamy układ związków matematycznych między zmiennymi dynamicznymi, współrzędnymi przestrzennymi, czasem i parametrami układu.

zmienne dynamiczne

Zmienne dynamiczne, ich liczba i charakter zależą od rodzaju analizowanego układu. W zagadnieniach biochemicznych są to najczęściej stężenia substancji wchodzących w reakcje, w zagadnieniach ewolucyjnych i ekologicznych — liczebności badanych obiektów lub populacji. Zależność zmiennych (x) od czasu (t) wyrażana jest w modelu przez pochodne tych zmiennych względem czasu (dx/dt). Na ogół zmienne dynamiczne mogą zależeć nie tylko od czasu, lecz i od miejsca w przestrzeni.

parametry układu

Parametrami układu mogą być np. prędkości reakcji chemicznych, śmiertelność lub przyrost naturalny badanej populacji, wielkości charakteryzujące oddziaływanie między komórkami, cząsteczkami chemicznymi, członkami badanych populacji itp. Konkretne wartości parametrów często nie są znane lub też zmieniają się w szerokim zakresie — w zależności od warunków, w jakich się znajduje badany układ.

Nie wchodząc na razie w szczegóły procedur stosowanych przy analizie modelu (poniżej), wymienimy tylko kolejne etapy modelowania matematycznego.

1. Zebranie dostępnych (zaczerpniętych z badań eksperymentalnych oraz ze znanych ogólnych prawidłowości) informacji o badanym procesie.

2. Podjęcie decyzji, jakie zmienne dynamiczne oraz jakie parametry będą rozważane w modelu.

3. Sformułowanie modelu, tj. związków między zmiennymi dynamicznymi, parametrami i czasem oraz położeniem w przestrzeni. Zależności od czasu i miejsca wyraża się przez pochodne względem czasu i współrzędnych geometrycznych.

4. Analiza modelu matematycznego, polegająca bądź na

a) rozwiązaniu układu równań różniczkowych i otrzymaniu jawnej postaci zależności każdej ze zmiennych od pozostałych zmiennych, czasu i miejsca w przestrzeni, bądź na

b) jakościowej analizie właściwości układu dynamicznego, prowadzącej do wniosków o ogólnym charakterze wzajemnych związków między zmiennymi i o ich zależności od parametrów, czasu i miejsca.

5. Sformułowanie wniosków dotyczących dynamicznych charakterystyk analizowanego zjawiska czy zespołu zjawisk. Wnioski takie mogą np. zawierać stwierdzenie charakteru zmienności w czasie poszczególnych zmiennych (np. występowanie oscylacji) lub stwierdzenie asymptotycznego (po długim czasie) zachowania się zmiennych (np. dążenie do wartości stacjonarnych odpowiadających trwałym, stabilnym stanom układu albo oddalenie się od wartości stacjonarnych odpowiadających stanom nietrwałym.

Ważne jest, że badając model, można przeanalizować wpływ poszczególnych parametrów na przebieg badanego zjawiska, a w szczególności — przewidzieć możliwość występowania nagłych zmian jakościowych przebiegu procesu (przełączanie).

6. Porównanie wniosków wynikających z analizy modelu ze znanymi faktami doświadczalnymi dotyczącymi badanego procesu lub wskazanie możliwych obserwacji i eksperymentów umożliwiających porównanie rzeczywistych charakterystyk procesu z przewidywaniem modelu. Po dokonaniu takiej konfrontacji model i założenia stanowiące jego podstawę mogą być uznane za prawidłowe, oczywiście w granicach przewidzianych przy wprowadzeniu uproszczeń i przy idealizacji rzeczywistej sytuacji.

W następnych paragrafach tego rozdziału omówimy metody modelowania matematycznego układów dynamicznych i wprowadzimy podział modeli na modele punktowe, modele uwzględniające wiek obiektów oraz modele uwzględniające układy przestrzennie niejednorodne.

Modele punktowe

najprostsze modele dynamiczne

Do najprostszych modeli matematycznych należą takie, w których zmienne dynamiczne nie zależą od współrzędnych przestrzennych. Takie modele matematyczne nazywamy punktowymi. Przed sformułowaniem układu równań stanowiących model matematyczny układu niezbędne jest podjęcie decyzji o wyborze i liczbie zmiennych dynamicznych. Jeżeli liczba zmiennych (x_i) wynosi n, wówczas model matematyczny badanego zjawiska można zapisać w postaci układu równań różniczkowych:

$$\frac{dx_i}{dt} = P_i(x_1, x_2, \ldots, x_n), \qquad (1)$$

gdzie w funkcjach P_i zawarte są parametry charakteryzujące warunki, w jakich zachodzą badane procesy. Jeżeli pochodne zmiennych dynamicznych względem czasu (dx_i/dt) są miarą tempa zmian wielkości x_i w czasie, to charakteryzują one dynamikę zachowania się tych wielkości.

Rozwiązanie tego układu równań w postaci jawnych zależności każdej ze zmiennych x_i od pozostałych zmiennych i czasu daje obraz zachowania się modelowanego układu w czasie. Wykażemy niżej, że nie zawsze taka postać rozwiązania jest niezbędna. Często wystarczające dla opisu modelowanego procesu jest podanie i analiza jego tzw. obrazu fazowego (o czym niżej).

układ autonomiczny

Jeżeli funkcje P_i nie zależą w sposób jawny od czasu, to układ taki nazywamy autonomicznym; w fizyce jest to równoważne badaniu układu odosobnionego. Jeżeli mamy do czynienia z układem znajdującym się w zmiennych warunkach zewnętrznych, wówczas mówimy o układzie nieautonomicznym; w układzie równań odpowiada to wprowadzeniu jawnej zależności od czasu. Dążenie do uwzględnienia w modelu wszystkich procesów zachodzących w badanym układzie biologicznym wiąże się nieuchronnie z wprowadzeniem dużej liczby zmiennych i — odpowiednio — dużej liczby równań stanowiących układ. Analiza takiego układu i efektywne wyprowadzenie wniosków, choćby tylko jakościowych, stanowi często zadanie trudne, niekiedy wręcz niewykonalne. Wynika to z trudności matematycznych, a przede wszystkim — z niedoskonałej znajomości charakterystyk poszczególnych procesów. Zdarza się co prawda, że nie znając szczegółowo jakiegoś procesu, wiemy jednak że jest to proces bardzo powolny w porównaniu z innymi, opisywanymi przez pozostałe równania układu, albo odwrotnie: wiemy, że jedna ze zmiennych dynamicznych zmienia się bardzo szybko, np. oscylując z okresem małym w porównaniu z interwałem czasowym charakterystycznym dla innych zmiennych. W takich wypadkach wolnozmienne wielkości można uznać za stałe i traktować jako parametry układu dynamicznego. Z drugiej strony — można przyjąć, że szybko zmieniające się wielkości osiągnęły już swoje wartości stacjonarne, odpowiadające chwilowym wartościom innych zmiennych. Umożliwia to związanie tych szybkozmiennych wielkości z pozostałymi za pomocą równań algebraicznych (przez przedstawienie tych zmiennych jako funkcji pozostałych zmiennych).

W wyniku takich rozważań i zastosowania odpowiednich procedur matematycznych otrzymujemy ostateczny układ dynamiczny, zawierający tylko te zmienne, których prędkości zmian w czasie są w przybliżeniu tego samego rzędu. Oznacza to, że jeżeli analizujemy zjawisko o pewnych określonych charakterystykach czasowych, nieistotne są procesy przebiegające znacznie wolniej lub znacznie szybciej.

Model matematyczny składa się zatem w rezultacie najczęściej z niewielu równań różniczkowych, często tylko z dwóch lub trzech.

najprostszy układ dynamiczny

Zajmiemy się najprostszym rodzajem układu dynamicznego, złożonym tylko z dwóch równań różniczkowych, które zapiszemy w postaci:

$$dx/dt = P(x, y), \quad dy/dt = Q(x, y). \tag{2}$$

Funkcje $P(x, y)$ i $Q(x, y)$ stanowią więc miarę tempa zmian zmiennych dynamicznych x i y. Odpowiednio równania te dzielimy stronami przez siebie, eliminujemy czas z układu równań (2) i otrzymujemy:

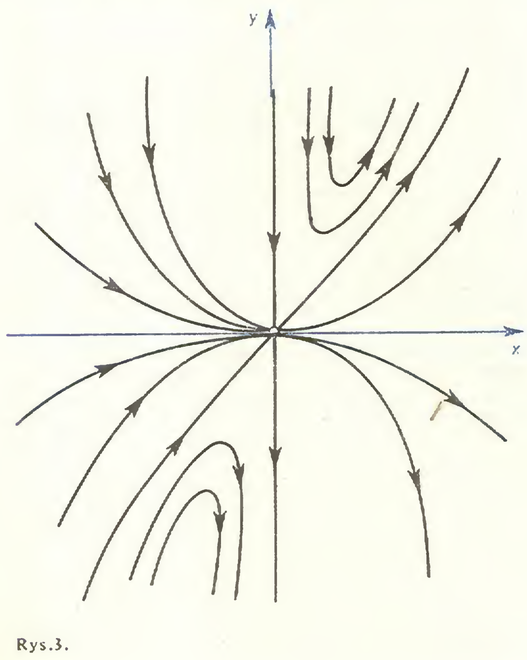
$$dy/dx = Q(x, y)/P(x, y) \tag{3}$$

pod warunkiem, że procedura ta nie dotyczy punktów, w których funkcja $P(x, y)$ osiąga wartość zero. Otrzymane w ten sposób równanie różniczkowe, wiążące zmienne dynamiczne x i y, można niekiedy rozwiązać analitycznie, tj. można znaleźć jawną postać związku $y = f(x)$, który spełnia równanie (3). Elementarna znajomość rachunku różniczkowego i całkowego prowadzi do wniosku, że rozwiązanie zapisane w postaci $y = f(x)$ nie jest jedyne; ogólne rozwiązanie równania (3) ma postać:

$$y = f(x, C), \tag{4}$$

gdzie C jest tzw. stałą całkowania. Wartość stałej całkowania C zależy od wartości początkowych, tj. wartości zmiennych x i y w określonej chwili, którą możemy nazwać chwilą początkową, $t = 0$. W ten sposób rozwiązaniem równania (3) jest cała rodzina związków między zmiennymi x i y; poszczególne funkcje, wchodzące w skład tej rodziny różnią się między sobą wartościami stałej C. Związek między zmiennymi x i y można przedstawić jako wykresy w układzie współrzędnych x i y. Rysunek 3 przedstawia przykład takiej rodziny krzywych, przy czym każdej krzywej odpowiada inna określona wartość stałej C. Taką płaszczyznę (przestrzeń trój- lub więcej wymiarową, jeżeli mamy do czynienia z więcej niż dwiema zmiennymi) nazywamy płaszczyzną fazową, a wykresy funkcji (4) nazywamy trajektoriami układu w przestrzeni fazowej. Trajektorie są krzywymi zorientowanymi, tzn. zaopatrzonymi w strzałki określające zwrot ruchu po trajektorii. Przechodzenie od punktu do punktu wzdłuż określonej trajektorii od-

płaszczyzna fazowa



Rys. 3.

powiada zmianom stanu modelowanego układu w czasie, jeżeli układ wystartował w chwili początkowej ze stanu, który odpowiada punktowi znajdującemu się na tej trajektorii, a warunki (parametry układu bądź wpływy czynników zewnętrznych) nie uległy w tym czasie zmianie. Zbiór trajektorii nosi nazwę obrazu fazowego modelu (2). Każdemu punktowi przestrzeni fazowej odpowiada określona para wartości zmiennych dynamicznych x i y, a zatem — określony stan badanego układu. Należy przy tym pamiętać, że współrzędne y i x nie mają nic wspólnego ze współrzędnymi przestrzennymi. Można wykazać, że przez każdy punkt płaszczyzny fazowej, z wyjątkiem punktów, w których $P(x, y) = 0$ i $Q(x, y) = 0$ (takie punkty nazywamy punktami osobliwymi) przechodzi tylko jedna trajektoria. Znalezienie punktów osobliwych i zbadanie ich charakteru (o czym niżej) stanowi istotny element badania układu dynamicznego.

obraz fazowy modelu

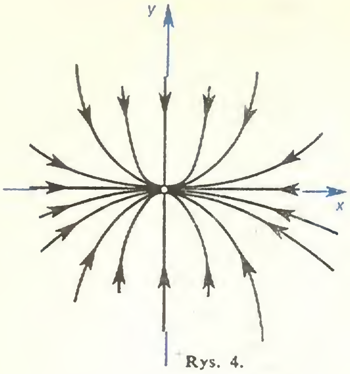
Jeżeli analityczne rozwiązanie równania (3) wiąże się ze znacznymi trudnościami matematycznymi, to zawsze jest możliwe skonstruowanie obrazu fazowego badanego układu przy pomocy rachunków numerycznych. Wykorzystuje się wówczas fakt, że wartość $dy/dx = Q(x, y)/P(x, y)$ odpowiadająca każdej parze wartości x i y daje kierunek stycznej do trajektorii przechodzącej przez punkt (x, y). Jednak nawet bez szczegółowych rachunków numerycznych można wykorzystać znajomość funkcji $P(x, y)$ i $Q(x, y)$ dla znalezienia ogólnych właściwości obrazu fazowego badanego układu. Równanie $P(x, y) = 0$ określa pewną krzywą zwaną izokliną stycznych pionowych; krzywa ta jest miejscem geometrycznym wszystkich punktów, w których styczne do trajektorii układu dynamicznego mają kierunek pionowy. Analogicznie równanie $Q(x, y) = 0$ określa izoklinę stycznych poziomych. Krzywe te nazywamy głównymi izoklinami obrazu fazowego badanego układu dynamicznego (zob. np. rys. 14). Punkt przecięcia głównych izoklin nazywamy punktem osobliwym (stacjonarnym). W stanie odpowiadającym temu punktowi pochodne dx/dt i dy/dt równają się zeru, a zatem zmienne x i y są stałe, nie ulegają zmianom. Stan taki można więc nazwać stanem równowagi, przy czym — podobnie jak w zagadnieniach statyki w mechanice — stan równowagi może być trwały (stabilny) lub nietrwały (niestabilny). Układ wyprowadzony z tego stanu (np. przez działanie czynników zewnętrznych) lub znajdujący się w chwili początkowej w pobliżu tego stanu dąży odpowiednio po odpowiedniej trajektorii do stanu równowagi trwałej lub oddala się coraz bardziej od stanu równowagi nietrwałej. Analiza charakteru punktów osobliwych na obrazie fazowym modelowanego układu stanowi zatem ważną informację o dynamice układu. Nie wchodząc w szczegóły matematycznych rozważań prowadzących do otrzymania obrazu fazowego, podamy poniżej kilka przykładów różniących się między sobą charakterem punktów osobliwych:

izokliny

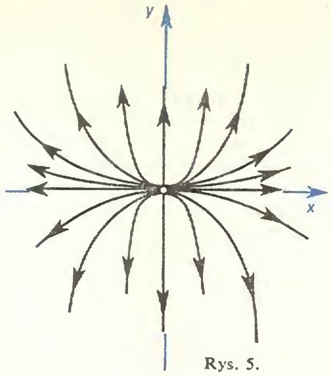
przykłady obrazów fazowych

a) Rysunek 4 przedstawia przykład obrazu fazowego, którego punkt osobliwy jest węzłem trwałym (na rysunku naszkicowano kilka spośród nieskończenie wielu trajektorii fazowych); wszystkie trajektorie zbiegają się w miarę upływu czasu do węzła trwałego, niezależnie od stanu początkowego układu. Stanowi to ważną informację w wypadku modelowania konkretnego układu. Stan odpowiadający węzłowi stabilnemu jest stanem równowagi trwałej odpowiadającym określonym wartościom zmiennych dynamicznych x i y.

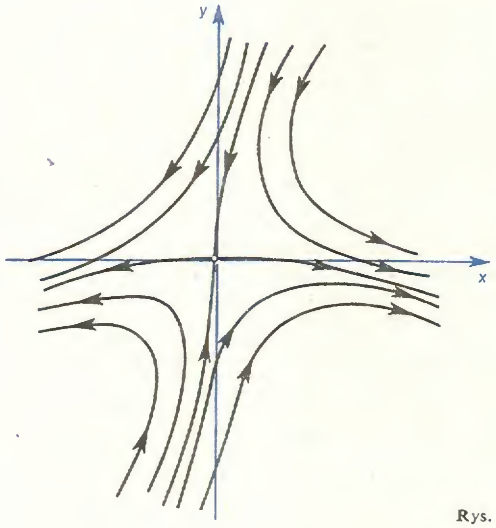
b) Rysunek 5 przedstawia przykład obrazu fazowego, którego punktem osobliwym jest węzeł nietrwały. Z biegiem czasu obydwie zmienne dynamiczne oddalają się od tego punktu, odpowiadającego stanowi równowagi nietrwałej. Stan układu odpowiadający punktowi osobliwemu jest nietrwały, co oznacza, że dowolnie mała, przypadkowa fluktuacja wyprowadzająca układ z tego stanu prowadzi do zmian nieod-



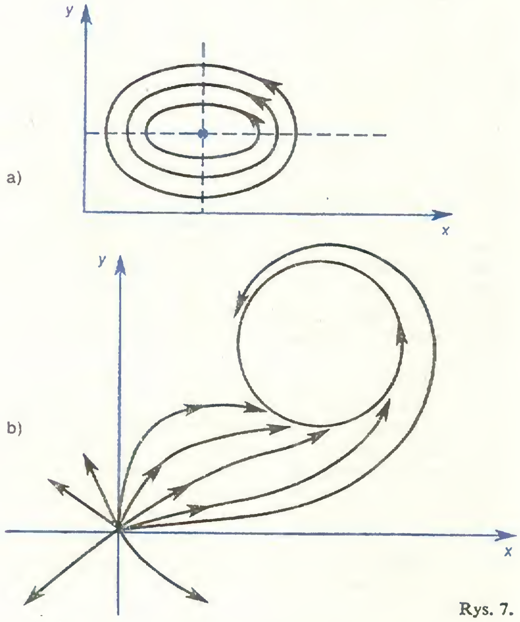
Rys. 4.



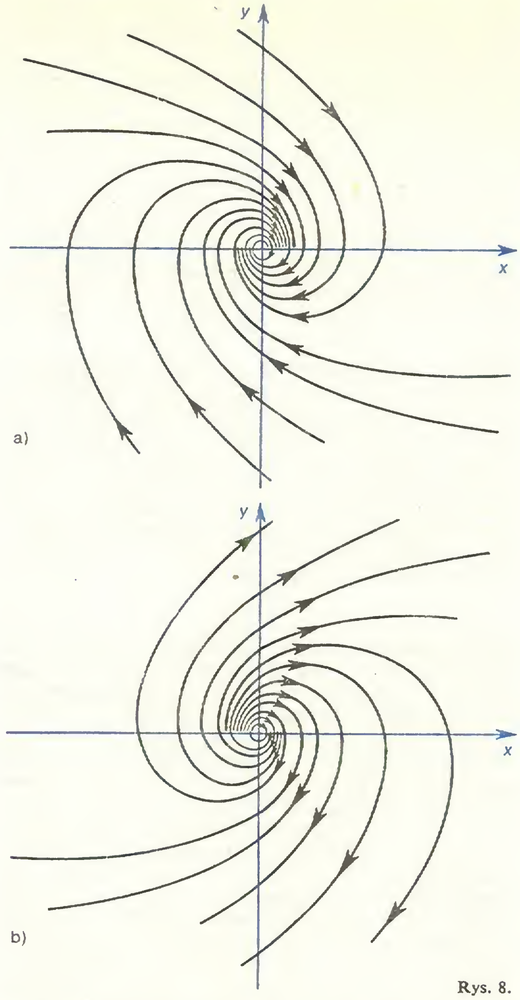
Rys. 5.



Rys. 6.



Rys. 7.



Rys. 8.

wracalnych — punkt stanowiący obraz stanu modelowanego układu oddala się po odpowiedniej trajektorii do nieskończoności lub do innego, trwałego punktu osobliwego, jeżeli taki istnieje na płaszczyźnie fazowej.

c) Rysunek 6 przedstawia przykład obrazu fazowego, którego punktem osobliwym jest siodło. Liniami pogrubionymi zaznaczono separatrysy układu. Łatwo zauważyć, że siodło jest nietrwałym punktem osobliwym. Przez analogię do poprzednich przykładów (a i b) można prześledzić losy punktu — obrazu na płaszczyźnie fazowej — w zależności od stanu początkowego układu.

d) Na rys. 7a podano przykład obrazu fazowego, którego punktem osobliwym jest tzw. środek. Przy omawianiu zastosowania maszyn analogowych w paragrafie poświęconym modelom matematycznym pokazano obraz fazowy tego typu otrzymany na ekranie oscyloskopu maszyny analogowej. Warto zauważyć, że zmienne dynamiczne takiego układu oscylują w czasie, periodycznie powracając do punktu początkowego, gdy odbędą pełny obieg po cyklu zamkniętym. Jeżeli układ modelowany zostanie wyprowadzony z tego cyklu (np. przez zwiększenie wartości zmiennej x — czy to przez doprowadzenie z zewnątrz pewnej ilości substancji, jeżeli x jest jej stężeniem, czy też przez doprowadzenie osobników określonego gatunku, jeżeli x jest liczebnością jego populacji), to punkt na płaszczyźnie fazowej rozpocznie ruch po nowej zamkniętej trajektorii. O takich trajektoriach mówimy, że są nietrwałe (o czym w rozdziale Modele oscylacyjne). Przy modelowaniu procesów spotykamy się niekiedy z tzw. cyklami granicznymi. Odpowiada to sytuacji, gdy trajektorie układu wychodzące z jakiegoś punktu osobliwego, np. niestabilnego węzła, dążą asymptotycznie, po dostatecznie długim czasie, do cyklu zamkniętego, a po osiągnięciu go obiegają taki cykl, co odpowiada trwałym oscylacjom zmiennych x i y. Sytuację taką przedstawia schematycznie rys. 7b.

e) Na rys. 8a, b podano przykłady obrazów fazowych układów, których punktami osobliwymi są odpowiednio ognisko trwałe i ognisko nietrwałe. Mamy tu do czynienia z oscylacjami obydwu zmiennych. Oscylacje te nie są jednak periodyczne, układ nie powraca do stanu początkowego, lecz dąży z biegiem czasu do punktu trwałego (a) lub do nieskończoności (b) czy innych punktów osobliwych. Podane przykłady stanowią ilustrację niektórych typów obrazów fazowych. Znajomość funkcji $P(x, y)$ i $Q(x, y)$ oraz odpowiedniego aparatu matematycznego umożliwia przeprowadzenie tzw. jakościowej analizy układu dynamicznego, stwierdzenie, czy istnieje i jaki ma charakter punkt osobliwy, oraz naszkicowanie przebiegu trajektorii fazowych, izoklin układu. Nawet gdy mamy do czynienia ze złożonym układem dynamicznym, można — nie poszukując jego pełnego, ścisłego rozwiązania — uzyskać bogate informacje o badanym układzie przeprowadzając jakościową analizę obrazu fazowego modelowanego układu.

Kilka przykładów zastosowania matematycznych modeli punktowych i ich analizy w zagadnieniach biologicznych podamy w rozdziale Wybrane przykłady modeli matematycznych w zagadnieniach biologicznych.

Modelowanie uwzględniające wiek obiektów

Modele takie stosuje się wówczas, gdy badane obiekty z biegiem czasu zmieniają swe własności. Uwzględnienie tego konieczne jest w zagadnieniach takich jak ekologia gatunków, wzrost i rozmnażanie komórek. Odpowiednio równania dynamiczne mają postać:

$$\frac{\partial x(t, T)}{\partial t} + \frac{\partial x(t, T)}{\partial T} = -Wx, \qquad (5)$$

gdzie x oznacza liczbę osobników w wieku T w chwili t, a W — współczynnik śmiertelności, który może zależeć od wieku T. (Symbol $\partial/\partial t$ oznacza tzw. pochodną cząstkową względem czasu). Lewa strona równania opisuje starzenie się osobników, prawa strona — ich śmierć. Uzupełnianie populacji jest uwzględnione w tzw. warunku granicznym:

warunek graniczny

$$x(t, 0) = \int_0^\infty \mu(T)x(t, T)dT, \qquad (6)$$

gdzie $\mu(T)$ jest współczynnikiem rozmnażania. W prostszych modelach stosuje się często rozkładanie populacji na grupy wiekowe. Równanie (6) przechodzi wówczas w układ równań, w których zmiennymi dynamicznymi są liczebności odpowiednich grup.

Modele niepunktowe, opisujące układy przestrzennie niejednorodne

Badanie zjawisk zachodzących w układach, w których zmienne i parametry charakteryzujące układ zależą nie tylko od czasu, ale i od współrzędnych przestrzennych, wymaga stosowania modeli niepunktowych. Aby uwzględnić niejednorodność przestrzenną i — w szczególności — zjawiska dyfuzji, wystarcza w układzie równań dynamicznych (1) dodać człony opisujące przepływ substancji w przestrzeni. Układ równań przybiera wówczas postać:

uwzględnienie niejednorodności przestrzennej

$$\frac{\partial x_i(t, \eta_1, \eta_2, \eta_3)}{\partial t} = P_i(x_1, x_2, \ldots, x_n) + D_i \sum_{j=i}^{n} \frac{\partial^2 x_i}{\partial \eta_j^2}, \qquad (7)$$

gdzie η_1, η_2, η_3 oznaczają współrzędne przestrzenne. Modele tego typu wykorzystuje się do opisu powstawania struktur przestrzennych, takich jak periodyczna struktura kręgosłupa czy warstwowa struktura gastruli w embriologii. Ważną właściwość modeli niepunktowych stanowi możliwość opisu przy ich pomocy procesu spontanicznego rozbicia pierwotnie jednorodnego obszaru na szereg jednorodnych podobszarów. Ma to miejsce wówczas, gdy w odpowiednim modelu punktowym (bez dodatkowych członów po prawej stronie równań) występują nietrwałe punkty osobliwe. W § Modele przełączania będzie mowa o przykładach takich modeli.

Znaczenie poszczególnych typów wyżej opisanych modeli zależy oczywiście od konkretnego zagadnienia, w jakim model jest stosowany, oraz od zadań stawianych przed modelem. Wydaje się jednak, że w obecnym stanie rozwoju biologii i biofizyki teoretycznej szczególnie ważną rolę odgrywają modele dynamiczne. Kilka takich przykładów podamy w rozdziale Wybrane przykłady modeli matematycznych w zagadnieniach biologicznych.

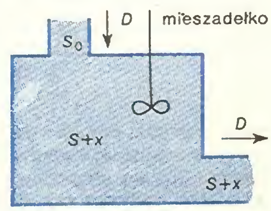
Wybrane przykłady modeli matematycznych w zagadnieniach biologicznych

Model hodowli ciągłej

model hodowli ciągłej mikroorganizmów

Jednym z najprostszych, a zarazem najszerzej stosowanych w praktyce modeli matematycznych jest model hodowli ciągłej mikroorganizmów. Model ten jest szczególnie przydatny przy planowaniu optymalnych warunków hodowli drobnoustrojów, których funkcją jest przemiana małowartościowego substratu na produkt bardziej dla nas wartościowy (np. parafiny na jadalne białka). Procesy takie na wielką skalę przeprowadza tzw. mikrobiologia przemysłowa. Podobne problemy występują w przemyśle farmaceutycznym oraz przy projektowanu układów umożliwiających długotrwałe przebywanie człowieka w przestrzeni kosmicznej.

W najprostszym wariancie modelu hodowli ciągłej uwzględnia się tylko zmiany w czasie stężenia substratu (S) oraz stężenia produkowanej biomasy (x). Rozważany proces ma następujący przebieg: do naczynia zwanego kultywatorem (rys. 9), w którym się odbywa proces, wprowadza się substrat o stężeniu S_0, z objętościową prędkością D. W kultywatorze znajdują się i rozmnażają interesujące nas w danym zagadnieniu drobnoustroje, a wypływa z naczynia (z taką samą prędkością objętościową D) mieszanina zawierająca nie zużyty substrat oraz drobnoustroje.



Rys. 9. Schemat kultywatora przepływowego

model matematyczny hodowli w warunkach przepływu

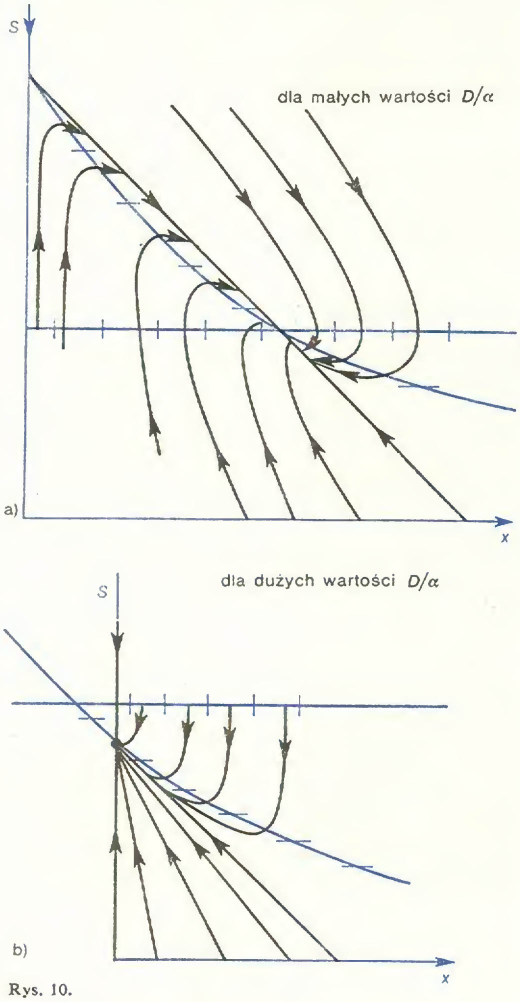
Model matematyczny hodowli ciągłej można zapisać w postaci układu dwóch równań różniczkowych

$$\frac{dx}{dt} = \mu(S)x - Dx,$$
$$\frac{dy}{dt} = \mu(S)x - DS + DS_0, \qquad (8)$$

gdzie

$$\mu(S) = \frac{\alpha S}{K_S + S}. \qquad (9)$$

Kształt zależności współczynnika μ od S jest znany z eksperymentalnych prac Monoda. Sens fizyczny tej zależności jest prosty: przy małych stężeniach substratu S współczynnik μ szybko się zmienia wraz z S, natomiast przy dużych stężeniach S, gdy substrat znajduje się w nadmiarze ($S \gg K_S$) — współczynnik $\mu(S)$ staje się praktycznie stały. Omawiana zależność jest podobna do znanego w kinetyce reakcji chemicznych prawa Michaelisa, opisującego zależność prędkości reakcji enzymatycznych od stężenia substratu. Możliwe są także inne postacie zależności μ od S. Obraz fazowy układu, którego model matematyczny stanowią równania (8) przedstawia rys. 10. Na rysunku zaznaczono główne izokliny I i II, punkt osobliwy układu jest węzłem trwałym. Analiza modelu prowadzi ponadto do wniosku, że istnieje pewna krytyczna wartość stężenia S_0, poniżej której hodowla ciągła staje się niemożliwa (mikroorganizmy wypłukują się z kultywatora). Analiza modelu pozwala ponadto wyznaczyć prędkość przepływu D oraz początkowe stężenie S_0, odpowiadające z góry danej efektywności produkcji, której miarą jest wielkość $Dx_s = (S_0 - S)/S$ (w warunkach stacjonarnych). Analizując dalej model można wreszcie uzyskać optymalny wariant pracy układu w początkowym stadium rozruchu.



Rys. 10.

Omówiony powyżej model jest modelem uproszczonym. W mikrobiologii przemysłowej stosowane są bardziej złożone modele, uwzględniające szereg innych czynników, takich jak efekt ciasnoty (wpływ oddziaływania między drobnoustrojami), rozkład hodowanych obiektów według wieku, przesunięcie w czasie (opóźnienie) zmian tempa rozmnażania w stosunku do wywołujących je zmian stężeń substratu. Jednakże u podstaw każdego z tych realnie stosowanych złożonych modeli leży przedstawiony powyżej uproszczony wariant modelu hodowli ciągłej.

Modele oscylacyjne

Typowymi przykładami modeli oscylacyjnych są modele dotyczące współistnienia gatunków, znane w literaturze pod nazwą modeli Lotki i Volterry. Przedstawiamy poniżej prosty model Volterry opisujący współistnienie populacji drapieżników i populacji ofiar przy założeniu, że ofiary (populacja (1)) mają pod dostatkiem pożywienia, pokarm zaś drapieżników (populacji (2)) stanowią przedstawiciele populacji (1).

model Volterry

Równania różniczkowe stanowiące model mają następującą postać:

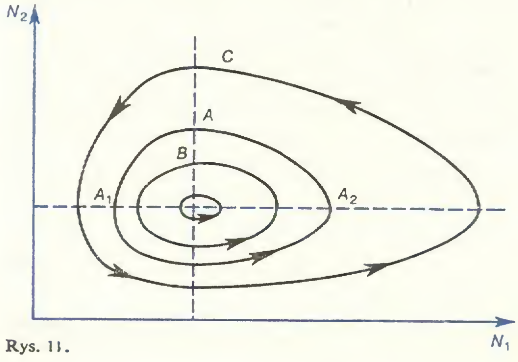
$$\frac{dN_1}{dt} = \alpha N_1 - \gamma N_1 N_2,$$
$$\frac{dN_2}{dt} = -\beta N_2 + \delta N_1 N_2. \qquad (10)$$

N_1 oznacza liczebność populacji (1) (ofiar), N_2 — populacji (2) (drapieżników). W modelu przyjęto założenie, że populacja (1) rozmnaża się ze stałą prędkością, ginie zaś w wyniku spotkań z drapieżnikami; ponadto — że populacja (2) może się rozmnażać tylko w warunkach, gdy nie brak jej pożywienia, którym są ofiary, ginie zaś tylko śmiercią naturalną. Założenia te łatwo można rozpoznać w poszczególnych członach układu równań.

Analiza zaproponowanego modelu nie jest trudna, wynik jej, w postaci obrazu fazowego układu, przedstawia rys. 11. Rozważany układ jest układem zachowawczym, ponieważ trajektorie układu są krzywymi zamkniętymi. Liczebność każdej populacji oscyluje w czasie; wynika to stąd, że procesy rozmnażania (zwiększania liczebności) populacji (2) są przesunięte w fazie, są opóźnione w stosunku do zmian ilości ich pożywienia, czyli populacji (1).

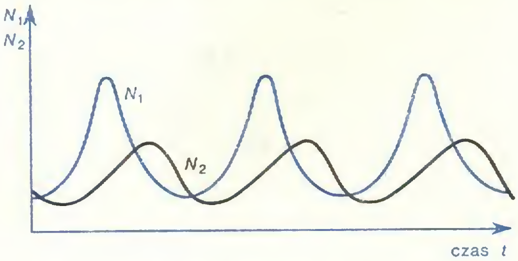
W przedstawionym tu, bardzo uproszczonym modelu pominięto działanie wszelkich czynników zewnętrznych, które w realnych warunkach mogą wywierać istotny wpływ na liczebność obydwu rozważanych populacji.

Warto zwrócić uwagę, że jednorazowe działanie czynników zewnętrznych, zmieniające liczebność jednego z gatunków, przenosi punkt na obrazie fazowym na inną trajektorię, przy czym efekt takiego działania zależy od fazy. I tak np. zwiększenie liczebności gatunku (1) w fazie A_1 (trajektoria A) przenosi układ na trajektorię wewnętrzną (B), o mniejszych oscylacjach, dla której maksymalna liczebność każdego gatunku jest mniejsza niż dla trajektorii A. Na odwrót — zwiększenie liczebności gatunku (1) w fazie A_2 przeprowadzi układ na trajektorię C, gdzie oscylacje



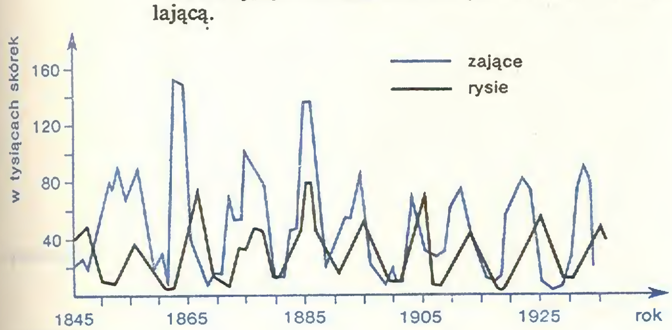
Rys. 11.

osiągają większe amplitudy, a maksymalna liczebność każdego gatunku jest większa niż dla początkowej trajektorii *A*. Takie oscylacje zaobserwowano w realnych warunkach, np. oscylacje liczebności populacji ryb w Morzu Śródziemnym oraz liczebności rysi i królików w Kanadzie. Drugi przykład zilustrowany jest wykresem (rys. 12 i 13), gdzie widoczny jest wyraźnie oscylujący przebieg liczebności obydwu omawianych gatunków, przy czym maksima liczebności ofiar wyprzedzają w czasie maksima liczebności drapieżników. Zauważmy, że zastosowany model matematyczny zawierał szereg upraszczających założeń oraz że dane przedstawione na rys. 13 dotyczą liczby skór zajęczych i rysich dostarczonych do punktów



Rys. 12. Zmiany w czasie liczby ofiar (N_1) i drapieżników (N_2) otrzymane w wyniku rozwiązania równania (10)

skupu przez myśliwych, co nie musi być ściśle proporcjonalne do liczebności populacji. Mimo tych zastrzeżeń zgodność przewidywań modelu z danymi obserwacyjnymi można uznać za jakościowo zadowalającą.



Rys. 13. Zmiany liczby skórek zajęczych i mysich w Kanadzie w latach 1845–1935

Natomiast porównanie ilościowe przewidywań wypływających z modelu z wynikami obserwacji wykazało, że model Volterry w wyżej podanej, najprostszej postaci nie daje dostatecznie dokładnego opisu zjawisk obserwowanych w badaniach ekologicznych. Wskazuje to na konieczność stosowania modeli bardziej złożonych, uwzględniających w szczególności czynniki takie, jak ograniczoność zapasów pożywienia ofiar, ograniczoność terenu, na którym współistnieją rozważane gatunki (efekt ciasnoty), i wiele innych. Modele takie istotnie szeroko się rozwija i opracowuje. Warto jednak podkreślić, że u podstaw tego rodzaju modeli leży przeważnie idea prostego, klasycznego modelu Volterry.

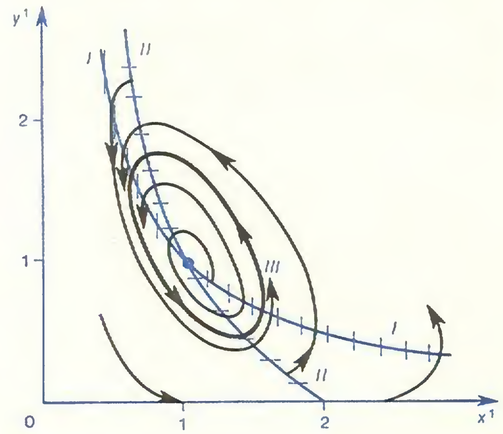
model procesu glikolizy

Model procesu glikolizy stanowi dobry przykład procesu wszechstronnie zbadanego, tzn. zarówno eksperymentalnie, jak i z punktu widzenia modeli matematycznych. Glikoliza jest jednym z najprostszych i dobrze zbadanych procesów odbywających się w cytoplazmie komórki. Proces ten polega na zamianie glukozy w kwas pirowinogronowy. Ponadto glikoliza odgrywa ważną rolę wśród procesów związanych z wytwarzaniem spirytusu etylowego. Przeprowadzono eksperymentalne badanie oscylacji w procesie glikolizy zarówno w cytoplazmie, jak i w układzie czystych enzymów i substratów glikolizy, przy czym w tej drugiej grupie eksperymentów otrzymano wiarogodne ilościowe dane doświadczalne (eksperymenty Chance'a i Hessa). W wyniku tych eksperymentów stwierdzono, że występują periodyczne zmiany w czasie stężeń produktów reakcji glikolizy.

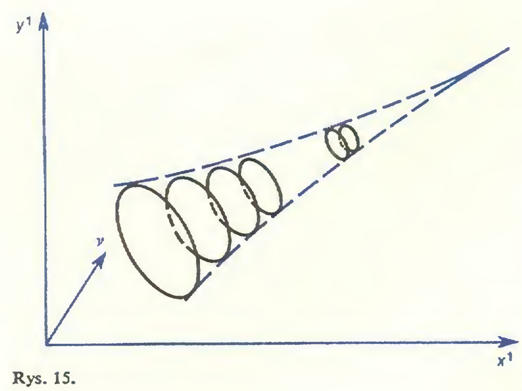
eksperymenty Chance'a i Hessa

Model matematyczny procesu zaproponował E. Sielkow. Za podstawę modelu przyjął ścisłe równania kinetyczne odnoszące się do całego łańcucha reakcji chemicznych stanowiących proces glikolizy i oparł się na znanych właściwościach enzymów biorących udział w tym procesie. Układ stanowiący model matematyczny składał się z wielu równań. Na podstawie reguł i procedur matematycznych wspomnianych w paragrafie. „Modele punktowe" udało mu się zredukować pierwotny model do 2 tylko równań różniczkowych z dwiema zmiennymi. W uproszczonym wariancie tego modelu, przeanalizowanym wcześniej przez Higginsa, zmiennymi dynamicznymi są: x, będące miarą stężenia sześciowęglowych cukrów, powstających ze stałą prędkością z glukozy (na początku łańcucha reakcji w procesie glikolizy), oraz y, będące miarą stężenia jednej z substancji — produktów końcowej części łańcucha procesu glikolizy. Nie podajemy tu pełnego zapisu modelu nawet w prostszym wariancie Higginsa, raczej kładziemy nacisk na jakościową stronę zagadnienia.

Obraz fazowy analizowanego układu po wprowadzeniu bezwymiarowych zmiennych x' i y' (proporcjonalnych odpowiednio do x i y) przedstawia rys. 14. Na rysunku zaznaczono główne izokliny *I* i *II* oraz krzywą zamkniętą *III*, stanowiącą tzw. cykl graniczny. Bardziej wnikliwa analiza modelu pozwoliła stwierdzić, że w wyniku zmian wartości parametrów występujących w modelu, a w szczególności — szybkości v, dopływu glukozy, model przewiduje bądź trwały, stabilny przebieg procesu (przy dużych wartościach v), bądź też oscylacyjny charakter procesu (przy małych



Rys. 14.



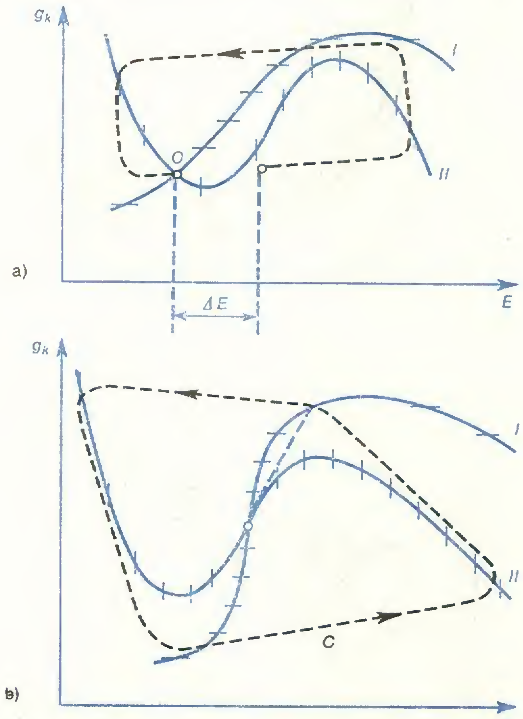
Rys. 15.

v), co odpowiada wspomnianemu już cyklowi granicznemu. Rysunek 15 ilustruje jakościowo tę zależność przebiegu procesu od parametru v. Wyniki modelu porównano z danymi eksperymentalnymi. Stwierdzono słuszność przewidywań modelu dotyczących przechodzenia układu od zachowania stabilnego do oscylującego oraz związku tych przejść ze zmianami wartości parametrów układu. Analiza opisanego modelu ma też bardziej ogólny aspekt; stanowi mianowicie pewien krok naprzód w zrozumieniu mechanizmu regulacji procesów biochemicznych. Nieco bardziej złożony wariant modelu zastosowano do opisu związku między procesem glikolizy i oddychaniem, co umożliwiło wyjaśnienie efektów włączania i wyłączania procesu glikolizy w warunkach anaerobowych (beztlenowych) i aerobowych (z dostępem tlenu).

model generacji impulsu nerwowego

Model generacji impulsu nerwowego zaproponowany przez Hodgkina i Huxleya w 1953 r. jest przykładem modelu, w którym punkt osobliwy jest węzłem. Model ten opisuje przechodzenie jonów potasu (K) i sodu (Na) przez membranę komórek nerwowych oraz zmiany potencjału membrany; zmiany potencjału membrany są uwarunkowane prądem elektrycznym, a zmiany przewodnictwa — zmianami stanu elementów przewodzących prąd w membranie. Nośnikami prądu są jony potasu i sodu, przy czym zmiany przewodnictwa jonów sodu są znacznie szybsze niż jonów potasu.

Wykorzystując reguły matematyczne, które prowadzą do redukcji liczby równań stanowiących model procesu, można sprowadzić model do dwóch tylko równań, w których zmiennymi będą napięcie E oraz przewodnictwo jonów potasu g_K. Obraz fazowy rozważanego układu, reagującego na wzbudzenie standardowego impulsu, przedstawia schematycznie rys. 16a. Przy zmianie napięcia o wartość E (tzw. wzbudzenie progowe) punkt-obraz na płaszczyźnie fazowej przechodzi do obszaru nietrwałego i porusza się po trajektorii oznaczonej krzywą przerywaną, po czym wraca do stanu początkowego. Trajektoria ta na niektórych odcinkach jest prawie równoległa do osi odciętych, co oznacza, że czas zmiany potencjału, zależny od elektrycznej pojemności membrany, jest



Rys. 16.

znacznie krótszy od czasu charakterystycznego dla zmian przewodnictwa. Rysunek 16a daje zatem obraz generacji impulsu nerwowego w normalnych warunkach. Przy zmianie parametrów związanych z elektrycznymi charakterystykami (np. z charakterystyką prądowo-napięciową) izokliny główne (*I* i *II*) na obrazie fazowym zmieniają swe położenie tak, że punkt osobliwy może przejść do obszaru nietrwałego. Sytuację tę ilustruje rys. 16b. Powstają wówczas nieustające drgania relaksacyjne, odbywające się wzdłuż cyklu *C*, zaznaczonego na rysunku. Zdarza się to w warunkach patologicznych. Znaczenie tego modelu dla medycyny polega przede wszystkim na tym, że umożliwia on opis procesów zarówno w stanach normalnych, jak i w sytuacjach patologicznych, a zarazem wskazuje na możliwe przyczyny efektów patologicznych.

model stanu patologicznego

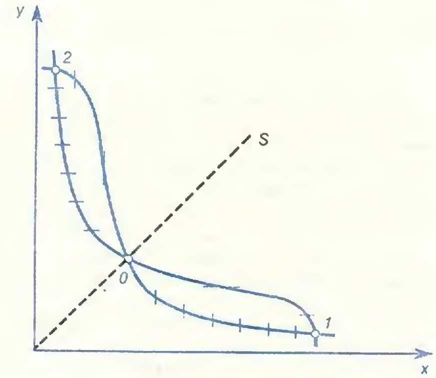
Opisany powyżej model generacji impulsu nerwowego należy do modeli punktowych, tzn. opisuje lokalną sytuację odpowiadającą małemu odcinkowi włókna nerwowego. Analiza zjawisk związanych z rozchodzeniem się impulsu nerwowego wymaga uwzględnienia w modelu matematycznym członu wyższego rzędu, odpowiadającego zmianom napięcia wzdłuż zmiennej przestrzennej; model staje się wówczas niepunktowy.

Modele przełączenia (tryggerowe)

Modele tego rodzaju są stosowane przy opisie procesów przełączeniowych w biologii, a zwłaszcza przy opisie sytuacji, w których obiekt biologiczny może dokonywać wyboru między dwoma możliwymi stanami (reżimami) pracy. Z modeli tego rodzaju korzysta się np. przy opisie procesów różnicowania tkanek w embriogenezie, procesów powstawania i różnicowania gatunków w ewolucji, procesów związanych z powstawaniem informacji biologicznej (we wczesnych stadiach ewolucji), a także przy opisie powstania asymetrii optycznej w biologii. Najważniejszą cechą modeli tego typu jest występowanie w nich co najmniej dwóch stanów stacjonarnych, oddzielonych od siebie na obrazie fazowym nietrwałą trajektorią (rys. 17).

embriogeneza

Do bardziej szczegółowo zbadanych biologicznych modeli przełączania należy model opisujący różnicowanie tkanek. Podstawowy problem wczesnej embrio-



Rys. 17.

genezy można przedstawić w uproszczonej, wyidealizowanej postaci w następujący sposób: dwie (sąsiednie) komórki znajdujące się w jednakowych warunkach i zawierające identyczną informację genetyczną po procesie różnicowania zaczynają pracować w różnych reżimach. Sytuacja wygląda zatem tak, jak gdyby została naruszona symetria przestrzenna początkowych i granicznych warunków dla stanu początkowego. Analiza odpowiedniego modelu matematycznego pozwala prześledzić proces różnicowania i wskazać przyczynę wspomnianego naruszenia symetrii. Modele matematyczne opisujące to zjawisko wychodzą

z założenia przełączeniowego (tryggerowego) schematu regulacji procesu biosyntezy, zaproponowanego przez Jacoba i Monoda. W schemacie tym przyjmuje się, że metabolity związane z jednym reżimem (białka oraz mRNA) uniemożliwiają syntezę metabolitów związanych z działaniem układu w drugim reżimie — i vice versa. Można to określić jako antagonistyczne oddziaływanie między metabolitami charakterystycznymi dla dwóch rozważanych reżimów. Analiza odpowiednio sformułowanego modelu matematycznego, w którym zmiennymi są stężenia dwóch wyżej wspomnianych grup metabolitów, prowadzi do obrazu fazowego przedstawionego na rys. 17, gdzie zaznaczono schematycznie kształt dwóch głównych izoklin oraz linię rozdzielającą (separatrysę) *S*. Układ biologiczny, którego stanowi odpowiada punkt-obraz po prawej stronie separatrysy, dąży nieuchronnie do stanu trwałego *1*, układ zaś w stanie początkowym po lewej stronie *S* dąży do stanu trwałego *2*. Komórka znajdująca się początkowo w stanie *0* (punkt stacjonarny nietrwały), bądź w stanie odpowiadającym jakiemuś punktowi na separatrysie może przejść do stanu *1* lub *2* pod wpływem dowolnie słabego czynnika zewnętrznego, przy czym przejście takie jest procesem nieodwracalnym.

Losy sąsiedniej (II) komórki, która jest związana z poprzednio omawianą (I) przez krążenie i wymianę metabolitów i substratów, zależą od tego, w jakim z dwóch możliwych reżimów działa komórka (I), która porzuciła stan nietrwały. Jeżeli wymiana substratów jest procesem szybkim, a dopływ ich jest ograniczony, to komórka (II) przechodzi do stanu przeciwnego niż komórka (I). Powstaje w ten sposób pewna struktura — układ komórek działających w różnych reżimach. W zależności od takich parametrów, jak zdolność do przenikania metabolitów i substratów między sąsiadującymi z sobą komórkami, mogą powstawać różne struktury, złożone z pojedynczych komórek lub ich zespołów, przy czym komórki w takim zespole pracują w jednym reżimie. Powstawanie tego rodzaju struktur nazywamy aktem różnicowania tkanek, oczywiście w bardzo schematycznym i uproszczonym ujęciu. Analiza powstawania struktur w wypadku bardzo dużej liczby komórek wymaga stosowania modeli niepunktowych. Warto podkreślić, że powstanie zróżnicowania, przestrzennej niejednorodności, jest wynikiem nietrwałości punktu osobliwego w odpowiednim modelu punktowym.

model powstawania i różnicowania gatunków

Podobne właściwości charakteryzują modele opisujące powstawanie i różnicowanie gatunków biologicznych. Z nietrwałym symetrycznym stanem stacjonarnym spotykamy się, gdy próbujemy skonstruować model matematyczny powstawania asymetrii optycznej w przyrodzie. Stwierdzono mianowicie, że aminokwasy wchodzące w skład żywych organizmów na Ziemi są optycznie prawoskrętne, cukry zaś — lewoskrętne. Wydaje się oczywiste, że żywe organizmy zawierające aminokwasy i cukry o przeciwnej aktywności optycznej mogłyby z równym powodzeniem żyć i rozwijać się, niemniej jednak nie występują na Ziemi. Tworząc model, który wyjaśnia powstawanie takiej asymetrii w przyrodzie, wychodzimy z założenia o antagonistycznym (wzajemnie zgubnym) oddziaływaniu organizmów zawierających odpowiednie substancje o przeciwnej aktywności optycznej. Stwierdzono bowiem, że lewoskrętne aminokwasy nie tylko nie są przyswajane, lecz nawet stanowią jakby truciznę dla organizmów zawierających prawoskrętne aminokwasy i — oczywiście — na odwrót. Wynika stąd nietrwałość stanu, w którym współistniałyby w równych ilościach lewo- i prawoskrętne żywe obiekty. Niewielka choćby przewaga liczbowa jednych nad drugimi prowadzi nieuchronnie do całkowitego zaniku organizmów drugiego typu. Można zatem sądzić, że jeżeli nawet we wczesnym stadium ewolucji powstały dwa „zwierciadlane" rodzaje żywych organizmów, to z biegiem czasu musiał pozostać praktycznie tylko jeden.

asymetria w przyrodzie

uniwersalność kodu genetycznego

Analogicznie rozumujemy, gdy chodzi o podstawowy dla współczesnej biofizyki problem, sprowadzający się do pytania, dlaczego kod genetyczny jest uniwersalny (identyczny w całej przyrodzie), mimo że można sobie wyobrazić inne, równoważne kody, w których np. aminokwasy byłyby poprzestawiane w porównaniu z obecnie istniejącym wariantem kodu (najkrócej — kodem można by nazwać ustaloną odpowiedniość między aminokwasami a trójkami nukleotydów). Odpowiedzi na postawione pytanie można szukać w antagonistycznym oddziaływaniu obiektów charakteryzujących się różnymi kodami. Wynikiem oddziaływań między nimi byłoby naruszenie regularnej syntezy białek u obydwu obiektów, co z kolei musiałoby być zgubne dla obydwu partnerów.

Model matematyczny uwzględniający to zjawisko przewiduje, że stacjonarny stan symetryczny jest nietrwały, co w konsekwencji prowadzi do nieuchronnego wyginięcia wszystkich żywych obiektów prócz pomyślnie rozwijającej się jednej grupy, charakteryzującej się jednym określonym kodem. Nie sposób jednak z góry przewidzieć, który kod odniesie zwycięstwo; zależy to od rodzaju i kierunku przypadkowej fluktuacji, wyprowadzającej zespół żywych organizmów we wczesnym stadium ewolucji z nietrwałego stanu początkowego, w którym współistniały żywe obiekty o różnych kodach. Wybór jednego wariantu spośród wielu możliwości jest równoważny z powstaniem uniwersalnego alfabetu dla całej żywej przyrody, stanowi zatem podstawę powstania informacji biologicznej. Aktowi powstawania informacji towarzyszą bowiem dwa podstawowe procesy: wybór (decyzja) i zapamiętanie. Obydwa te procesy występują w układach przełączania (tryggerowych).

Inne rodzaje zastosowania modelowania matematycznego

Na zakończenie przeglądu zagadnień związanych z modelowaniem matematycznym procesów biologicznych warto poruszyć jeszcze dwa tematy nie omawiane w zasadniczym tekście artykułu.

Układy otwarte

Uważano dawniej, że termodynamika układów otwartych pomoże w znalezieniu ogólnych zasad formułowania modeli matematycznych w zagadnieniach biologicznych. Okazało się jednak, że termodynamika układów bliskich stanu równowagi napotyka poważne trudności przy próbach zastosowania takiego podejścia do procesów biologicznych (próby takie podejmował J. Prigogine, który wraz z grupą innych naukowców pracuje od całego szeregu lat nad problemami termodynamiki tzw. układów otwartych), zwłaszcza przy opisie reżimów oscylacyjnych. Termodynamika układów dalekich od stanu równowagi nie jest jeszcze dostatecznie rozwiniętą dziedziną. Można sądzić, że analiza konkretnych modeli, szczególnie niektórych spośród omawianych w niniejszym artykule, będzie pewną pomocą w znalezieniu ogólnych zasad termodynamiki układów dalekich od równowagi. Można mieć nadzieję, że w razie pomyślnego rozwoju tego kierunku zasady takie będą pożyteczne i owocne przy formułowaniu nowych modeli procesów biologicznych.

termodynamika układów otwartych

Dziedziny mniej związane z biofizyką

zastosowanie w medycynie

Przykład tego kierunku badań może stanowić wykorzystanie modeli matematycznych w medycynie, gdzie sformułowanie i analiza odpowiedniego modelu pomaga niekiedy w wyborze optymalnego wariantu

procedury leczenia środkami farmaceutycznymi lub fizykoterapeutycznymi. Rozkład w czasie dozowania stosowanych leków czy działań terapeutycznych ma niekiedy podstawowe znaczenie, zwłaszcza gdy przebieg choroby ma charakter oscylujący. W takich wypadkach ten sam środek stosowany w różnych momentach (w różnych fazach) może prowadzić do istotnie różnych, a niekiedy wręcz przeciwnych rezultatów. Jako przykład takiej sytuacji można wymienić wykorzystanie modelowamia matematycznego w strategii leczenia periodycznie występujących nawrotów np. malarii.

strategia szczepień

Często stosuje się modele matematyczne w zagadnieniach epidemiologii, gdzie przedmiotem modelu jest rozprzestrzenianie się i skutki choroby zakaźnej, bądź strategia postępowania przy zwalczaniu choroby, np. przez masowe szczepienia ochronne. Jako przykład można tu wymienić modele mające na celu wybór optymalnej strategii szczepień przeciwko różyczce, która — w przypadku kobiet we wczesnym okresie ciąży — może prowadzić do poważnych wad rozwojowych u potomstwa. Szczepienie ochronne zapobiega zapadnięciu na różyczkę w ciągu pewnego ograniczonego czasu po szczepieniu. Skrajne warianty strategii stosowania szczepień ochronnych przewidują poddawanie szczepieniu albo tylko dorastających dziewcząt, albo powszechne szczepienia dzieci obojga płci; ponadto istnieje szereg wariantów pośrednich, takich jak np. szczepienia powtarzane co kilka lat w wybranych populacjach dzieci i młodzieży. Uwzględnienie w modelu matematycznym konkretnego wariantu strategii szczepień ochronnych takich efektów, jak zaraźliwość choroby w danej populacji, trwałość i skuteczność szczepienia i wreszcie kosztów związanych z przeprowadzaniem szczepień, pozwala na ocenę skuteczności danego wariantu strategii postępowania; ilościową miarą tej skuteczności może być np. średnia wartość prawdopodobieństwa niezapadnięcia na różyczkę we wczesnym okresie ciąży. Wynikiem analizy modelu odpowiadającego określonemu wariantowi strategii postępowania jest oszacowanie jego skuteczności i kosztu; tego rodzaju dane pozwalają na podjęcie decyzji o wyborze lub odrzuceniu danej strategii postępowania.

strategia wykorzystywania bogactw

Inny przykład dziedziny, w której się coraz częściej stosuje modelowanie matematyczne, stanowi strategia wykorzystywania naturalnych bogactw czerpanych z fauny i flory naszego globu (np. optymalizacja odłowów rybnych w basenach morskich czy długoterminowe prognozy dotyczące możliwości zagwarantowania pożywienia stale rosnącej liczbie mieszkańców Ziemi).

Rola modeli matematycznych wzrosła ostatnio w związku z szybkim rozwojem techniki i przemysłu, a co za tym idzie — ze stale wzrastającym wpływem działania człowieka na biosferę. Uniknąć katastrofalnych skutków tego wpływu można tylko przy właściwej strategii gospodarowania człowieka na Ziemi i właściwym wpływaniu na biosferę. Nie sposób wyobrazić sobie poważnych prac w tej dziedzinie bez stosowania modeli matematycznych.

D. S. Czernawski *Model matematyczny powstawania życia*, Post. Fiz., 26, 153 (1975); D. S. Czernawski i in. *Co to jest biofizyka matematyczna?* Warszawa 1974; I. Prigogine i in. *Termodynamika ewolucji*, Post. Fiz., 26, 253 (1975); I. Rashevsky *Mathematical Biophysics*; *Physico-Mathematical Fundations in Biology*, New York 1960.

FIZYKA ZIEMI

Fizyka skorupy i wnętrza Ziemi · Fizyka atmosfery · Fizyka przestrzeni okołoziemskiej · Magnetyzm ziemski · Fizyka morza (dynamika, optyka, akustyka) · Fizyka wód śródlądowych · Geofizyka poszukiwawcza ·

Fizyka skorupy i wnętrza Ziemi

Renata Dmowska

W latach sześćdziesiątych i siedemdziesiątych naszego stulecia nastąpił znaczny rozwój wiedzy o wnętrzu Ziemi. Obraz procesów zachodzących w jej wnętrzu i kształtujących zjawiska powierzchniowe zaczął się nareszcie wyłaniać z gęstwiny danych i teorii, zarówno tych znanych od lat, jak i najnowszych. Nastąpiło to dzięki wprowadzeniu do geofizyki w ciagu ostatnich kilkunastu lat nowych metod doświadczalnych i teoretycznych oraz możliwości prowadzenia badań na dużo szerszą skalę niż to było możliwe dotychczas. Początkiem przełomu było zgromadzenie Międzynarodowej Unii Geodezji i Geofizyki w Helsinkach w 1960 r., gdzie postanowiono skoncentrować uwagę badaczy na zewnętrznej warstwie Ziemi o grubości 1000 km, obejmującej skorupę i górny płaszcz. Przedstawiony wówczas ogólnoświatowy program badań geofizycznych skorupy i górnego płaszcza, tzw. Projekt Górnego Płaszcza (ang. Upper Mantle Project), był stymulatorem wielu interesujących odkryć i teorii. Po raz pierwszy przeprowadzono systematyczne badania różnic w budowie górnego płaszcza, rejonu, w którym koncentrują się najistotniejsze procesy geodynamiczne zachodzące we wnętrzu Ziemi i mające bezpośredni wpływ na zjawiska na jej powierzchni. Zgromadzona ogromna ilość danych dotyczących skorupy i górnego płaszcza umożliwiła powstanie globalnych teorii geodynamicznych.

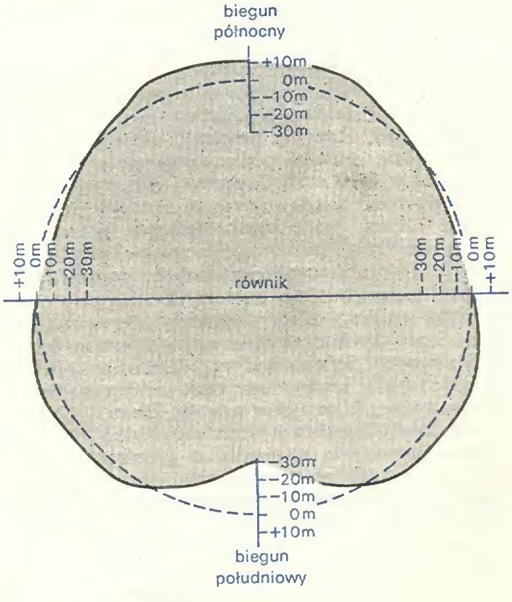
Na przełomie lat 1971 i 1972 zamknięto Projekt Górnego Płaszcza i przystąpiono do realizacji nowego, szerokiego programu badań, nazwanego Projektem Geodynamicznym (ang. *Geodynamical Project*). Celem programu jest kontynuowanie badań zjawisk zachodzących we wnętrzu Ziemi oraz, o ile to możliwe, skonstruowanie jednolitej, globalnej teorii, opisującej procesy geodynamiczne przekształcające naszą planetę od momentu jej powstania aż do chwili obecnej.

kształt Ziemi

Kształt Ziemi opisywany jest zwykle przy użyciu dwu powierzchni: sferoidy, tzn. elipsoidy obrotowej możliwie najlepiej odpowiadającej powierzchniom morza, i geoidy, tzn. powierzchni jednakowego potencjału siły ciężkości. Spłaszczenie sferoidy f określone jest wzorem:

$$f = (a-c)/a,$$

gdzie a jest średnim promieniem równikowym i wynosi 6378,160 km, c — promieniem biegunowym Ziemi i wynosi 6356,775 km. Spłaszczenie sferoidy wynosi 1/298,25. Ta ostatnia wielkość została wyznaczona z obserwacji orbit sztucznych satelitów Ziemi. Metodami geodezji satelitarnej stwierdzono poza tym, że geoida ma kształt gruszkowaty. Gruszkowate odchylenia geoidy od sferoidy wynoszą mniej niż 20 m, natomiast wybrzuszenie równikowe sferoidy wynosi ponad 20 km. Średnie odchylenia geoidy od sferoidy wzdłuż dowolnego południka ilustruje rys. 1; natomiast rys. 2 pokazuje mapę odchyleń geoidy od sferoidy sporządzoną na podstawie danych satelitarnych i powierzchniowych pomiarów grawimetrycznych. Przewyższenia geoidy nad sferoidą odpowiadają większym gęstościom materiału we wnętrzu Ziemi, a obniżenia — obszarom o mniejszych gęstościach materiału, przy czym anomalie gęstości dotyczą przypuszczalnie płaszcza Ziemi.

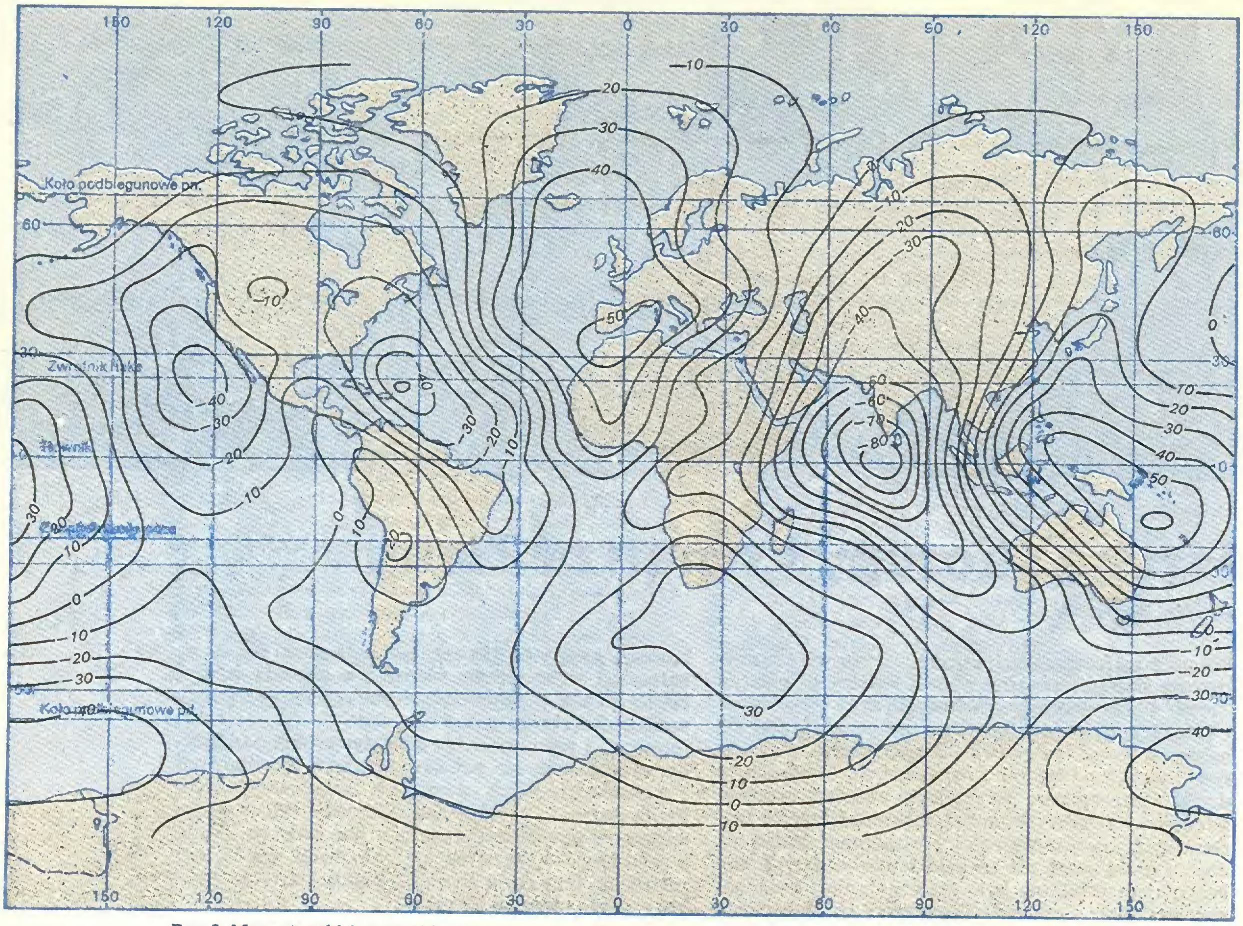


Rys. 1. Odchylenia geoidy od sferoidy

Budowa Ziemi

sejsmologia

Gałąź geofizyki, dotycząca powstawania i propagacji fal we wnętrzu Ziemi, zwana sejsmologią, dostarczyła podstawowych danych o budowie jej wnętrza, prawie całkowicie niedostępnego badaniom bezpośred-



Rys. 2. Mapa ziemskich anomalii grawimetrycznych. Wartości anomalii podane w metrach, w odniesieniu do sferoidy o spłaszczeniu 1/298,25

nim. Powstała na początku bieżącego stulecia sejsmologia instrumentalna umożliwiła rejestrację, a następnie analizę pola falowego generowanego w ogniskach trzęsień Ziemi. Rozkład prędkości fal sejsmicznych z głębokością ujawnił podstawowe cechy budowy Ziemi; już w 1910 r. zbudowano trójwarstwowy model jej wnętrza, składający się ze skorupy, płaszcza i jądra. Dalsze prace sejsmologiczne, uzupełnione analizą ziemskiego pola magnetycznego i powierzchniowych anomalii grawimetrycznych, doprowadziły do rozpoznania bardziej subtelnych szczegółów wewnętrznej struktury Ziemi i ujawniły cały szereg granic oddzielających obszary o różnych własnościach mechanicznych.

sejsmologia współczesna

Sejsmologia współczesna umożliwiła wykrycie wielu szczegółów budowy horyzontalnej górnego tysiąca kilometrów wnętrza Ziemi, a obecnie ujawnia ona, dotychczas częściowo tylko zbadane, wielosetkilometrowe anomalie o granicach pionowych w górnym płaszczu Ziemi, zarysy których, wbrew oczekiwaniom, nie pokrywają się z zarysami struktur skorupy ziemskiej.

Własności fizyczne wnętrza Ziemi

Określenie budowy i własności fizycznych wnętrza Ziemi jest wynikiem rozwiązania tzw. zagadnienia odwrotnego, tzn. określenia struktury wewnętrznej za pomocą obserwacji dokonywanych na powierzchni i, dodatkowo, dzięki wykorzystywaniu wyników laboratoryjnych pomiarów własności fizycznych znanych skał i minerałów w warunkach wysokich ciśnień i temperatur. Większość informacji pochodzi z analizy grawitacyjnego pola Ziemi oraz z obserwacji sejsmologicznych, ponadto z badań powierzchniowego strumienia cieplnego i z obserwacji magnetycznego pola Ziemi. Dane dotyczące wnętrza Ziemi, a w szczególności jej najgłębszych warstw, powinny być traktowane z dużą ostrożnością i uważane raczej za pewne oszacowania niż dane dokładne.

własności mechaniczne

Analiza objętościowych i powierzchniowych fal sejsmicznych, a także swobodnych drgań Ziemi, generowanych przez wyjątkowo silne trzęsienia Ziemi, prowadzi do określenia rozkładów modułu ściśliwości K, modułu ścinania μ i gęstości ϱ w funkcji promienia ziemskiego r. Dodatkowo jest wykorzystywane równanie Adamsa–Williamsona, wiążące prędkości fal sejsmicznych v_P i v_S z gęstością ϱ:

równanie Adamsa–Williamsona

$$\frac{d\varrho}{dr} = \frac{g\varrho}{v_P^2 - {}^4/_3 v_S^2};$$

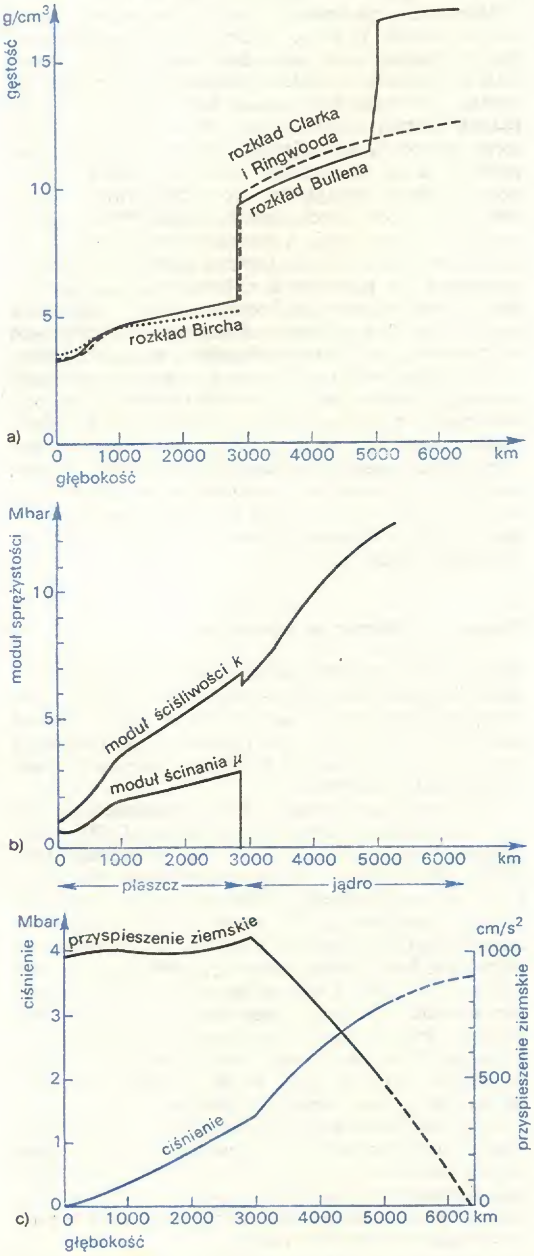
równanie to dotyczy jednak ośrodka jednorodnego ściskanego adiabatycznie w polu naprężeń hydrostatycznych i nie może być stosowane w obszarach przejść fazowych oraz w rejonach, w których rozkład temperatury nie jest adiabatyczny. Ze znanych rozkładów ϱ obliczone są rozkłady przyspieszania ziemskiego g i ciśnienia p we wnętrzu Ziemi, wszystkie te parametry przedstawione są na rys. 3 w funkcji głębokości. Należy tu zauważyć, że konstruowanie modeli dotyczących rozkładu gęstości we wnętrzu Ziemi opiera się na różnego rodzaju dodatkowych założeniach i hipotezach co do własności fizykochemicznych wnętrza Ziemi, gdyż rozwiązanie jednoznaczne tego problemu nie jest możliwe.

Informacje dotyczące rozkładu temperatury we wnętrzu Ziemi są silnie ograniczone. Znajomość rozkładu temperatury nie jest możliwa bez pełnej znajomości rozkładu źródeł ciepła i mechanizmów przeno-

rozkład temperatury

szenia ciepła, których to informacji nie jesteśmy w stanie otrzymać, jednakże określenie w pewnym zakresie rozkładu temperatury we wnętrzu Ziemi jest możliwe. Jako dane wyjściowe służą wyniki pomiarów powierzchniowego strumienia cieplnego i założone rozkłady pierwiastków promieniotwórczych oraz przewodnictwa cieplnego. Ponieważ średni strumień cieplny na oceanach i kontynentach jest jednakowy, przy równoczesnym różnym rozkładzie pierwiastków promieniotwórczych (grubsza kontynentalna skorupa ziemska sugeruje obecność większej ilości tych pierwiastków pod kontynentami), rozkłady temperatury dla oceanów i kontynentów są zwykle różne do głębokości kilkuset kilometrów, na której to głębokości temperatura jest przypuszczalnie jednakowa pod kontynentami i oceanami.

Ponieważ materiał płaszcza, z wyjątkiem pewnych lokalnych inkluzji, znajduje się w stanie stałym, temperatura topnienia materiału stanowi niewątpliwie górną granicę rozkładu temperatury w płaszczu.



Rys. 3. Rozkład we wnętrzu Ziemi: a) gęstości, b) modułu sprężystości, c) przyspieszenia ziemskiego i ciśnienia

Teoria ciała stałego wiąże temperaturę topnienia T_m ośrodka jednorodnego z prędkościami fal w tym ośrodku v_P i v_S relacją:

$$T_m = C(1/v_P^3 + 1/v_S^3)^{-1/3}.$$

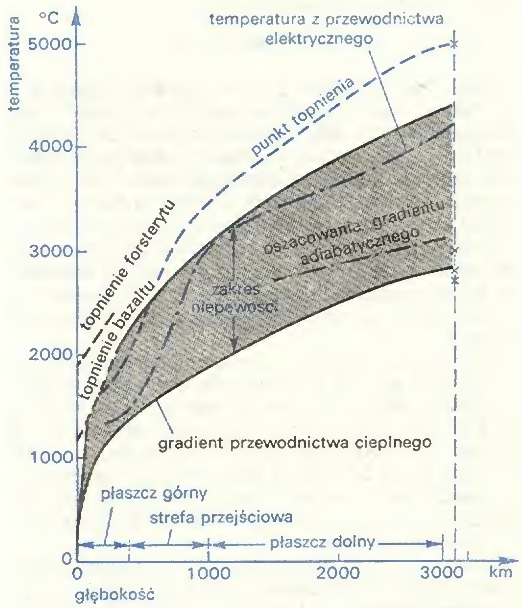
Związek ten pozwala na określenie temperatury topnienia aż do granicy płaszcz–jądro i pomimo, iż równanie to nie jest prawdziwe w rejonie przejść fazowych, gdyż współczynnik C silnie zależy od masy cząsteczkowej, uzyskana krzywa (Uffen, 1952 r.) stanowi dosyć realistyczne oszacowanie górnej granicy temperatury w dolnym płaszczu Ziemi.

Temperatura na granicy płaszcz–jądro może być określona przez ekstrapolację temperatury topnienia żelaza w obszar wysokich ciśnień. Metoda ta opiera się na wykorzystaniu empirycznego związku między temperaturą topnienia a ciśnieniem (podanego przez Simona w 1953 r.):

$$p = a[(T-T_0)^c - 1],$$

gdzie a i c są stałymi, a T_0 jest temperaturą odniesienia. Otrzymane wyniki są jednak wątpliwe, gdyż jądro Ziemi prawdopodobnie nie jest zbudowane z czystego żelaza.

Rozkład temperatury we wnętrzu Ziemi może być również uzyskany z danych o przewodnictwie elektrycznym, jednakże przy bardzo silnym i nierealnym założeniu co do stałości fizykochemicznego składu płaszcza Ziemi. Wszystkie omówione tu oszacowania temperatury we wnętrzu Ziemi przedstawiono na rys. 4.



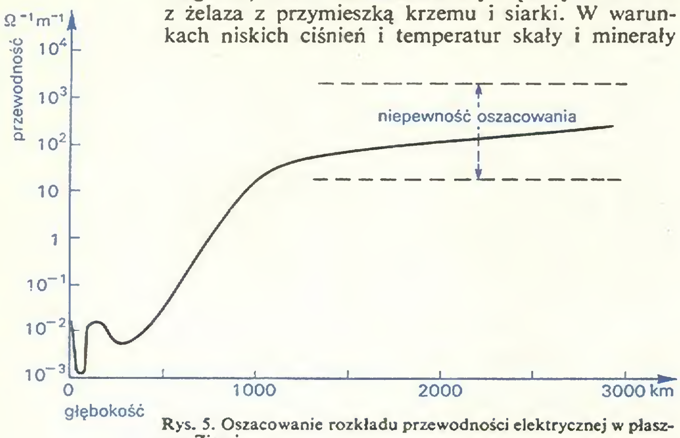
Rys. 4. Oszacowanie rozkładu temperatury w płaszczu Ziemi

własności elektryczne

Rozkład przewodności elektrycznej we wnętrzu Ziemi, uzyskany przez analizę zmian magnetycznego pola ziemskiego oraz przez zastosowanie metody magnetotellurycznej, pozwalającej określać przewodność elektryczną skał w zakresie głębokości 10–100 km, przedstawiony jest na rys. 5. Ostatnia z wymienionych metod polega na równoczesnym pomiarze pola magnetycznego i indukowanego w Ziemi pola elektrycznego.

Przewodność elektryczna wody morskiej wynosi około 4 $(\Omega\ m)^{-1}$, a nasyconych wodą skał osadowych od ok. 10^{-3} do 1 $(\Omega\ m)^{-1}$. Przewodność suchych skał skorupy, leżących poniżej skał osadowych, wynosi od ok. 10^{-6} do 10^{-3} $(\Omega\ m)^{-1}$. Zewnętrzne warstwy Ziemi są więc słabo przewodzące, oprócz silnie przewodzącej warstewki stworzonej przez oceany i skały osadowe.

Płaszcz Ziemi do głębokości ok. 400 km jest zbudowany prawdopodobnie z krzemianów, wśród których dominuje bogaty w magnez oliwin; w dolnym płaszczu fazę stabilną tworzą prawdopodobnie tlenki magnezu, żelaza i metakrzemiany. Jądro jest złożone z żelaza z przymieszką krzemu i siarki. W warunkach niskich ciśnień i temperatur skały i minerały



Rys. 5. Oszacowanie rozkładu przewodności elektrycznej w płaszczu Ziemi

zachowują się praktycznie jak izolatory, tzn. ich pasmo walencyjne jest wypełnione, a pasmo przewodnictwa oddzielone od pasma walencyjnego skończoną przerwą energetyczną. Ze względu na aktywacyjny charakter przewodnictwa zależność przewodności σ od temperatury jest następująca:

zależność przewodności od temperatury

$$\sigma = \sigma_0 e^{-A/kT},$$

gdzie T jest temperaturą, k — stałą Boltzmanna, a σ_0 — pewną stałą. Zależność tę potwierdzono laboratoryjnie dla skał, przy czym A równe jest $2E$ dla przewodnictwa jonowego (E opisuje energię aktywacji) i E dla przewodnictwa domieszkowego i elektronowego. σ i E zwykle zależą od ciśnienia i temperatury. Najsilniejszy wzrost przewodności z temperaturą następuje przy $T = E/2k$. W temperaturze dużo niższej przewodność jest niewielka w porównaniu z wartością σ_0, a w temperaturach wyższych od $T = E/k$ asymptotycznie zbliża się do maksymalnie możliwej wartości, czyli σ_0.

przewodność oliwinu

Przewodność oliwinu wzrasta od wartości niższej od 10^{-5} $(\Omega\ m)^{-1}$ w temperaturze pokojowej do 10^{-2} w 1000°C. Przewodnictwo domieszkowe ($\sigma_0 = 10^{-4}$–$10^{-2}\,\Omega^{-1}\,m^{-1}$) dominuje w zakresie temperatur do kilkuset °C. Powyżej około 700°C przewodnictwo elektronowe ($\sigma_0 \approx 5\,\Omega^{-1}\,m^{-1}$) wzrasta silnie i przewyższa przewodnictwo domieszkowe. Powyżej ok. 1100°C przewodnictwo jonowe ($\sigma_0 \approx 10^{-4}\,\Omega^{-1}m^{-1}$) zaczyna silnie dominować.

Przewodnictwo domieszkowe jest spowodowane obecnością obcych atomów w sieci kryształu, co powoduje nadmiar elektronów lub dziur, które poruszają się w sieci przy przyłożeniu siły elektromotorycznej. Ten typ przewodnictwa przypuszczalnie występuje w suchych skałach skorupy i najwyższych partii płaszcza.

Przewodnictwo jonowe jest silnie hamowane przez ciśnienie. Przewodnictwo to nie gra prawdopodobnie żadnej roli poniżej głębokości 500 km, gdyż prawdopodobnie temperatura jest tam zbyt niska na to, by przewodnictwo to mogło przewyższyć przewodnictwo elektronowe, wywołane ciśnieniem panującym na tej głębokości. Jednakże ten typ przewodnictwa może tłumaczyć względnie wysoką przewodność warstwy zaczynającej się na głębokości ok. 75 km. Warstwa ta jest przypuszczalnie identyczna z kanałem niskich prędkości, przy czym oba efekty spowodowane są bliskością punktu topnienia w tym obszarze. Przewodnictwo jonowe jest przypuszczalnie dominujące w magmie, a także w obszarach o temperaturach bliskich temperaturze topnienia, co może również tłumaczyć pewne lokalne anomalie przewodności w górnych obszarach płaszcza.

Przewodnictwo elektronowe jest przypuszczalnie dominującym procesem przewodzenia w strefie przejściowej i w dolnym płaszczu.

Trzęsienia ziemi

W ciągu jednego roku nawiedza Ziemię nie mniej niż milion wstrząsów sejsmicznych, z czego ok. 100 000 rejestrują sejsmografy. Na szczęście są to w większości trzęsienia bardzo słabe, wykrywalne jedynie przez sejsmografy o powiększeniach przekraczających 100 000 razy. Jedynie kilkadziesiąt trzęsień rocznie powoduje widoczne zmiany na powierzchni Ziemi (il. 203 i 204, tabl. 54; il. 205 i 206, tabl. 55), a ujmując rzecz staty stycznie — tylko jedno z nich ma przebieg katastrofalny (zob. str. 822).

Mieszkamy na terenie, który należy do obszarów asejsmicznych. W części Karpat leżącej w granicach Polski obserwuje się wprawdzie bardzo słabą aktywność sejsmiczną, ale duże trzęsienia ziemi nam nie zagrażają. W różnych częściach kuli ziemskiej istnieją jednak obszary o silnej sejsmiczności. Tam trzęsienia ziemi powodują śmierć setek tysięcy ludzi i ogromne zniszczenia na powierzchni. Obok wybuchów wulkanów, powodzi, huraganów i tornado zjawiska te należą do najważniejszych klęsk żywiołowych z uwagi na częstość występowania i szerokie rozprzestrzenienie. Ocenia się, że w wyniku trzęsień zginęło w obecnym tysiącleciu co najmniej 5 milionów ludzi; niektóre źródła podają liczby dwukrotnie, a nawet trzykrotnie wyższe. Badanie procesów fizycznych zachodzących w obszarze ogniska trzęsienia ziemi, a także rejestracja i analiza związanych z tymi procesami anomalii pewnych parametrów fizycznych (prędkości fal sejsmicznych, namagnesowania i przewodności elektrycznej skał, koncentracji radonu w wodach gruntowych, ilości i gęstości występowania słabych trzęsień ziemi itp.), występujących w obszarze przyszłego trzęsienia ziemi, stanowi jeden z podstawowych celów sejsmologii i jest pierwszym krokiem ku prawidłowym prognozom katastrofalnych trzęsień ziemi.

Procesy fizyczne w ognisku

Trzęsieniem ziemi nazywamy proces nagłego wyzwolenia energii w stosunkowo niewielkim obszarze we wnętrzu Ziemi zwanym ogniskiem trzęsienia. Obszar położony bezpośrednio nad ogniskiem jest nazywany obszarem epicentralnym i w nim najsilniej dają się odczuć skutki trzęsienia.

obszar epicentralny

Trzęsienia ziemi klasyfikuje się w zależności od głębokości ogniska w sposób następujący: 0–70 km — trzęsienia płytkie, 70–300 km — trzęsienia średniogłębokie, 300–720 km — trzęsienia głębokie. Na głębokościach większych od 720 km nie wykryto żadnych trzęsień ziemi. Ponad 75% energii wyzwala się w trzęsieniach płytkich i jedynie około 3% — w trzęsieniach głębokich. Średnio rocznie w trzęsieniach ziemi wyzwala się ok. 10^{18} J, przy czym większość tej energii w czasie jednego lub kilku największych trzęsień. Wyzwalana energia tylko częściowo przekształca się w energię fal sejsmicznych; duża jej część przekształca się bezpośrednio w ciepło na skutek tarcia w płaszczyźnie rozrywu w ognisku trzęsienia.

trzęsienia płytkie, średniogłębokie i głębokie

Trzęsienie ziemi trwa zwykle krótko — od kilku sekund do kilkunastu minut, często jednak po głównym silnym wstrząsie następują wstrząsy następcze (repliki), nieraz przez kilka dni lub nawet kilka miesięcy, o sile zwykle malejącej, czasem jednak prawie równej sile trzęsienia początkowego.

Trzęsienia ziemi należą do klasy zjawisk związanych z przekroczeniem pewnej granicy krytycznej (w tym

wypadku granicy wytrzymałości materiału skalnego we wnętrzu Ziemi) i mogą być wywoływane przez bardzo wiele różnorodnych czynników, takich jak zmiana ciśnienia atmosferycznego, naprężenia pływowe, pobliskie trzęsienia ziemi, niemożliwe do przewidzenia zjawiska pełzania materiału w górnym płaszczu lub skorupie ziemskiej, czy też chwilowy wzrost naprężeń tektonicznych.

Energia wyzwalająca się w sposób gwałtowny w czasie trzęsienia została uprzednio nagromadzona w procesie powolnej akumulacji naprężeń tektonicznych we wnętrzu Ziemi. Pierwotnym źródłem tej energii, a także źródłem energii wielu innych procesów geodynamicznych, jest energia cieplna wnętrza Ziemi.

trzęsienia typu tektonicznego

Większość trzęsień ziemi to trzęsienia typu tektonicznego, polegające na niestacjonarnym rozwoju jednej lub kilku szczelin; w wypadku trzęsień powierzchniowych szczeliny te wychodzą na powierzchnię i są obserwowane w postaci głębokich uskoków przesuwczych na powierzchni Ziemi. Przypuszczalnie możliwy jest również inny mechanizm powstania ogniska, sugerowany głównie dla trzęsień głębokich, polegający na szybkim przejściu fazowym materiału skalnego w rejonie ogniska ze stanu metastabilnego, w jakim znalazł się na skutek powolnych ruchów mas skalnych we wnętrzu Ziemi, do stanu stabilnego, odpowiadającego aktualnie panującym w tym rejonie warunkom temperatury i ciśnienia.

Fale sejsmiczne

ognisko

płaszczyzna ogniskowa

W ognisku materiał skalny ulega nagłemu przesunięciu i zniszczeniu wzdłuż pewnej płaszczyzny, zwanej płaszczyzną ogniskową, a sam proces w ognisku jest przyczyną powstania fal sejsmicznych. W czasie trzęsień tektonicznych fale emitowane są w trakcie niestacjonarnego rozprzestrzeniania się szczeliny.

Ze względu na panujące we wnętrzu Ziemi ciśnienie już poniżej górnych 20–30 km szczeliny te są wyłącznie szczelinami ścinania; w obrębie skorupy ziemskiej możliwe bywają również szczeliny otwarte, obserwowane często w rejonach epicentralnych silnych powierzchniowych trzęsień ziemi.

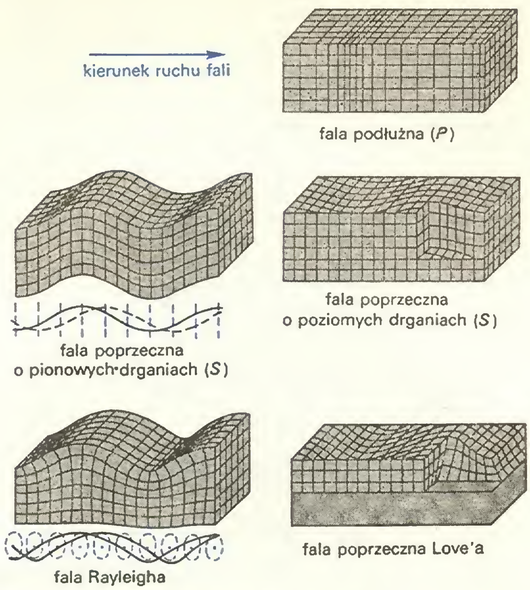
W ognisku generowane są różnego rodzaju fale, głównie fale objętościowe: fale podłużne (zgęszczeń i rozrzedzeń), zwane falami *P* (od łac. *primae* — 'pierwsze', gdyż ich prędkość jest większa niż innych rodzajów fal i dochodzą one najwcześniej do miejsca obserwacji) i fale poprzeczne (fale ścinania), zwane falami *S* (od łac. *secundae* — 'drugie'). Po dojściu do powierzchni swobodnej Ziemi fale objętościowe mogą generować fale powierzchniowe Love'a lub fale powierzchniowe Rayleigha, w zależności od budowy warstwy przypowierzchniowej (rys. 6).

układ sił w ognisku trzęsienia

Jak wynika z analizy pola falowego generowanego w trakcie trzęsienia, ruch mas skalnych po przeciwnych stronach płaszczyzny ogniskowej może być opisany prawidłowo jako wynik działania pary sił (rys. 7a) lub podwójnej pary sił bez momentu (rys. 7b). Przestrzeń wokół ogniska trzęsienia można podzielić na cztery jednakowe obszary, w których na przemian pierwszy ruch fali *P* (znak fazy) odpowiada zgęszczeniu lub rozrzedzeniu. Jedna z płaszczyzn rozdzielających te obszary to właśnie płaszczyzna ogniska; druga z nich jest prostopadła do pierwszej, a obie tworzą układ tzw. płaszczyzn nodalnych. Pole falowe fal *P* emitowanych w ognisku trzęsienia jest identyczne dla obu modeli mechanizmów trzęsień, tzn. pary sił (rys. 7c) i podwójnej pary sił bez momentu (rys. 7d), natomiast pole falowe fal *S* jest charakterystycznie różne dla obu modeli (rys. 7e, f). Tak więc analiza pierwszych ruchów w fali *P* pozwala wyłącznie na określenie położenia płaszczyzn nodalnych, z których jedna odpowiada płaszczyźnie ogniska, ale nie wiadomo która, dopiero równoczesna analiza pierwszych ruchów fali *S* umożliwia wydzielenie z płaszczyzn nodalnych płaszczyzny ogniskowej.

płaszczyzny modalne

W trzęsieniach ziemi wywołanych szybką przemianą fazową znak pierwszej fazy fali *P* jest jednakowy w całej przestrzeni, a więc również i na powierzchni Ziemi,

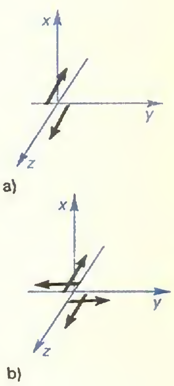


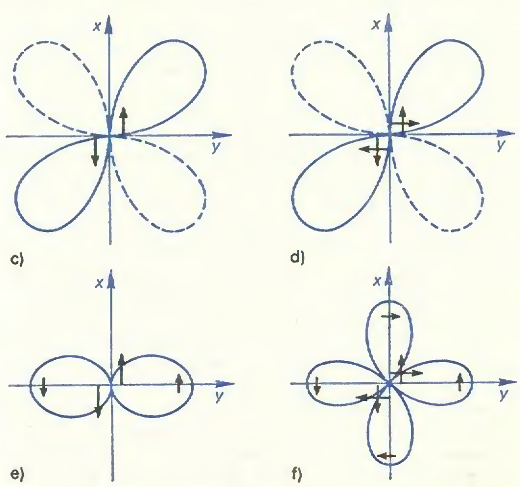
Rys. 6. Podstawowe rodzaje fal sejsmicznych

przy czym znak ten jest określony kierunkiem przejścia fazowego.

Badanie położeń płaszczyzn ogniskowych trzęsień ziemi stanowi źródło cennych informacji o rozkładzie naprężeń działających we wnętrzu Ziemi i o współczesnych ruchach geodynamicznych.

W ośrodku sprężystym mającym powierzchnię swobodną mogą rozprzestrzeniać się — oprócz fal objęto-





Rys. 7. Układ sił w ognisku trzęsienia: a) pojedyncza para sił (układ sił typu I), b) podwójna para sił (układ sił typu II), c) i d) pole fal *P* generowane przez pojedynczą parę sił (c) i podwójną parę sił (d), e) i f) pole fal *S* generowane przez pojedynczą parę sił (*e*) i podwójną parę sił (*f*)

fale powierzchniowe

ściowych — także fale powierzchniowe. Są to fale długie o znacznych amplitudach i one to powodują najsilniejszy wstrząs podczas trzęsienia. Amplitudy tych fal maleją odwrotnie proporcjonalnie do pierwiastka z odległości od epicentrum, podczas gdy amplitudy fal objętościowych maleją odwrotnie proporcjonalnie do samej odległości, tak więc na dużych odległościach fale powierzchniowe zapisywane są wyraźniej, niż fale objętościowe. Amplitudy fal powierzchniowych ma-

leją z głębokością według prawa wykładniczego. Fale powierzchniowe wykazują silną dyspersję (prędkości ich silnie zależą od długości fali). Analiza długookresowych fal powierzchniowych jest jedną z najcenniejszych metod badania wewnętrznej struktury Ziemi, a przede wszystkim budowy skorupy i górnych warstw płaszcza.

Dwa najprostsze typy fal powierzchniowych to fale Rayleigha i fale Love'a (→ Akustyczne fale powierzchniowe i ich zastosowanie).

fale Rayleigha

Fale Rayleigha są jedynym rodzajem fal powierzchniowych możliwych w jednorodnej przestrzeni sprężystej. Cząstki ośrodka zataczają elipsy w płaszczyźnie pionowej równoległej do kierunku rozchodzenia się fal (rys. 6). Amplituda fali maleje wykładniczo wraz ze wzrostem odległości od powierzchni swobodnej. W ośrodku o stałej Poissona równej 0,25 prędkość fal Rayleigha v_R wynosi 0,92 v_S, gdzie v_S jest prędkością fali *S*.

Fale Rayleigha ulegają dyspersji, jeżeli własności mechaniczne oraz gęstość ośrodka zmieniają się wraz z odległością od powierzchni swobodnej. Fale o długości λ są czułe na własności sprężyste i gęstość ośrodka do głębokości $\lambda/5$, tak więc prędkość fal Rayleigha na powierzchni Ziemi zależy od długości fali, przy czym im większa jest długość fali, tym głębsze warstwy wpływają na jej prędkość.

Każdą falę można rozłożyć na fale płaskie jednorodne i fale niejednorodne, z nich tylko fale niejednorodne powodują powstawanie fal powierzchniowych Rayleigha. Płaskie jednorodne fale padając na powierzchnię swobodną nie dają fal powierzchniowych i odwrotnie, im większa jest krzywizna frontu fali padającej, tym silniejsze są fale powierzchniowe. Ponieważ przy ogniskach głębokich lub przy dużych odległościach od epicentrum trzęsienia front fali jest prawie płaski, to powstające fale Rayleigha są bardzo słabe, natomiast silne fale Rayleigha są obserwowane w wypadku ognisk niegłębokich i w rejonach blisko epicentrum.

fale Love'a

Najprostszym typem struktury, w której mogą rozchodzić się fale Love'a, jest jednorodna warstwa z jedną powierzchnią swobodną, a drugą leżącą na jednorodnej półprzestrzeni sprężystej, przy czym prędkość fali *S* w warstwie jest mniejsza niż w półprzestrzeni. Drgania odbywają się w płaszczyźnie poziomej, prostopadłej do kierunku rozchodzenia się fal. W obrębie półprzestrzeni amplituda fali maleje wykładniczo z odległością od granicy oddzielającej półprzestrzeń od warstwy. Fale Love'a ulegają dyspersji, przy czym ich prędkość fazowa zmienia się od wartości prędkości fali *S* w warstwie (dla bardzo krótkich fal Love'a) do prędkości fal *S* w półprzestrzeni (dla bardzo długich fal Love'a). Fale Love'a istnieją również w ośrodkach o bardziej skomplikowanej budowie, tzn. w ośrodkach o większej liczbie warstw, jeżeli tylko prędkość fal *S* rośnie wraz z odległością od powierzchni swobodnej, przy czym krzywe dyspersji tych fal odzwierciedlają strukturę ośrodka.

Skomplikowana struktura wnętrza Ziemi jest przyczyną tego, że rejestrowane pole falowe jest bardzo złożone i oprócz fal objętościowych i powierzchniowych obserwowane są również i inne typy fal. Ostatnio np. (1974 r.) wykryto fale charakterystyczne dla ośrodków o złożonej strukturze wewnętrznej, jak ośrodki mikromorficzne lub mikropolarne. Fale te rozchodzą się w rejonach bliskich epicentrum niektórych trzęsień, w obszarach o silnie zakłóconej tektonice i wyraźnej strukturze blokowej. Fale te charakteryzuje m.in. rotacja elementów ośrodka przy przejściu fali i one to przypuszczalnie wywołują dobrze znane efekty skręcenia kolumn, cokołów pomników i tym podobnych elementów architektonicznych w obszarach epicentralnych niektórych silnych trzęsień ziemi.

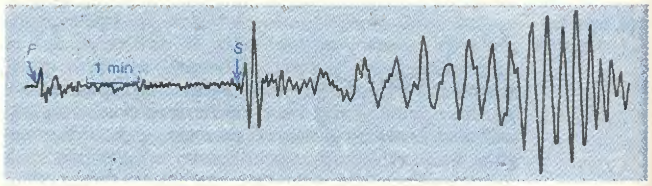
fale sejsmiczne w ośrodkach o złożonej strukturze

Analiza pola falowego rejestrowanego na powierzchni Ziemi jest głównym źródłem informacji o strukturze wnętrza oraz współczesnych ruchach geodynamicznych zachodzących we wnętrzu naszej planety.

Drgania, które docierają do powierzchni Ziemi w wyniku trzęsienia, składają się z ruchów o różnych kierunkach, z różnymi amplitudami i różnymi okresami. Bezpośrednie rejestrowanie takich przestrzennych drgań jest niemożliwe i dlatego rozkłada się je na 3 wzajemnie prostopadłe składowe i notuje oddzielenie. Całość zapisu umożliwia w pełni odtworzenie drgań przestrzennych cząstek podłoża.

Zapis trzęsienia ziemi

Zapis trzęsienia ziemi, zwany sejsmogramem, składa się zazwyczaj z trzech zasadniczych części: fazy wstępnej (fale objętościowe), fazy głównej (fale powierzchniowe) i „ogona" (rys. 8).



Rys. 8. Sejsmogram trzęsienia ziemi w Turcji (1909 r.) zarejestrowany w stacji Pułkowo w Rosji (wg Galicyna)

Bezpośrednio biegnące fale *P* docierają jedynie do odległości epicentralnej $\Delta = 103°$ (ok. 11 600 km); w odległościach większych (do 130°) dają się zauważyć jedynie słabe ślady fal ugiętych na powierzchni jądra. Dopiero w odległości 142–145° pojawiają się znowu bardzo wyraźne fale podłużne, jakkolwiek znacznie opóźnione (tzw. fale *PKP*). Strefa, do której fale sejsmiczne nie docierają wskutek odchylenia przez jądro ziemskie, nazywa się „strefą cienia".

faza wstępna

Gdy odległości epicentralne są małe faza wstępna trwa krótko: ok. 1 min przy $\Delta = 500$ km i ok. 2 min przy $\Delta = 2000$ km, po czym pojawiają się silniejsze impulsy pochodzące od fal powierzchniowych. Wraz ze wzrostem odległości epicentralnej wzrasta czas trwania fazy wstępnej. Maksymalna długość fazy wstępnej dochodzi do niespełna 1 h przy $\Delta = 20\,000$ km.

faza główna

W fazie głównej najdłuższymi okresami drgań odznaczają się zawsze fale Love'a. Gdy odległości epicentralne są średnie, ich okresy wynoszą 12–20 s, gdy duże — przekraczają nawet 1 min. Przy trzęsieniach średniogłębokich i najgłębszych (głębokość ogniska 700 km) faza główna jest bardzo słabo zaznaczona, bywa też czasem zupełnie niedostrzegalna.

Gdy odległości epicentralne są małe, nie ma zasadniczych różnic w zewnętrznym wyglądzie sejsmogramu trzęsienia płytkiego i głębokiego, prócz tego, że grupy fal *P* i *S* dla trzęsień głębokich są zwykle wyraźniejsze i mają krótszy okres.

W polu falowym generowanym przez słabe trzęsienia ziemi dominują fale o krótkich okresach rzędu 1 s lub mniejszym. Wielkie trzęsienia ziemi generują fale zarówno krótkie — jak i długookresowe, przy czym fale *P* mogą mieć okresy nawet rzędu 25 s, a niektóre rodzaje fal powierzchniowych okresy osiągające wielkość trzech minut.

Fale krótkookresowe są tłumione silniej we wnętrzu Ziemi niż długookresowe.

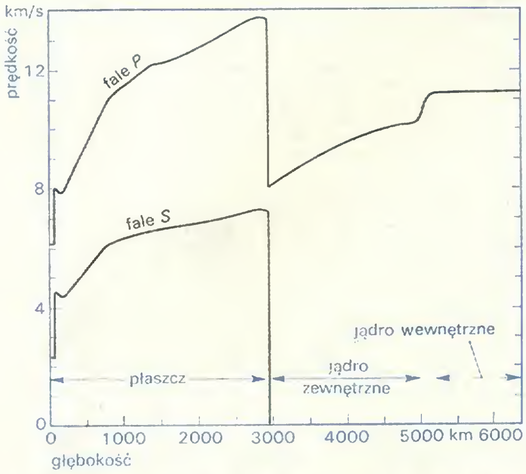
Badanie wewnętrznej budowy Ziemi

Kształt zapisu pola falowego, generowanego w ognisku trzęsienia ziemi, określony jest przez trzy podstawowe czynniki, a mianowicie: proces w ognisku trzęsienia (prędkość rozrywu w ognisku, tarcie w płaszczyźnie rozrywu, wielkość naprężeń tektonicznych wyzwolonych w czasie trzęsienia, wielkość i kształt płaszczyzny rozrywu itp.), własności reologiczne i strukturę ośrodka, przez który przeszły fale na drodze ognisko–rejestrator, a także charakterystykę samego urządzenia rejestrującego. Przyjęcie pewnych

założeń modelowych dotyczących procesu w ognisku trzęsienia i znajomość charakterystyk amplitudowych i częstościowych sejsmografu pozwala na wyodrębnienie informacji o wewnętrznej budowie Ziemi.

prędkość fal sejsmicznych

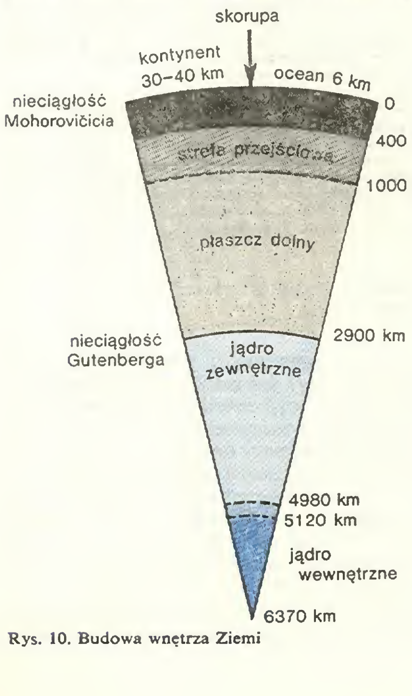
Prędkość fal sejsmicznych jest funkcją własności mechanicznych ośrodka, tzn. gęstości, ściśliwości i sztywności skał we wnętrzu Ziemi. Znajomość prędkości fal sejsmicznych (rys. 9) umożliwia określenie zmian tych parametrów wraz ze zmianą głębokości oraz pozwala na stawianie hipotez dotyczących stanu wnętrza Ziemi.



Rys. 9. Rozkład prędkości fal sejsmicznych P i S we wnętrzu Ziemi

fale P i S

Pierwszym ważnym krokiem w tej dziedzinie było wyróżnienie przez R. Oldhama w 1897 r. dwóch rodzajów objętościowych fal sejsmicznych (fal P i S). Analiza zapisów trzęsień ziemi ujawniła istnienie ciekłego jądra Ziemi, obszaru o mniejszej prędkości fali P, otoczonego grubym, półplastycznym płaszczem. Granicę pomiędzy jądrem a płaszczem charakteryzuje skokowa zmiana prędkości fal sejsmicznych (nieciągłość Gutenberga, rys. 10). Nieciągłość ta może być wywołana zarówno różnym składem chemicznym poszczególnych warstw jak i zmianą struktury fazowej tego samego materiału skalnego. Skokowa zmiana prędkości występuje również pomiędzy skorupą, a płaszczem Ziemi (nieciągłość Mohorovičicia), w którym prędkości fal P i S są znacznie większe, niż w skorupie.

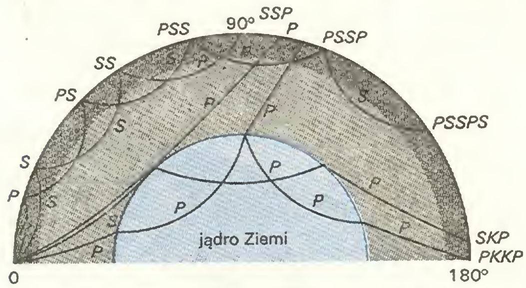


Rys. 10. Budowa wnętrza Ziemi

Analiza zapisów sejsmicznych umożliwia również wykrycie stałego jądra wewnętrznego.

Ostatnim odkryciem dotyczącym rozpoznania zasadniczych elementów struktury Ziemi było wykrycie istnienia silnie plastycznej warstwy w górnym płaszczu (warstwa ta ma różną grubość dla fal P i S) o mniejszych prędkościach fal sejsmicznych — astenosfery. Analizując zapisy trzęsień zarejestrowane w różnych punktach powierzchni Ziemi można ustalić tory fal sejsmicznych w jej wnętrzu (rys. 11), a tym samym uzyskać pewne informacje o gęstości i własnościach mechanicznych ośrodka, przez który przechodzą.

tory fal sejsmicznych



Rys. 11. Tory niektórych typów fal sejsmicznych we wnętrzu Ziemi, gdzie P — oznacza fale podłużne idące przez wnętrze Ziemi bezpośrednio z ogniska; S — analogiczne fale poprzeczne; PS — fale odbite od powierzchni Ziemi, biegnące początkowo jako podłużne, a po odbiciu jako poprzeczne, SP — fale odbite od powierzchni Ziemi, biegnące początkowo jako poprzeczne, a po odbiciu jako podłużne, PP — raz odbite od powierzchni Ziemi fale podłużne, SS — raz odbite od powierzchni Ziemi fale poprzeczne, PPP, SSS, PSP itd. — dwukrotnie odbite fale; ich charakter jest podany przez odpowiednie litery, PKP — fala podłużna przechodząca przez jądro Ziemi bez odbicia wewnątrz jądra, ale załamująca się na jego granicy (litera K — od niemieckiego *kernwellen* oznacza fale, które przeszły przez jądro), PKS — analogiczna fala podłużna wychodząca z jądra jako fala poprzeczna, $PKKP$ — fala podłużna przechodząca przez jądro z jednym odbiciem wewnątrz jądra (przez jądro przechodzą fale zawsze jako P, gdyż jego część jest w stanie ciekłym)

Cennym źródłem wiadomości o budowie skorupy i głównych warstw płaszcza jest analiza krzywych dyspersji fal powierzchniowych.

Współczesna sejsmologia coraz dokładniej rejestruje szczegóły wewnętrznej budowy Ziemi i jej warstw, ukazuje różnice w strukturze wgłębnej między oceanem a kontynentem, a także między różnymi regionami Ziemi. I tak np. grubość skorupy waha się od kilkunastu kilometrów w rejonie oceanów do kilkudziesięciu kilometrów w niektórych rejonach kontynentalnych, przy czym różnice w strukturze dotyczą nie tylko skorupy Ziemi, lecz sięgają również głęboko w obszary górnego płaszcza.

sejsmograf strunowy

W przeciwieństwie do sejsmografów dawniejszych, o wielkich masach, koniecznych ze względu na rejestrację mechaniczną, ostatnio buduje się sejsmografy o masach niewielkich. Jako przykład może służyć strunowy sejsmograf Wooda i Andersona, w którym masa wynosi zaledwie 0,7 g. Jest ona przymocowana ekscentrycznie do struny wolframowej, która działa w tym wypadku jak oś obrotu. Sejsmograf ten ma rejestrację optyczną i powiększa 400–2000 razy.

skala magnitud

Najstarszą metodą określania wielkości trzęsienia była skala intensywności oparta na efektach powierzchniowych wstrząsu. Niestety efekty te nie są proporcjonalne do energii trzęsienia, gdyż silnie zależą od głębokości ogniska oraz od lokalnych warunków tektonicznych, tak więc skala ta nie nadaje się do prawidłowej klasyfikacji trzęsień ziemi. Obecnie w celu sklasyfikowania energii trzęsienia stosuje się skalę instrumentalną, tzw. skalę magnitud. Wielkość zwana magnitudą określana jest na podstawie amplitudy maksymalnych ruchów cząsteczek gruntu i jest związana z energią trzęsienia. Dla trzęsień głębokich, generujących niewiele fal powierzchniowych, wprowadzono inną skalę magnitud, związanych z maksymalną am-

plitudą fal objętościowych. Obie skale są ze sobą związane tak, że możliwe jest porównywanie trzęsień płytkich i głębokich.

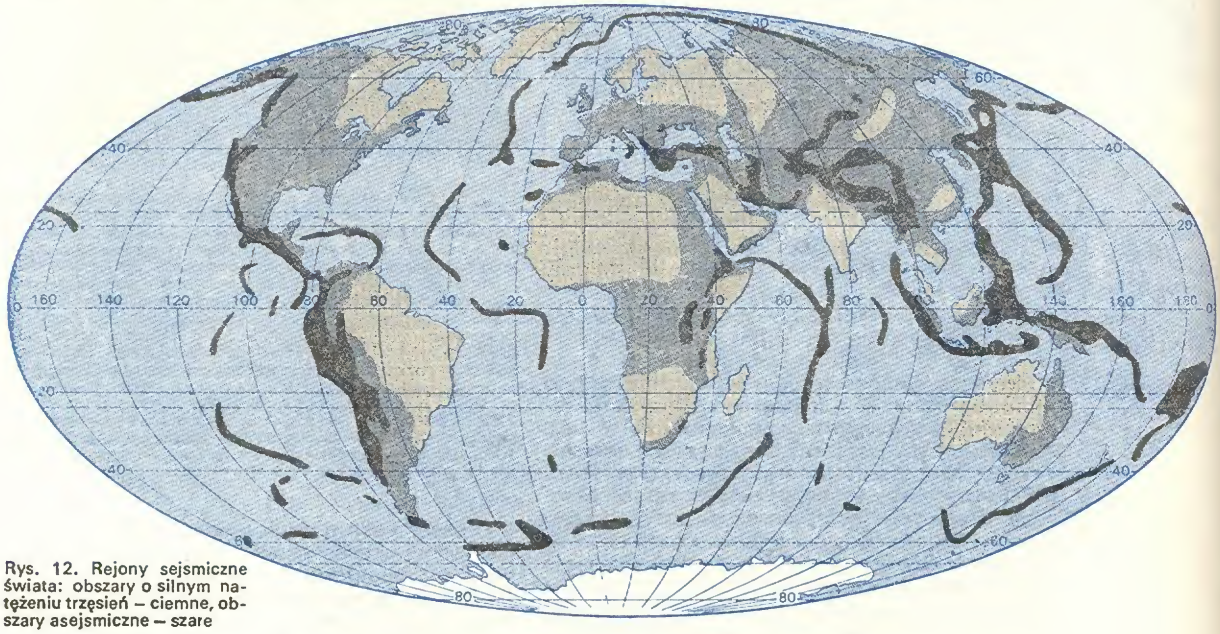
Energia najsilniejszych obserwowanych trzęsień ziemi jest szacowana na 10^{18} dżuli.

Sejsmiczność Ziemi

Mapy rozmieszczenia obszarów epicentralnych na całym świecie wykazują istnienie dwu głównych stref sejmicznych: okołopacyficznej oraz śródziemnomorskiej (rys. 12).

strefy sejsmiczne

Strefa okołopacyficzna (tzw. ando-japońsko-malajska) biegnie dookoła Oceanu Spokojnego. Do strefy tej włączone są nie tylko wybrzeża lądu stałego i wyspy, ale także znaczne obszary morskie, w których silnym trzęsieniom ulega dno, zwłaszcza w obrębie głębokich rowów oceanicznych. Strefa ta jest siedliskiem 80% trzęsień, w tej liczbie wielu trzęsień katastrofalnych, 40% płytkich ognisk, 90% średniogłębokich i wszystkich głębokich.



Rys. 12. Rejony sejsmiczne świata: obszary o silnym natężeniu trzęsień – ciemne, obszary asejsmiczne – szare

Strefa śródziemnomorska zorientowana jest prawie równoleżnikowo; zawiera ona ogniska średniej głębokości.

obszary asejsmiczne

Przeciwieństwem stref sejsmicznych są obszary asejsmiczne, pokrywające wielkie tarcze lądowe spoczywające na starych cokołach kontynentalnych, w których ustały już ruchy górotwórcze. Do obszarów asejsmicznych należą: basen Oceanu Spokojnego (z wyjątkiem Wysp Hawajskich), płyta kanadyjska, płyta brazylijska, płyta eurazjatycka, Afryka (z wyjątkiem wybrzeży Morza Śródziemnego i górnego dorzecza Nilu), Antarktyda i inne drobniejsze połacie kuli ziemskiej.

trzęsienia katastrofalne

W strefach sejsmicznych co jakiś czas zdarzają się trzęsienia ziemi o przebiegu katastrofalnym. Jedną z najcięższych katastrof sejsmicznych w historii było trzęsienie ziemi lizbońskie w 1755 r. Obszar dotknięty zniszczeniami był niezwykle rozległy i objął południowo-zachodnią Europę i północno-wschodnią Afrykę. Zginęło ok. 60 000 mieszkańców Lizbony, miasta, które liczyło wówczas 235 000 mieszkańców. Zniszczeniu uległo 3/4 miasta. Na wodach stawów i jezior, nawet znajdujących się w znacznej odległosci od epicentum, trzęsienie wzbudziło sejsze (długotrwałe oscylacje poziomu wody o częstościach równych częstościom własnym danego zbiornika); oscylacje jeziora w Loch Lomond w Szkocji trwały półtorej godziny. Rzeka Trave w Lubece podniosła nagle swój poziom o 1,5 m. Obszar, na którym wstrząsy były wyraźnie odczuwalne, oceniany jest na co najmniej $1{,}5 \cdot 10^6$ km².

Katastrofalne trzęsienia ziemi, ze względu na ogrom strat i zniszczeń, silnie pobudzały rozwój nauki o trzęsieniach ziemi. Pierwszym naukowo dobrze udokomentowanym silnym trzęsieniem ziemi było trzęsienie assamskie w 1897 r. o epicentrum leżącym w prowincji Assam, między Himalajami a Zatoką Bengalską. Trzęsienie trwało niecałe 3 minuty, a obszar całkowitego zniszczenia objął 100 tys. km². Szereg domów zbudowanych na bardzo miękkim gruncie pogrążyło się w nim aż do poziomu dachów. W pobliżu epicentrum powstały liczne uskoki, największy z nich miał długość 20 km, przy czym jedna jego strona górowała o 10 m nad drugą. Wstrząsy następcze w tym rejonie trwały ponad półtora roku.

Wielkie trzęsienie ziemi, które nawiedziło Kalifornię 18 kwietnia 1906 r. (il. 205 a, tabl. 55) zasługuje na szczególną uwagę ze względu na wyjątkowe rozmiary ruchów tektonicznych, które były jego przyczyną. Widoczne na powierzchnj przemieszczenia poziome wzdłuż znanego geologom uskoku tektonicznego San Andreas (il. 204, tabl. 54) ciągnęły się na przestrzeni około 430 km, największe z nich wynosiły 6,5 m. Obszar najsilniej zniszczony miał kształt wąskiego pasa ciągnącego się wzdłuż uskoku. Trzęsienie trwało jedynie ok. 40 s, a wstrząsy następcze trwały ok. półtora roku; niektóre z nich osiągnęły duże natężenie. Bardzo silnym było również trzęsienie w 1923 r., którego ognisko leżało pod dnem morskim zatoki Sagami, w odległości około 90 km od Tokio i 65 km od Jokohamy. Jak to miało miejsce podczas trzęsienie w San Francisco, rozmiary katastrofy zostały kilkakrotnie powiększone przez pożary. W samej Jokohamie zginęło 27 tys. osób, 40 tys. osób zostało rannych, a 70 tys. domów uległo całkowitemu zniszczeniu. Jedenastometrowa fala morska (tzw. tsunami) spowodowana trzęsieniem wyrządziła wiele szkód na wybrzeżu. Trzęsienie to stanowiło jedną z najcięższych katastrof sejsmicznych w dziejach ludzkości, zginęło 99 tys. osób. Równocześnie było ono pierwszym z wielkich trzęsień dokładnie zarejestrowanych przez nowoczesne sejsmografy we wszystkich obserwatoriach świata.

Największe straty ponosi się, gdy epicentum znajduje się w pobliżu dużych skupisk ludzkich. W 1960 r. na skutek trzęsienia ziemi prawie całkowicie został zniszczony Agadir i zginęła 1/3 mieszkańców miasta (11 tys. osób). Trzęsienie to było 6 tys. razy słabsze, niż trzęsienie z San Francisco, ale epicentrum trzęsienia leżało w odległości paru zaledwie kilometrów od granic miasta. Ostatnio takie katastrofalne trzęsienie ziemi miało miejsce we Włoszech 10 stycznia 1981 r., epicentrum znajdowało się w pobliżu miejscowości Avellino. Zginęło tam ok. 3 tys. osób.

Prognozowanie trzęsień ziemi, sztuczne trzęsienia

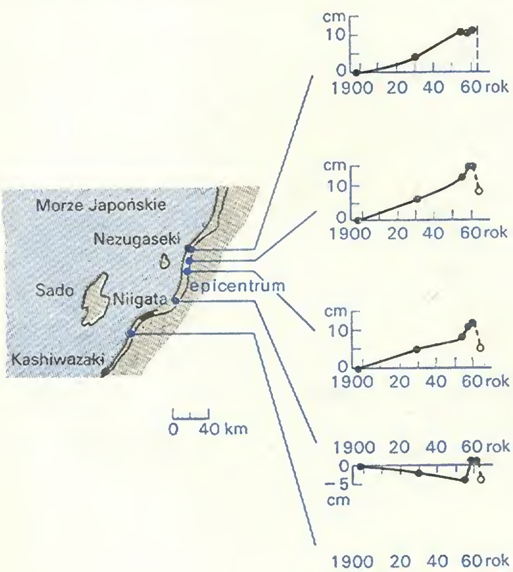
Niektóre zjawiska sejsmiczne w regionach o prostej strukturze tektonicznej i dokładnie znanym mechanizmie trzęsień mogą być całkowicie prawidłowo przewidziane, podczas gdy w złożonej sytuacji tektonicznej możliwe są jedynie prognozy oparte na probabilistyce i dokładnej znajomości historii sejsmicznej danego regionu. Ten ostatni typ prognoz oparty jest na dokładnej rejestracji miejsc występowania i siły poszczególnych trzęsień ziemi, co w zestawieniu z tektoniką danego obszaru pozwala na wykreślenie map sejsmiczności, z uwzględnieniem rejonów niebezpiecznych sejsmicznie. Zakłada się tu stałość reżimu sejsmicznego w obrębie danej jednostki tektonicznej, tzn. zakła-

da się, że jeżeli w danym regionie miało miejsce silne trzęsienie ziemi, to może się ono powtórzyć z tą samą siłą w obrębie tej samej jednostki tektonicznej. Ze względu na występowanie tzw. cykli sejsmicznych mapy sejsmiczności uwzględniają również powtarzalność silnych trzęsień.

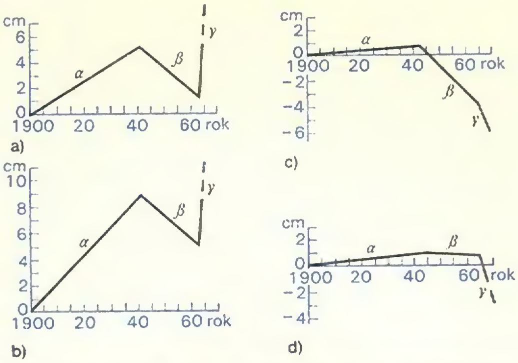
Mapy sejsmiczne umożliwiają obecnie precyzyjną prognozę miejsca i siły przyszłego trzęsienia ziemi, natomiast jedyną nadzieją na prawidłowe prognozy momentu wystąpienia takiego zjawiska jest znalezienie zjawisk fizycznych, poprzedzających trzęsienie, czyli zjawisk towarzyszących akumulacji naprężeń w rejonie ogniska przyszłego trzęsienia ziemi. Zjawiska te są obecnie intensywnie badane; mniej systematyczne poszukiwania były prowadzone już od prawie stu lat, jednakże bez widocznych wyników. Zjawiska te są związane ze wzrostem naprężeń w pewnym rejonie wnętrza Ziemi. Należą do nich: zmiany prędkości fal *P* i *S*, spowodowane wzrostem ilości drobnych szczelin w obszarze przyszłego ogniska (przy czym prędkość fal *P* ulega zmianie nawet do 20%, a prędkość fal *S* — o kilka procent), powolna ciągła deformacja powierzchni Ziemi mająca miejsce przed niektórymi silnymi płytkimi trzęsieniami ziemi, zmiany przewodnictwa elektrycznego mas skalnych w rejonie przyszłego ogniska, zmiany niektórych parametrów ziemskiego pola magnetycznego, anomalne zmiany poziomu wody (rys. 13–15 prezentują rejestracje niektórych z tych efektów przed i w trakcie silnych trzęsień ziemi).

zjawiska poprzedzające trzęsienie

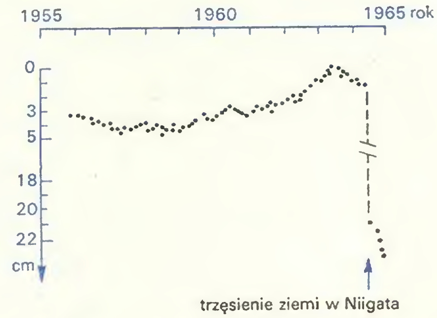
Współcześnie (Nur, 1972 r.) opracowano model wyjaśniający wiele z obserwowanych zjawisk poprzedzających silne płytkie trzęsienia ziemi. W modelu tym wzrost naprężeń prowadzi do obserwowalnego rozszerzenia (dylatacji) naprężonego regionu, co w następstwie powoduje przepływ wody zawartej w porach i szczelinach skał otaczających ten obszar w głąb rozszerzonego regionu. Ten przepływ wody powoduje wzrost ciśnienia cieczy w porach skał w naprężonym obszarze oraz nasycenie wodą porów i szczelin, co może być bezpośrednią przyczyną trzęsienia. Obecność dodatkowej ilości cieczy w naprężonym rejonie może wywołać zmianę przewodnictwa elektrycznego skał, a globalna deformacja mogłaby być przyczyną obserwowanych deformacji powierzchniowych, poprzedzających niektóre silne płytkie trzęsienia ziemi. Model ten tłumaczy również obserwowane zmiany poziomu wód gruntowych. Obecnie prowadzone systematyczne badania nad zjawiskami poprzedzającymi silne trzęsienia ziemi mają na celu m.in. udokumentowanie prawdziwości tego modelu lub też znalezienie innego, właściwego opisu obserwowanych procesów fizycznych.



Rys. 13. Anomalne zmiany wysokości n.p.m. zanotowane przed trzęsieniem Ziemi w Niigata w 1964 r.



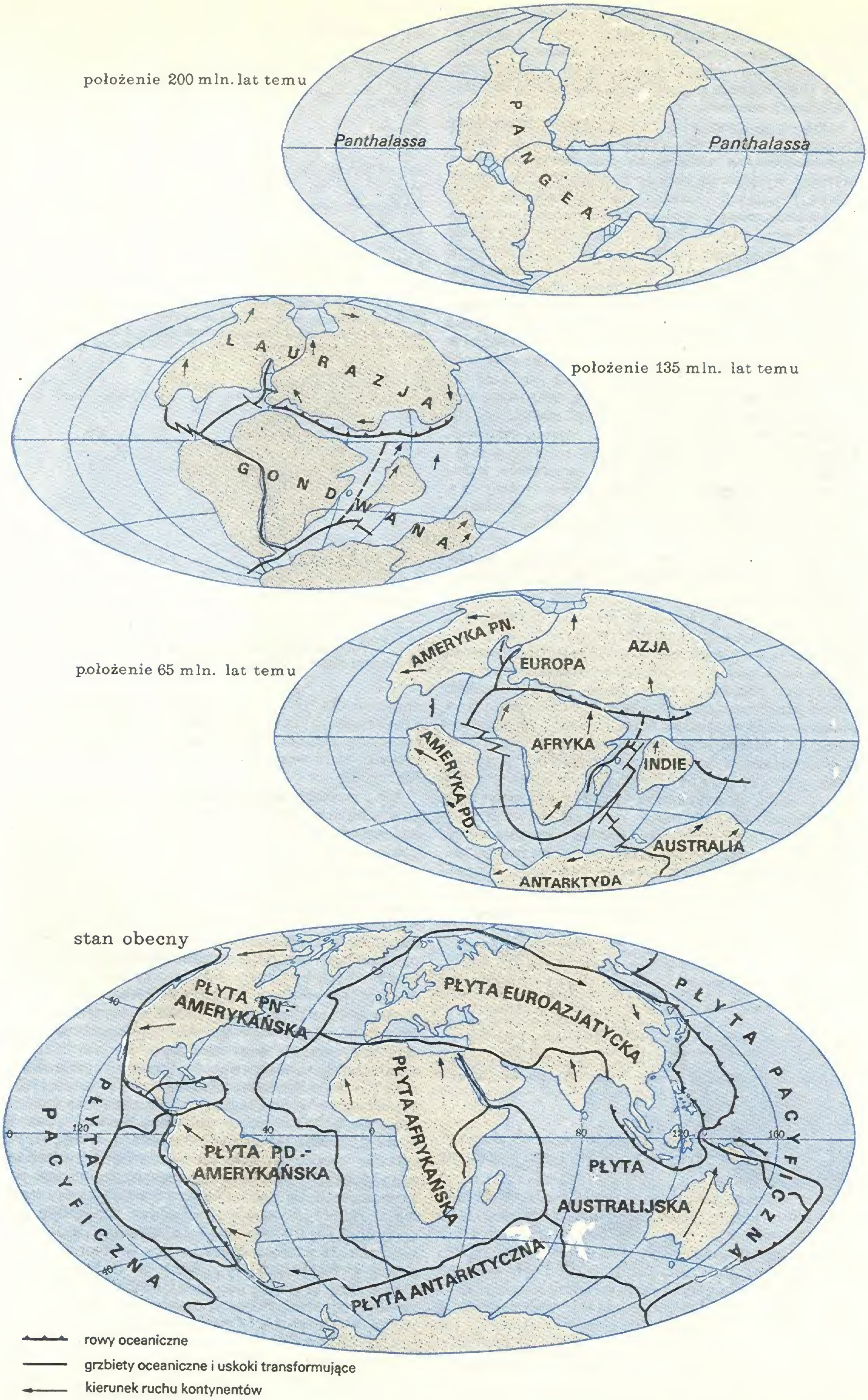
Rys. 14. Anomalne zmiany wysokości n.p.m. zarejestrowane przed taszkenckim trzęsieniem ziemi w 1966 r., pokazujące trzy fazy α, β, γ deformacji skorupy ziemskiej: a) i b) pomiary były robione przy samym epicentrum, c) i d) daleko od epicentrum



Rys. 15. Zmiany średniej miesięcznej poziomu morza przed trzęsieniem ziemi w Niigata w 1964 r.

sztuczne trzęsienia

Seria nieoczekiwanych odkryć, poczynionych w ciągu ostatnich kilkunastu lat, a konkretnie od wczesnych lat sześćdziesiątych, rzuciła światło na możliwości kontrolowania trzęsień ziemi oraz wywoływania sztucznych trzęsień. Odkrycia te dotyczą zjawisk sejsmicznych wyzwalanych przy tworzeniu dużych zbiorników wodnych oraz przy wpompowywaniu cieczy w głąb szybów (jest to metoda usuwania szkodliwych odpadów radioaktywnych i chemicznych). Po raz pierwszy zjawiska takie zarejestrowano we wczesnych latach 40-ych bieżącego stulecia przy budowie zapory wodnej Hoover Dam na jeziorze Mead w USA; powstały zbiornik wodny wyzwolił w uprzednio asejsmicznym rejonie wiele słabych trzęsień ziemi o epicentrach wyraźnie koncentrujących się wokół zbiornika. Analogiczne zjawisko wystąpiło w rejonie zbiornika Koyna w Indiach, napełnianego w latach 1962–65, przy czym w dwa lata po zakończeniu napełniania zbiornika (1967 r.) miało tam miejsce silne trzęsienie ziemi (177 osób zginęło, a 2300 zostało rannych). To ostatnie trzęsienie było potwierdzeniem wcześniej zauważonego braku korelacji czasowej pomiędzy poziomem wody w zbiorniku, a powstającymi trzęsieniami ziemi; jak obecnie wiadomo, wzmożona sejsmiczność ma zwykle pewne opóźnienie czasowe w stosunku do wzrostu ciężaru wody w zbiorniku, jednakże w niektórych wypadkach to opóźnienie jest bliskie zeru. Wyzwalanie zjawisk sejsmicznych w rejonie dużych zbiorników wodnych było obserwowane jeszcze wielokrotnie. Obecnie wiadomo już, że nie zawsze budowa dużego zbiornika wodnego wyzwala w danym rejonie zjawiska trzęsień ziemi (jako przykład służyć może budowa zapory wodnej Grand Coulee, USA, gdzie nie zarejestrowano żadnych trzęsień ziemi). O ile zjawiska te występują, są one zwią-



Rys. 16. Wędrówka kontynentów

zane przypuszczalnie z uaktywnieniem istniejących w tym obszarze starych uskoków tektonicznych na skutek ciężaru wody w zbiorniku oraz przenikania cieczy w pory i szczeliny otaczających zbiornik mas skalnych.

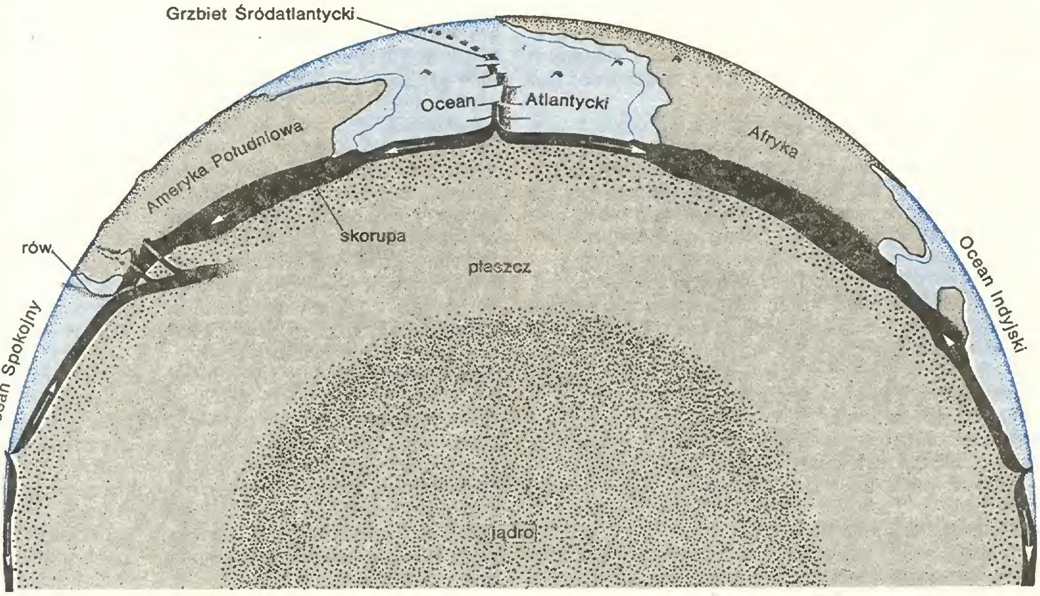
Tego samego typu efekty związane są z wpompowywaniem cieczy w głąb głębokich szybów, co zarejestrowano po raz pierwszy w Denver w stanie Colorado (USA) w latach 1961–65 (głębokość szybu wynosiła ponad 3,5 km). Wyzwalane trzęsienia są zwykle trzęsieniami słabymi, co sugeruje możliwość kontrolowanego powolnego wyzwalania energii sejsmicznej w rejonach o dużym niebezpieczeństwie katastrofalnych trzęsień ziemi, jednakże zebrane dotychczas informacje są jeszcze niewystarczające do stosowania takich metod w praktyce.

Globalne teorie geodynamiczne

Już wiele lat temu zauważono szczególne podobieństwo linii brzegowych poszczególnych kontynentów, głównie zwrócono uwagę na prawie identyczny zarys zachodniego wybrzeża Afryki i wschodniej linii brzegowej Ameryki Południowej. W 1620 r. F. Bacon wyraził pogląd, iż trudno przypuścić, by zbieżność taka była całkowicie przypadkowa, chociaż nie mógł on sobie wyobrazić mechanizmu, który byłby odpowiedzialny za to zjawisko. W połowie XVII w. Francuz F. Placet zasugerował, iż rozbicie kontynentów na istniejące obecnie miało miejsce w czasie czterdziestodniowej powodzi opisywanej przez *Biblię*. Pogląd ten utrzymał się aż do początków XIX w. i znalazł swe odbicie w pracach niem. naukowca A. von Humboldta. Przez cały XIX w. gromadzono dowody ruchu kontynentów polegające na przyporządkowaniu sobie identycznych warstw skalnych i skamieniałości organicznych występujących na brzegach obu tak odległych od siebie lądów jak Afryka i Ameryka Południowa, jednakże dopiero w XX w. zaczęto na serio rozważać możliwość ruchu kontynentów jako całości.

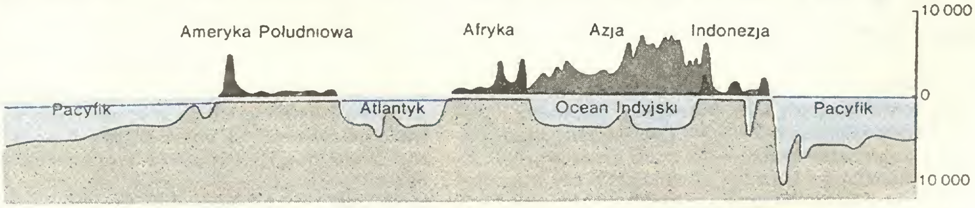
teoria Wegenera

Po raz pierwszy myśl tę zaprezentował geolog amerykański Frank B. Taylor w latach 1908–10, jednakże idea ta jest znana głównie z prac meteorologa i teoretyka niem. A. Wegenera, z których pierwszą przedstawił w 1912 r. W swojej pracy „Pochodzenie kontynentów i oceanów", opublikowanej w 1915 r., Wegener przedstawił teorię ruchu lekkich granitowych mas kontynentalnych po cięższym bazaltowym podłożu den oceanicznych, ruchu wywołanego ruchem obrotowym Ziemi. Rozpad pierwotnego superkontynentu miał miejsce wg Wegenera „miliony lat temu". Konserwatywnie mysiący geologowie odrzucili zdecydowanie tę teorię. Nie zrażony negatywnym przyjęciem Wegener wielokrotnie ponownie przedstawiał udoskonalone wersje swej teorii ruchu kontynentów. Zmarł w 1930 r. podczas wyprawy meteorologicznej na Grenlandię, w czasie, gdy jego teoria była ciągle całkowicie negowana.



Rys. 17. Przesuwanie się płyt tektonicznych i powstawanie Grzbietu Śródatlantyckiego

Sytuacja nie uległa zmianie aż do II wojny światowej, po której, w wyniku rozwinięcia i zastosowania nowych technik instrumentalnych badania den oceanicznych, a także rejestrowania fal sejsmicznych, uzyskano bardzo wiele nowych informacji. Przede wszystkim zwrócono uwagę na ogromne łańcuchy górskie wyrastające z dna oceanów (zob. rys. 18). Największym i najciekawszym z nich okazał się Grzbiet Śródatlantycki, ciągnący się na przestrzeni 19 tys. km,



Rys. 18. Przekrój schematyczny przez skorupę ziemską wzdłuż równika — ukazuje ukształtowanie kontynentów i den oceanicznych. Widać Grzbiet Południowoatlantycki i głębokie rowy w Oceanie Spokojnym.

Grzbiet Śródatlantycki

rozsuwanie się dna oceanów

system większy niż Andy, Góry Skaliste i Himalaje wzięte razem. Przebieg podwodnego grzbietu zgadzał się z zarysem dodatniej anomalii ziemskiego strumienia cieplnego, grzbietowi towarzyszyła także słaba aktywność sejsmiczna i wulkaniczna. Przez środek grzbietu przebiegała długa, wąska dolina, tzw. bruzda grzbietowa. Wysunięto hipotezę, że nowa skorupa ziemska powstaje właśnie w rejonie tej bruzdy, wypływając powoli z głębszego wnętrza Ziemi, a następnie oziębia się i rozsuwa na boki, przy czym procesowi temu towarzyszą trzęsienia ziemi i wybuchy wulkanów. Zgodnie z tą teorią dna oceanów rozszerzają się, powodując rozsuwanie się kontynentów z prędkością kilku centymetrów rocznie. W innych miejscach tego systemu, w rejonie głębokich rowów oceanicznych, związanych również z głębokimi trzęsieniami ziemi, stara skorupa ziemska znika we wnętrzu Ziemi, jest wciągana w głąb, pod kontynenty.

badania paleomagnetyczne

Teoria rozprzestrzeniania się den oceanicznych znalazła swe potwierdzenie w badaniach paleomagnetycznych. Badania paleomagnetyczne to pomiary tej składowej namagnesowania skały, która powstała w momencie tworzenia się skały i oddaje wartość i kierunek pola magnetycznego w tym momencie czasu. Prowadzone intensywnie w latach sześćdziesiątych badania paleomagnetyczne wykazały, że skorupa ziemska rzeczywiście tworzy się w rejonie bruzdy śródgrzbietowej, ulegając następnie rozsunięciu na boki; pomiary te pozwoliły również na określenie prędkości tego procesu. Równocześnie analizowanie wieku badanych skał pozwoliło na uzyskanie obrazu powolnych, globalnych wędrówek kontynentów.

Pangea

Panthalassa

Laurazja

Gondwana

Zgodnie z tą teorią jeszcze 200 mln lat temu obszary lądowe tworzyły jeden ogromny superkontynent, Pangeę (tak jak postulował Wegener), oblany jednym tylko oceanem, Panthalassą. Około 180 mln lat temu ten gigantyczny kontynent rozpadł się początkowo na dwa bloki (rys. 16): północny — Laurazję, złożony z Ameryki Północnej, Europy i Azji, oraz południowy — Gondwanę, złożony z Afryki, Ameryki Południowej, Antarktydy, Australii i Indii. 135 mln lat temu pojawił się grzbiet rozdzielający Afrykę i Amerykę Południową, które zaczęły oddalać się od siebie. Afryka oddzieliła się od Antarktydy, również Indie oderwały się i przesunęły o 8 tys. km na północ, zderzając się z Azją około 40 mln lat temu. W wyniku zderzenia uległ wypiętrzeniu łańcuch Himalajów. Teoria przewiduje również dalszy ruch kontynentów; Australia ma kontynuować posuwanie się na północ, oba oceany: Atlantycki i Indyjski mają się nadal rozszerzać, a Morze Śródziemne kurczyć.

Współcześnie najszerzej akceptowaną globalną teorią dotyczącą procesów geodynamicznych jest tzw. tektonika płyt, powstała w końcu lat sześćdziesiątych, a właściwie jej zmodernizowana wersja, zwana nową tektoniką globalną, opracowywana obecnie. Zasadniczą ideą jest istnienie kilkunastu różnej wielkości płyt (lub bloków) tektonicznych, na obrzeżach których zewnętrzna warstwa Ziemi ulega silnej deformacji. Rysunek 16 przedstawia jak zmieniały się te płyty.

Grubość poszczególnych płyt szacowana jest na 50–100 km. Aktywność tektoniczna koncentruje się w wąskich strefach brzegowych, przy czym odróżnia się trzy podstawowe rodzaje granic pomiędzy płytami: obszary, w których ma miejsce rozszerzanie się granicy i gdzie powstaje nowa skorupa ziemska (grzbiety śródoceaniczne), obszary kompresji, gdzie skorupa ulega zniszczeniu w wyniku zachodzenia na siebie dwu płyt, oraz tzw. uskoki przesuwcze (rys. 17, 18), gdzie płyty poruszają się równolegle do siebie i skorupa ani nie jest produkowana, ani też nie ulega destrukcji.

Teoria płyt zawiera w sobie wcześniejsze teorie rozszerzania się den oceanicznych w wyniku tworzenia się nowej skorupy w regionach grzbietów śródoceanicznych oraz teorię wędrówek kontynentów, zasugerowaną przez Wegenera. W swym pierwotnym kształcie teoria tektoniki płyt, a także obecna nowa tektonika globalna stanowią próbę globalnego opisu i interpretacji wielu różnorodnych zjawisk, jednak bez wyjaśnienia ich pierwotnego mechanizmu. Jakkolwiek dotychczas sugerowano wiele bardzo różnych mechanizmów mających stanowić siłę napędową obserwowanych zjawisk; sprawa ta jest jeszcze nie wyjaśniona i stanowi jeden z najciekawszych otwartych problemów, przed którymi stoi współczesna nauka. Za najciekawsze rozwiązanie uważa się obecnie teorię konwekcji cieplnej występującej w niektórych obszarach górnego płaszcza; równocześnie jednak istnieje wiele innych teorii. Problem ten powinien uzyskać rozwiązanie już w ciągu najbliższych kilku lat, w czasie realizowanego obecnie Projektu Geodynamicznego. W ramach tego programu prowadzone są intensywne badania całości zjawisk geodynamicznych, a także szczególnie dokładnie badane są najciekawsze z punktu widzenia tych zjawisk obszary globu, jak np. rejon Grzbietu Śródatlantyckiego, tzn. obszar, w którym powstaje nowa skorupa ziemska (zob. rys. 17). W ramach trzyletniego francusko-amerykańskiego planu badań, nazwanego FAMOUS (French-American Mid-Ocean Undersea Study), w latach 1973–1975 przebadano dokładnie wybrany obszar dna oceanicznego o powierzchni 156 km² w rejonie Wysp Azorskich. Projekt FAMOUS potwierdził teorię rozszerzania się dna oceanicznego (w tym rejonie prędkość tego procesu wynosi 2–3 cm na rok) oraz dostarczył ogromnej ilości danych, zebranych w trakcie powierzchniowych i podwodnych wypraw eksploracyjnych. Dane te, opracowywane obecnie, umożliwią być może lepsze zrozumienie skomplikowanych procesów geodynamicznych.

K. Balińska-Wuttke *Powstanie i budowa kontynentów*, Warszawa 1970; R. Fraser *Ziemia, planeta, na której żyjemy*, Warszawa 1968; E.W. Janczewski *Zarys sejsmologii ogólnej i stosowanej*, Warszawa 1955; *Planeta Ziemia*, Warszawa 1962; E. Stenz, M. Mackiewicz *Geofizyka ogólna*, Warszawa 1964.

Fizyka atmosfery

Krzysztof Haman

Atmosfera ziemska — gazowa otoczka Ziemi jest podstawowym składnikiem środowiska człowieka. Warunki atmosferyczne miały — i w dalszym ciągu mają — zasadnicze znaczenie dla tak kluczowych dziedzin działalności ludzkiej jak rolnictwo czy transport i wyciskają swoje piętno na różnych przejawach życia społecznego. Próby poznania praw rządzących zarówno pojedynczymi zjawiskami pogody jak i ich właściwościami statystycznymi czyli klimatem, a także próby kształtowania ich w sposób pożądany, towarzyszą chyba całej historii ludzkości niezależnie od technicznych możliwości ich realizacji. Badania atmosfery prowadzone w ostatnich dziesięcioleciach przyniosły w tej dziedzinie poważne sukcesy, a także doprowadziły do wykrycia w atmosferze i wyjaśnienia wielu nowych zjawisk wykraczających poza potocznie rozumianą pogodę. Jednakże bardzo wiele — czasem bardzo pospolitych — zjawisk atmosferycznych jak chociażby zwykły deszcz kryje w sobie sporo dotąd niewyjaśnionych zagadek. Trzeba też pamiętać, że atmosfera ziemska należy do najbardziej zagrożonych działalnością człowieka elementów środowiska.

Wszystko to powoduje rosnące zainteresowanie naukami o atmosferze ziemskiej, wśród których na czołowe miejsce wysuwa się fizyka atmosfery.

Przez termin fizyka atmosfery rozumieć będziemy

całokształt wiedzy o zjawiskach fizycznych zachodzących w atmosferze ziemskiej. Fizyka atmosfery jest nauką złożoną wewnętrznie, odzwierciedlającą zarówno strukturę samej fizyki jak i strukturę atmosfery. Mamy więc z jednej strony np. mechanikę, termodynamikę, elektryczność atmosfery, z drugiej zaś — działy fizyki atmosfery poświęcone kompleksowym badaniom określonych zjawisk atmosferycznych, jak np. fizyka chmur lub fizyka wysokich warstw atmosfery. Na szczególne wyodrębnienie ze względu na znaczenie praktyczne zasługuje meteorologia — dział fizyki atmosfery zajmujący się zespołem zjawisk potocznie zwanych pogodą. Zjawiska te koncentrują się głównie w dolnych kilkudziesięciu lub nawet kilkunastu kilometrach atmosfery i są natury głównie mechanicznej i termodynamicznej.

Pomocną przy czytaniu tego hasła, poświęconego głównie problemom meteorologii, może być il. 8 z tabl. 1, zawierająca szereg informacji o atmosferze.

Zasady fizycznego opisu atmosfery

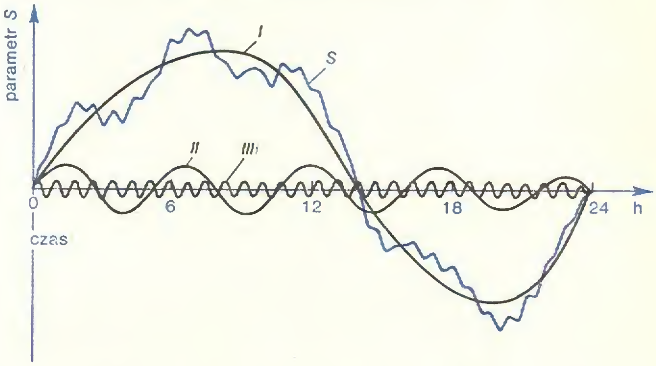
Zjawiska fizyczne w atmosferze opisuje się ilościowo podając wartości pewnych parametrów fizycznych (np. temperatury, ciśnienia, prędkości wiatru itp.). Najczęściej podaje się je w funkcji czasu i przestrzeni, czyli w postaci tzw. pól. Pierwotnym źródłem informacji o tych zjawiskach jest — jak we wszystkich naukach przyrodniczych — obserwacja i pomiar. W odróżnieniu od innych działów fizyki, w których dominują pomiary i obserwacje dokonywane podczas eksperymentów laboratoryjnych — kontrolowanych i powtarzalnych, zjawiska atmosferyczne bada się na ogół obserwując procesy naturalne, nie poddające się kontroli obserwatora i powtarzalne o tyle, o ile leży to w charakterze zjawiska. Sytuacja taka jest charakterystyczna w geofizyce, a także w astrofizyce i astronomii. W geofizyce, a zwłaszcza w licznych działach fizyki atmosfery, dochodzi jeszcze jedna trudność — wiele zjawisk ze względu na ich rozległość nie mieści się w polu widzenia pojedynczego obserwatora i ich obraz uzyskuje się dopiero przez syntezę obserwacji wykonanych przez różnych obserwatorów. W wyniku tej syntezy otrzymuje się mapy lub przekroje pól parametrów atmosferycznych (w formie graficznej lub cyfrowej), których najpospolitszym przykładem może być np. mapa ciśnienia przedstawiona w telewizyjnej prognozie pogody. Mapy te uzyskiwane są przez interpolację między zwykle dość odległymi w czasie i przestrzeni pomiarami. W zależności od gęstości tych pomiarów oraz sposobów ich nanoszenia i interpolacji otrzymane mapy mogą być bardziej lub mniej szczegółowe czyli — jak się to w geofizyce zwykle mówi — mogą przedstawiać składowe pól o różnych skalach, które służą do opisywania zjawisk różnych skal.

pola parametrów fizycznych

skala zjawisk atmosferycznych

Pojęcie skali pola i ogólnie skali odgrywa wielką rolę w naukach geofizycznych. Ścisła definicja tego pojęcia jest dość skomplikowana, ale jego sens ogólny sprowadza się do charakterystyki przestrzennych rozmiarów zjawiska (skala przestrzenna) oraz czasu jego trwania (skala czasowa); skala charakteryzuje jednocześnie gęstość sieci obserwacyjnej niezbędnej do dostatecznie dokładnej obserwacji danego zjawiska.

Dla przykładu rozpatrzymy skalę czasową temperatury. Obserwując bieg temperatury przez dłuższy czas stwierdzimy, że można przedstawić go jako sumę średniej wieloletniej temperatury w danym miejscu oraz wahań sezonowych związanych ze zmianami pór roku, wahań dobowych związanych z cyklem dnia i nocy, a także nieregularnych oscylacji, w których przedziały czasu pomiędzy kolejnymi przejściami — od wzrostu do spadku temperatury — wynoszą od kilku do kilkunastu minut; za pomocą bardzo czułych i szybko reagujących termometrów dałoby się wykryć całkiem znaczne wahania temperatury zachodzące w czasie kilku sekund lub mniejszym. Jest przy tym rzeczą oczywistą, że scharakteryzowanie rocznego przebiegu temperatury jest możliwe przy pomiarach wykonywanych np. raz na dzień, biegu dobowego przy pomiarach wykonywanych co godzina, podczas gdy wyznaczenie pozostałych wymienionych składowych wypadkowego biegu temperatury wymaga pomiarów znacznie częstszych — co najmniej kilku w każdym okresie o ustalonym kierunku zmian temperatury (rys. 1). Liczbę charakteryzującą w przybliżeniu czas, w ciągu którego dana składowa biegu temperatury zachowuje znak pochodnej — tzn. wzrasta lub maleje — będziemy nazywać skalą czasową tej składowej.



Rys. 1. Obserwowany przebieg parametru meteorologicznego S można zwykle rozłożyć (choć nie w sposób jednoznaczny) na sumę szeregu składowych o różnych skalach czasowych. Krzywa I — skala czasowa rzędu doby, II — skala czasowa rzędu godzin, III — skala czasowa rzędu minut. Przebieg na rysunku jest fikcyjny i stanowi jedynie wyjaśnienie. Podobnie można by przedstawić składowe o różnych skalach przestrzennych

W podobny sposób można wprowadzić pojęcie skali przestrzennej (pionowej lub poziomej) danej składowej pola temperatury lub innego parametru fizycznego. Zwróćmy uwagę, że tak wprowadzona skala nie jest dokładnie określona i składowa pola o danej skali może być wyodrębniona na różne sposoby prowadzące do nieco odmiennych wyników. W badaniach geofizycznych składowe pola o różnych skalach rozpatrujemy zazwyczaj jako odrębne obiekty badań, ponieważ dysponując ograniczonymi możliwościami zbierania i przetwarzania informacji o polach parametrów geofizycznych musimy wybierać pomiędzy dokładniejszą analizą mniejszego (w sensie rozciągłości w czasie i przestrzeni) obiektu lub mniej dokładną — większego. Ponadto składowe te mają zazwyczaj odmienne przyczyny fizyczne. Dlatego rozpatrując różne prawidłowości należy zawsze brać pod uwagę, do jakich skal danego zjawiska one się odnoszą.

badanie składowych pola o różnych skalach

Nakładające się na siebie pola różnych skal można porównać do interferujących fal elektromagnetycznych o odmiennych długościach i okresach. W odniesieniu do fal elektromagnetycznych stosuje się na ogół zasadę prostego sumowania, tak że rozchodzenie się jednej fali nie ma wpływu na rozchodzenie się innych. Procesy atmosferyczne różnych skal natomiast na ogół na siebie oddziałują, np. pola skali większej zmieniają swoje zachowanie, gdy nakładają się na nie pola skali mniejszej i odwrotnie. Uwzględnienie tych oddziaływań jest jednym z najtrudniejszych problemów w fizyce atmosfery (i w geofizyce w ogóle).

Budowa i skład atmosfery

Atmosfera ziemska znajduje się stale w stanie bardzo bliskim lokalnej równowagi hydrostatycznej w tym sensie, że różnica ciśnienia atmosferycznego pomiędzy

rozkład ciśnienia

dwoma poziomami jest zawsze bardzo bliska ciężarowi zawartego pomiędzy nimi pionowego słupa powietrza o jednostkowej powierzchni przekroju poprzecznego. Dokładność tego przybliżenia wynosi ok. 10^{-4}% i poważniejsze odchylenia (z reguły nie przekraczające 1%) zdarzają się jedynie w niewielkich obszarach bardzo intensywnych ruchów powietrza. Wynika to stąd, że przyspieszenia w ruchach atmosfery o znaczniejszej rozciągłości przestrzennej są z reguły małe w porównaniu z przyspieszeniem grawitacyjnym. Fakt ten prowadzi do charakterystycznego pionowego rozkładu ciśnienia (a więc i masy) w atmosferze; przykładem

Typowy wysokościowy rozkład ciśnienia w atmosferze

Wysokość km n.p.m.	Ciśnienie Pa	% masy atmosfery powyżej
0	$1,01 \cdot 10^{5}$	100
10	$2,65 \cdot 10^{4}$	26,2
20	$5,5 \cdot 10^{3}$	5,4
50	$9 \cdot 10^{1}$	0,1
100	$1,74 \cdot 10^{-2}$	$1,7 \cdot 10^{-5}$
200	$1,95 \cdot 10^{-4}$	
500	$2,9 \cdot 10^{-6}$	
1000	$1,9 \cdot 10^{-8}$	

takiego rozkładu mogą być dane przytoczone w tabeli. Ogólnie można przyjąć, że ciśnienie do wysokości ok. 100 km — a wraz z nim i masa atmosfery pozostająca ponad danym poziomem — spada dziesięciokrotnie co 16–20 km. Podobnie przedstawiają się zmiany gęstości w atmosferze. Zatem troposfera (średnio pierwsze 10 km atmosfery) zawiera ok. 75% masy atmosfery, a wraz ze stratosferą (średnio pierwsze 50 km) — ponad 99,9%. Typowy skład suchego powietrza atmosferycznego w pobliżu powierzchni Ziemi przedstawia tabela.

Koncentracja głównych składników atmosfery suchej do wysokości 90 km

Składnik	% masy powietrza suchego	Składnik	% masy powietrza suchego
Azot N_2	75,527	Neon Ne	$1,25 . 10^{-3}$
Tlen O_2	24,143	Krypton Kr	$3,3 \cdot 10^{-4}$
Argon A	1,282	Hel He	$7,24 \cdot 10^{-5}$
Dwutlenek węgla CO_2	$4,5–6 \cdot 10^{-2}$	Wodór H	$3,48 \cdot 10^{-6}$

procesy wpływające na skład atmosfery

Koncentracja poszczególnych składników atmosfery kształtuje się lokalnie w wyniku działania procesów dwóch typów — produkcji (dodatniej, a także ujemnej — czyli likwidacji) oraz transportu. Produkcja jest wynikiem działalności źródeł (dodatnich i ujemnych) takich, jak reakcje chemiczne, przemiany fazowe, adsorpcja i wydzielanie przez minerały i wodę oceaniczną, fotosynteza itp. Transport jest superpozycją unoszenia przez uporządkowany ruch powietrza, mieszania przez nieuporządkowane ruchy różnych skal oraz dyfuzji molekularnej. Produkcja przy pominięciu pozostałych procesów prowadzi do lokalnych zmian koncentracji w obszarze źródeł; mieszanie — do powstania mieszaniny o składzie jednorodnym przestrzennie, dyfuzja zaś (której intensywność rośnie wraz ze średnią drogą swobodną molekuł) do ustalenia równowagi hydrostatycznej każdego składnika z osobna, co prowadzi z kolei do względnego wzrostu koncentracji składników lżejszych wraz ze wzrostem wysokości.

skład atmosfery

Stwierdzany pomiarami wysoki stopień jednorodności atmosfery w odniesieniu do jej głównych składników (O_2, N_2, Ar) w warstwie do wysokości 50–60 km świadczy o tym, że dominującym czynnikiem transportu jest mieszanie, zaś wydajność źródeł jest stosunkowo mała. Powyżej tej wysokości zaczynają się zaznaczać procesy dysocjacji pod wpływem promieniowania Słońca oraz rekombinacji molekuł głównych gazów atmosferycznych. Prowadzi to do lokalnych zmian składu atmosfery — pojawienia się większych ilości tlenu i azotu atomowego. Względna wydajność tych procesów traktowanych jako źródła jest znaczna, lecz wobec wielkiego rozrzedzenia wysokich warstw atmosfery ich udział w łącznym bilansie gazów atmosferycznych jest niewielki. Powyżej 120 km droga swobodna molekuł staje się tak duża, że dyfuzja zaczyna dominować jako mechanizm transportu i w atmosferze zaczyna stopniowo wzrastać koncentracja gazów lekkich — wodoru i helu, które stają się głównymi składnikami atmosfery na wysokości powyżej 1000 km.

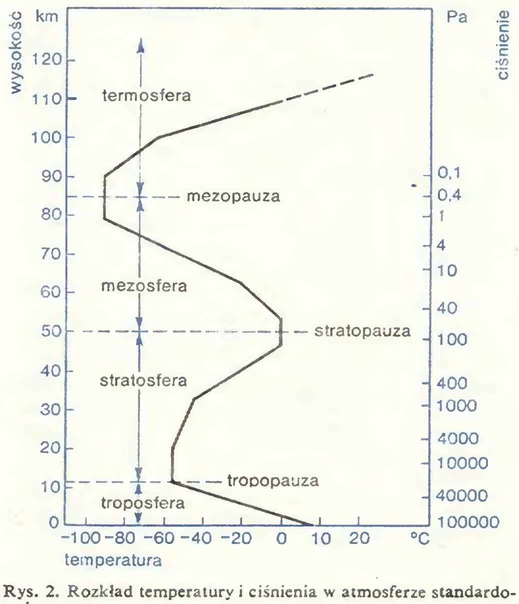
Głównymi źródłami azotu, tlenu i dwutlenku węgla są procesy fotosyntezy i przemiany materii w szacie roślinnej lądowej i morskiej a ewolucja tej szaty jest sprzężona z ewolucją składu atmosfery. Na uwagę zasługują też ze względu na swoje znaczenie biologiczne lub pośredni wpływ na ważne zjawiska atmosferyczne niektóre składniki atmosfery występujące w ilościach śladowych (O_3, NH_4, SO_2, N_2O).

Szczególnym składnikiem atmosfery jest para wodna — jedyny gaz atmosferyczny przechodzący w warunkach atmosferycznych w stan ciekły i stały. W odniesieniu do pary wodnej procesy produkcji i likwidacji (parowanie i kondensacja wody) często dominują nad mieszaniem — stąd znaczne niejednorodności jej koncentracji zawierającej się w granicach 0–4%. Dzięki dużym wartościom ciepła topnienia i kondensacji przemiany fazowe wody są istotnym czynnikiem energetycznym w atmosferze, a związane z nimi zjawiska — powstawanie chmur i opadów — należą do najważniejszych dla człowieka procesów pogodowych.

Ważnym składnikiem atmosfery są też mikroskopijne cząstki ciał stałych i cieczy unoszące się w powietrzu — tzw. aerozol atmosferyczny. Choć jego koncentracja rzadko przekracza 10^{-6}, odgrywa on ważną rolę w licznych procesach atmosferycznych oraz może mieć bezpośrednie działanie biologiczne.

pionowa struktura atmosfery

W fizyce atmosfery stosuje się różnego rodzaju podziały atmosfery na warstwy. Najczęściej stosowanym kryterium jest pionowy rozkład temperatury, według którego wyróżnia się kolejno: 1) troposferę — warstwę atmosfery ziemskiej, w której temperatura maleje wraz ze wzrostem wysokości niemal jednostajnie; rozciąga się ona od powierzchni Ziemi do wysokości ok.



Rys. 2. Rozkład temperatury i ciśnienia w atmosferze standardowej

7 km w okolicach podbiegunowych, a do ok. 16 km w pobliżu równika; 2) stratosferę — warstwę atmosfery ziemskiej o temperaturze prawie stałej (stratosfera dolna) lub rosnącej wraz z wysokością (stratosfera górna); 3) mezosferę — warstwę atmosfery ziemskiej, w której ponownie następuje spadek temperatury wraz z wysokością i 4) termosferę — warstwę atmosfery, w której ponownie wzrasta temperatura (do kilku tys. K); rozciąga się ona aż do granic atmosfery. Powierzchnie rozgraniczające te warstwy nazywają się odpowiednio: tropopauza, stratopauza i mezopauza. Standardowy rozkład temperatury w atmosferze przedstawia rys. 2. Rozkłady rzeczywiste mogą dość znacznie różnić się od niego w zależności od położenia geograficznego, pory dnia lub roku. W atmosferze ziemskiej wyróżnia się również warstwy wg innych niż temperatura kryteriów fizycznych: np. jonosfera jest warstwą powyżej 80 km, którą charakteryzuje znaczny stopień jonizacji, ozonosfera — warstwą na wysokości 25–40 km, która odznacza się znacznie podwyższoną koncentracją ozonu (ok. 10^{-5} wobec ok. 10^{-8} przy powierzchni Ziemi); egzosfera — warstwą powyżej 300 km, która charakteryzuje się ucieczką najszybszych molekuł gazów atmosferycznych w przestrzeń międzyplanetarną. Wyodrębnia się też tzw. planetarną warstwę graniczną, obejmującą pierwsze 1,5–2 km atmosfery, w której ruch powietrza w sposób istotny jest zakłócany przez tarcie o powierzchnię Ziemi.

Promieniowanie słoneczne źródłem energii w atmosferze

Pierwotnym źródłem energii, dzięki której zachodzą zjawiska fizyczne w atmosferze jest promieniowanie słoneczne. Dla pewnej grupy zjawisk (np. zjawisk optycznych, zorzy polarnej) jest to bezpośrednie źródło energii, jednak dla większości zjawisk badanych obecnie przez fizykę atmosfery jest to głównie źródło energii cieplnej, która zostaje następnie zmieniona w inne formy energii w różnych procesach termodynamicznych. Promieniowanie słoneczne docierające do górnych granic atmosfery jest w przybliżeniu (dosyć dokładnym) promieniowaniem ciała doskonale czarnego o temperaturze ok. 6000 K nieco zdeformowanym przez zjawiska zachodzące w atmosferze słonecznej. Niemal cała energia tego promieniowania jest skoncentrowana w promieniowaniu widzialnym i bliskiej podczerwieni. Promieniowanie krótkofalowe (nadfioletowe i krótsze) oraz promieniowanie korpuskularne, odgrywa znaczną rolę w fizyce wysokich warstw atmosfery, lecz stanowi niewielki ułamek całej energii słonecznej docierającej do Ziemi.

pochłanianie, rozpraszanie i odbicie

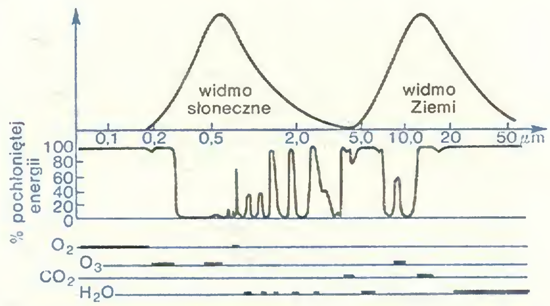
W atmosferze ziemskiej promieniowanie słoneczne ulega częściowemu pochłanianiu, rozpraszaniu i odbiciu na drobnoskalowych niejednorodnościach gęstości powietrza, cząstkach chmurowych i aerozolu atmosferycznym — częściowo zaś dociera do powierzchni Ziemi. Pochłaniane jest głównie promieniowanie nadfioletowe (przez tlen i ozon w górnych warstwach atmosfery); znaczniejsze pochłanianie w szerokim przedziale widmowym występuje niekiedy przy dużej koncentracji aerozolu. Ogólnie pochłanianie promieniowania słonecznego w atmosferze jest niewielkie i wynosi ok. 10%. Rozpraszanie przebiega różnie w różnych kierunkach i zależy od długości fali świetlnej. W kierunku poprzecznym do padających promieni najsilniej jest rozpraszane promieniowanie o falach najkrótszych — stąd przewaga niebieskiej części widma w barwie nieba, które jest barwą światła rozproszonego. Odbicie następuje głównie od cząstek dużych w porównaniu z długością fali świetlnej — głównie od cząstek chmurowych. Promieniowanie odbite od obiektów atmosferycznych i od powierzchni Ziemi w przestrzeń kosmiczną oraz promieniowanie rozproszone wstecz, nie bierze udziału w procesach atmosferycznych jako źródło energii. Stosunek natężenia tego promieniowania do natężenia promieniowania padającego nosi nazwę albedo. Średnie albedo Ziemi dla widma słonecznego wynosi ok. 0,3.

albedo

Nie odbite promieniowanie słoneczne zostaje pochłonięte przez obiekty atmosferyczne i powierzchnię Ziemi stając się składnikiem ich bilansu cieplnego. Z kolei Ziemia i atmosfera promieniują same. Skład widmowy promieniowania powierzchni Ziemi odpowiada w przybliżeniu (z dosyć dużą dokładnością) promieniowaniu ciała doskonale czarnego o temperaturze tej powierzchni — niemal cała jego energia przypada na daleką podczerwień (maksimum dla fal o długości 10–12 μm). W tym obszarze widmowym leżą intensywne pasma pochłaniania pary wodnej i CO_2 (rys. 3). Przypowierzchniowa warstwa powietrza grubości kilkudziesięciu metrów pochłania niemal całkowicie promieniowanie leżące w tych pasmach. Wypromieniowanie z powierzchni Ziemi może się odbywać jedynie w drodze tzw. transferu promieniowania, tzn. pochłaniania promieniowania przez kolejne warstwy atmosfery i retransmitowania go dalej jako własnego promieniowania cieplnego o odpowiednio zmienionym składzie widmowym. Bezpośrednio w przestrzeń kosmiczną powierzchnia Ziemi promieniuje jedynie w przedziale tzw. okna atmosferycznego — przedziału widmowego ok. 9–12 μm wolnego od pasm pochłaniania głównych składników atmosfery. Dopiero powyżej 6–10 km możliwe staje się bezpośrednie wypromieniowanie z atmosfery w przestrzeń kosmiczną większych ilości energii. Ponieważ retransmisja promieniowania odbywa się zarówno w kierunku warstw leżących wyżej jak i w kierunku powierzchni Ziemi, znaczna część energii wypromieniowanej przez powierzchnię Ziemi wraca do niej z powrotem. W rezultacie temperatura powierzchni Ziemi ustala się na poziomie wyższym niżby to zachodziło przy braku atmosfery. Efekt ten wynikający ze znacznej przezroczystości atmosfery dla promieniowania słonecznego i nieprzezroczystości dla podczerwonego promieniowania Ziemi nosi tradycyjną nazwę efektu inspektowego, ponieważ uważano dawniej (nie całkiem słusznie), że jest on analogiczny do procesu podnoszenia temperatury w szklarniach i inspektach ogrodniczych.

promieniowanie Ziemi

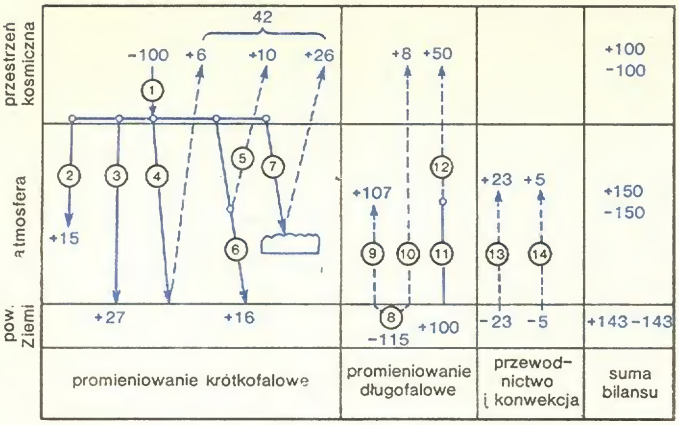
efekt inspektowy



Rys. 3. Widmo promieniowania ciała doskonale czarnego w temperaturze 6000 K (co w przybliżeniu odpowiada promieniowaniu słonecznemu) i 270 K (co w przybliżeniu odpowiada promieniowaniu Ziemi), oraz położenie pasm pochłaniania najważniejszych składników atmosfery. Atmosfera jest przezroczysta dla znacznej części promieniowania słonecznego i niemal nieprzezroczysta dla promieniowania Ziemi

bilans cieplny

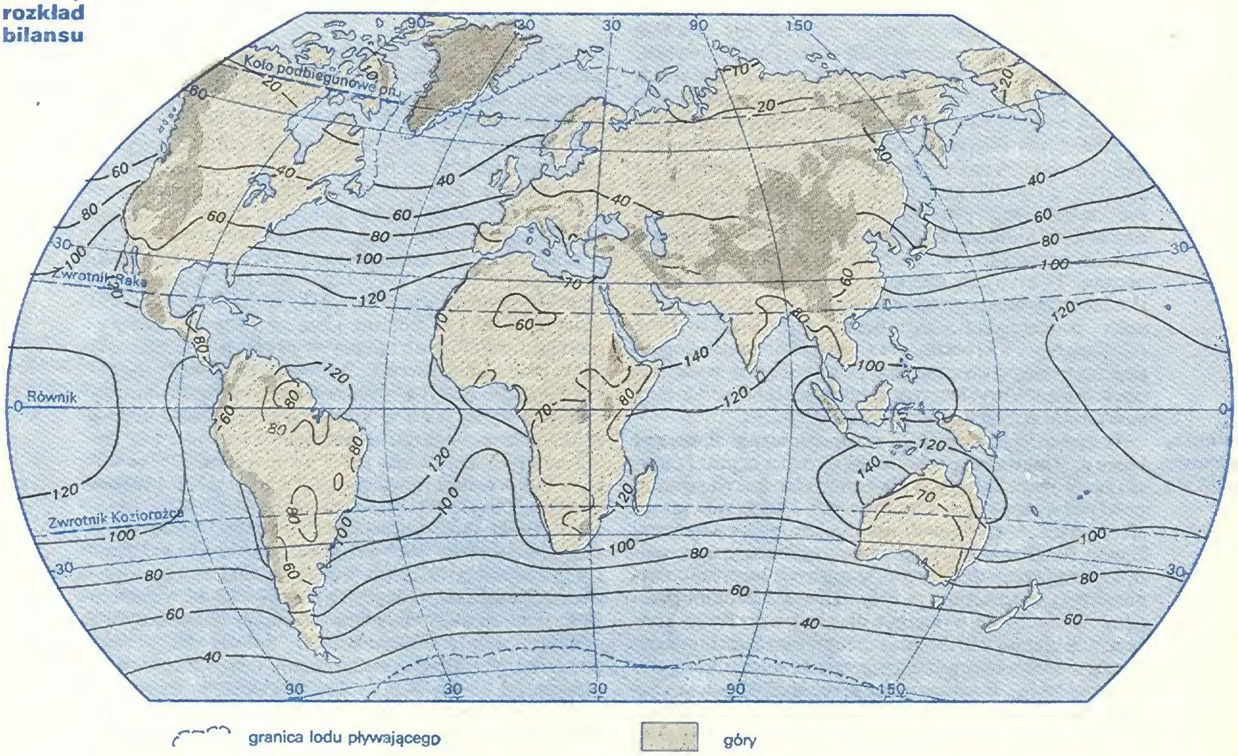
Pionowy rozkład temperatury w atmosferze jest wynikiem równowagi pomiędzy transferem promieniowania podczerwonego, transportem ciepła przez konwekcję i turbulencję, wydzielaniem ciepła utajonego w przemianach fazowych wody oraz w pewnym stopniu (głównie w wyższych warstwach atmosfery) także bezpośrednim pochłanianiem promieniowania słonecznego i molekularnym przewodnictwem cieplnym (rys. 4). W troposferze dominują trzy pierwsze czynniki, w stratosferze dolnej — transfer promienio-



Rys. 4. Schemat średniego bilansu cieplnego Ziemi i atmosfery w jednostkach umownych (stała słoneczna = 100); *1* promieniowanie słoneczne dochodzące do górnych granic atmosfery, *2* pochłanianie w atmosferze, *3* pochłanianie na powierzchni Ziemi, *4* odbicie od powierzchni Ziemi, *5* i *6* rozpraszanie w atmosferze, *7* odbicie od chmur, *8* promieniowanie powierzchni Ziemi, z tego *9* pochłonięte w atmosferze, *10* przepuszczone przez atmosferę, *11* i *12* promieniowanie atmosfery, *13* transport ciepła w postaci utajonej, *14* transport ciepła w postaci „odczuwalnej" (wg S. P. Chromowa i L. I. Mamontowej)

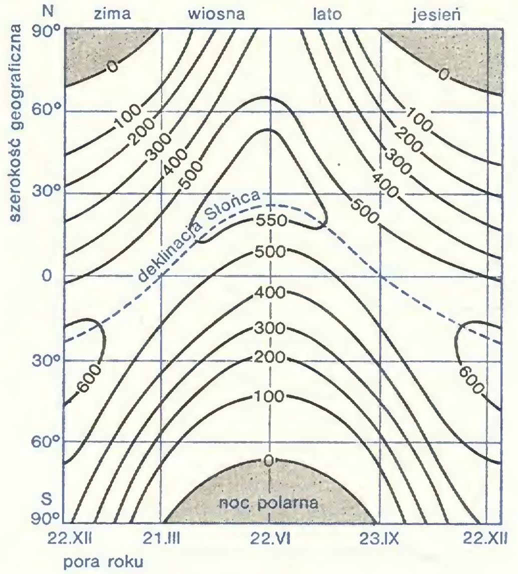
wania, do którego w stratosferze górnej dołącza się pochłanianie nadfioletowego promieniowania słonecznego przez ozon. Wraz ze zmniejszaniem się gęstości powietrza, coraz większego znaczenia nabiera pochłanianie promieniowania krótkofalowego i korpuskularnego Słońca w procesach fotochemicznych oraz przewodnictwo molekularne; czynniki te stają się dominującymi w bilansie cieplnym termosfery. Tem-

geograficzny rozkład bilansu



peratura powierzchni Ziemi, która ma zasadnicze znaczenie w kształtowaniu się warunków atmosferycznych w troposferze, jest konsekwencją bilansu cieplnego, w którym poza promieniowaniem i wymianą ciepła z atmosferą przez przewodnictwo i konwekcję istotną rolę odgrywa transport ciepła w głąb podłoża oraz jego ciepło właściwe. Parametry te, podobnie jak albedo, mogą wykazywać znaczne różnice lokalne (np. pustynia i ocean).

Lokalne różnice w dopływie promieniowania słonecznego zależą od położenia geograficznego, pory dnia i roku oraz od stanu przezroczystości atmosfery (zachmurzenia, zapylenia oraz albedo). Pomijając osłabienie w atmosferze, największą ilość promieniowania w skali doby uzyskują rejony podbiegunowe w czasie dnia polarnego — najmniejszą, te same rejony w czasie nocy polarnej (rys. 5). W ciągu roku najwięcej



Rys. 5. Rozkład dopływu bezpośredniego promieniowania słonecznego do górnych granic atmosfery w zależności od pory roku i szerokości geograficznej (cal/(cm²·d))

Rys. 6. Geograficzny rozkład rocznego bilansu promieniowania powierzchni Ziemi w kcal/cm² (1 kcal = 0,41868·10⁴J). Zależy on nie tylko od szerokości geograficznej, ale także zachmurzenia i przezroczystości atmosfery

promieniowania otrzymują obszary międzyzwrotnikowe. Także bilans promieniowania pochłanianego i emitowanego wykazuje znaczne zróżnicowanie (rys. 6). Obszary podzwrotnikowe (powierzchnia Ziemi i atmosfera łącznie) mają natomiast w ciągu roku średnio nadwyżki promieniowania, podczas gdy okolice podbiegunowe — niedobór. Aby lokalnie temperatura utrzymywała się w obserwowanych granicach, bilans cieplny musi być wyrównywany inną formą transportu energii. Jest nią przede wszystkim tzw. unoszenie ciepła przez ruchy powietrza. Ruchy powietrza transportują ciepło zarówno w formie „odczuwalnej" tzn. związanej bezpośrednio z temperaturą powietrza jak i utajonej w parze wodnej. Ciepło utajone jest oddawane atmosferze podczas kondensacji w chmurach. Występuje przy tym sprzężenie zwrotne, ponieważ właśnie przestrzenne niejednorodności w bilansie cieplnym powodują powstawanie różnic temperatur, które umożliwiają zamianę energii cieplnej w inne postaci energii. W szczególności dzięki nim funkcjonuje cyrkulacja atmosferyczna i oceaniczna, stanowiąca podłoże niemal wszystkich zjawisk będących przedmiotem badań współczesnej fizyki atmosfery.

Dynamika atmosfery

Ruchy powietrza atmosferycznego rozpatruje się zwykle w układzie odniesienia związanym z obracającą się wokół osi Ziemią. Są one wynikiem działania różnorodnych sił, które zazwyczaj odnosi się do jednostki masy lub jednostki objętości powietrza.

siły działające w atmosferze

Najważniejsze z nich to:

— gradient ciśnienia,

— siła ciążenia — mająca największe znaczenie w statyce oraz ruchach pionowych atmosfery;

— siła Coriolisa związana z dobowym obrotem Ziemi — skierowana prostopadle do wektora prędkości i do osi obrotu Ziemi.
Praktyczne znaczenie ma jedynie część jej składowej poziomej, która w odniesieniu do jednostki masy jest dana wzorem $(4\pi/T)v\sin\varphi$, gdzie T długość doby, v — prędkość pozioma powietrza, φ — szerokość geograficzna. W dalszym ciągu przez termin „siła Coriolisa" rozumieć będziemy właśnie tę wielkość. Jest ona skierowana prostopadle do wektora prędkości w prawo na półkuli północnej — w lewo na południowej.

— tarcie wewnętrzne (lepkość molekularna), które jest wynikiem molekularnej dyfuzji pędu makroskopowego. Z energetycznego punktu widzenia prowadzi ono ostatecznie do zamiany energii kinetycznej w cieplną (dyssypacja energii).

Efektem działania tych sił jest przyspieszenie ruchu powietrza. W pewnych sytuacjach siły te w znacznym stopniu się równoważą, tak że ich wypadkowa jest znacznie mniejsza od głównych składników; mówi się wówczas, że składniki te są w stanie kwazirównowagi.

Ponieważ w dynamice atmosfery z reguły rozważa się ruchy jedynie powyżej pewnej skali, której reprezentację umożliwia zdolność rozdzielcza stosowanej mapy, należy uwzględnić wpływ na ruch atmosfery ruchów mniejszych skal (tzw. ruchów podskalowych).

lepkość turbulencyjna powietrza

Fizycznie oddziaływanie to jest równoważne pojawieniu się dodatkowych sił — podobnie jak wzajemne oddziaływanie ruchu makroskopowego gazu z ruchem molekularnym równoważne jest pojawieniu się sił tarcia wewnętrznego. Jeżeli efekty te są spowodowane ruchami bardzo małych skal — rzędu kilkuset metrów i mniej — analogia ta sięga tak daleko, że można wręcz mówić o tzw. lepkości turbulencyjnej powietrza. Charakteryzujący ją współczynnik lepkości turbulencyjnej jest z reguły 10^3–10^5 razy większy od współczynnika lepkości molekularnej.

Rola poszczególnych sił jest różna w różnych kategoriach ruchów. Tak więc np. tarcie molekularne (poza wysokimi warstwami atmosfery) ma istotne znaczenie jedynie w ruchach bardzo małych skal (centymetry i mniej), którymi zajmuje się raczej aerodynamika niż geofizyka (choć ma pośredni wpływ również na procesy większych skal). Tarcie turbulencyjne jest istotnym czynnikiem w ruchach skal małych i średnich (do kilkudziesięciu kilometrów), w ruchach zaś o większych skalach przestrzennych odgrywa rolę głównie w obszarach, w których pionowe gradienty prędkości wielkoskalowych są duże (warstwy tarcia). Odbywa się to głównie w pobliżu powierzchni Ziemi (planetarna warstwa graniczna), choć zdarza się także na wysokościach większych. W obszarach poza warstwami tarcia czyli w tzw. atmosferze swobodnej, w odniesieniu do ruchów o skalach największych (powyżej 1000 km w poziomie), występuje na dużych obszarach Ziemi tzw. równowaga kwazigeostroficzna, tzn. stan kwazirównowagi między siłą Coriolisa i siłą gradientu ciśnienia (można pominąć siłę tarcia).

własności ruchów różnych skal

oddziaływanie międzyskalowe

Wzajemne oddziaływanie ruchów różnych skal jest często charakteryzowane przez rozmiary i kierunek przekazywania energii kinetycznej od skali do skali. Na przykład ruchy o skalach poniżej kilkuset metrów (nazywane potocznie turbulencją atmosferyczną) charakteryzuje przekazywanie energii kinetycznej ku coraz mniejszym skalom aż do ruchów molekularnych, w których ulega dyssypacji (zamianie na ciepło). W tym sensie lepkość turbulencyjna wpływa hamująco na ruchy skal większych (odbierając energię kinetyczną), podobnie jak lepkość molekularna hamuje ruchy makroskopowe. Przy większych skalach (rzędu setek i tysięcy kilometrów) obserwuje się jednak także przekazywanie energii kinetycznej skalom większym. Chcąc w tym przedziale skal traktować oddziaływanie skal mniejszych na większe jako analog lepkości, należałoby niekiedy mówić o lepkości ujemnej.

niestabilność

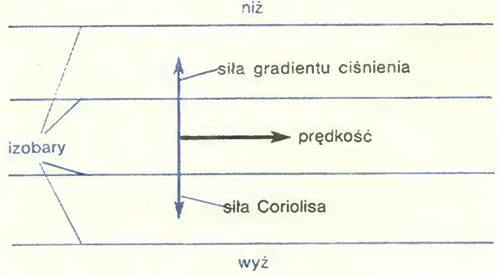
W oddziaływaniu ruchów skal większych na mniejsze ważną funkcję spełnia zjawisko niestabilności. Polega ono na tym, że obszary dużych gradientów temperatury i prędkości mogą być niestabilne względem ruchów skal mniejszych, tzn. powstające w tych obszarach przypadkowe, niewielkie zaburzenia ruchu o odpowiedniej skali mają tam tendencje do samorzutnej intensyfikacji. Fizyczne mechanizmy zjawiska niestabilności są dość złożone i nie będą tu szczegółowo omawiane.

równowaga kwazihydrostatyczna

Analizę dynamiki atmosfery ułatwia występowanie w atmosferze stanów kwazirównowagi. Na przykład rozkład ciśnienia w atmosferze jest lokalnie bardzo bliski rozkładowi hydrostatycznemu, dzięki czemu można w wielu zagadnieniach wyznaczać z dostateczną dokładnością pole gęstości (a za pomocą równania stanu także pole temperatury) z pola ciśnienia. W ten sposób zmniejsza się liczba parametrów, których ewolucję trzeba analizować niezależnie.

równowaga kwazigeostroficzna

Podobne znaczenie ma równowaga kwazigeostroficzna (rys. 7), dzięki której wiatr rzeczywisty można w wielu rozważaniach zastępować tzw. wiatrem geostroficznym, wyznaczonym z warunku równoważenia poziomej składowej sił gradientu ciśnienia przez siłę Coriolisa (rys. 7). Wiatr geostroficzny wieje równo-



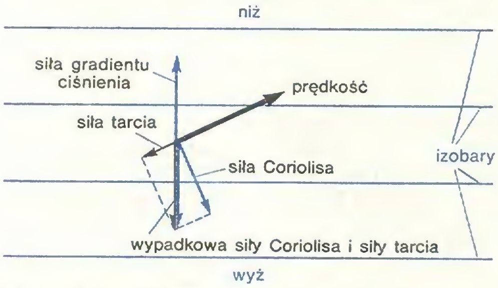
Rys. 7. Równowaga geostroficzna — w wietrze geostroficznym gradient ciśnienia (spadek na jednostkę odległości w kierunku prostopadłym do izobar) jest równoważony przez siłę Coriolisa

legle do izobar w ten sposób, że na półkuli północnej niskie ciśnienie pozostaje na lewo od kierunku, w którym wieje wiatr, a na południowej — na prawo. Równowaga kwazigeostroficzna w atmosferze swobodnej jest realizowana z dokładnością 10–15%, a więc znacznie mniejszą niż równowaga hydrostatyczna. W sytuacjach, w których można ją założyć, pole prędkości wiatru można wyznaczyć z pola ciśnienia. Znając więc ewolucję pola ciśnienia można, korzystając z kwazihydrostatyczności i kwazigeostroficzności wielkoskalowych ruchów powietrza, wyznaczyć w przybliżeniu ewolucję wielkoskalowego pola temperatury i wiatru. Stąd wynika znane powszechnie znaczenie ciśnienia w prognozie pogody.

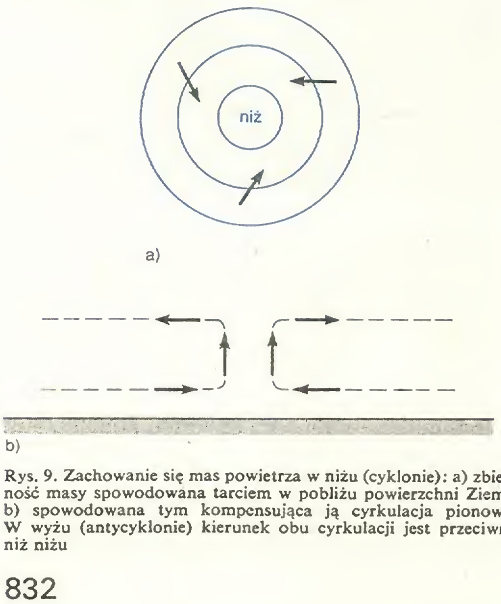
cyklony i antycyklony

Charakterystycznymi układami wielkoskalowej cyrkulacji poziomej są niże i wyże baryczne zwane też cyklonami i antycyklonami; są to układy ciśnienia z izobarami w postaci krzywych zamkniętych z minimum lub maksimum ciśnienia w środku. Wobec kwazigeostroficzności ruchu cyrkulacja w niżach odbywa się niemal dokładnie wzdłuż izobar, przeciwnie do ruchu wskazówek zegara (na półkuli północnej), w wyżach zaś — zgodnie z tym ruchem.

W planetarnej warstwie granicznej (pierwsze 1–2 km) siły tarcia turbulencyjnego nie mogą już być pominięte wobec siły Coriolisa. Pozioma kwazirównowaga sił gradientu ciśnienia, siły Coriolisa i siły tarcia (która z reguły ma składową skierowaną przeciwnie do kierunku wiatru) prowadzi do odchylenia wektora prędkości w tej warstwie w kierunku niższych ciśnień, tzn. do środka układu w niżu, na zewnątrz zaś układu — w wyżu (rys. 8). Spowodowany tym napływ względnie odpływ masy powietrza od centrum układu musi być skompensowany ruchem pionowym (rys. 9), jeżeli układ ma się utrzymać przez dłuższy czas.



Rys. 8. Równowaga geostroficzna z tarciem. Przy istnieniu równowagi siły tarcia gradientu ciśnienia i siły Coriolisa wiatr jest skręcony w kierunku mniejszych ciśnień



Rys. 9. Zachowanie się mas powietrza w niżu (cyklonie): a) zbieżność masy spowodowana tarciem w pobliżu powierzchni Ziemi; b) spowodowana tym kompensująca ją cyrkulacja pionowa. W wyżu (antycyklonie) kierunek obu cyrkulacji jest przeciwny niż niżu

Wynikiem tego efektu są powolne wielkoskalowe ruchy powietrza wstępujące — w niżach i zstępujące — w wyżach. Ma to ogromne znaczenie dla kształtowania pogody w tych układach, ponieważ ruchy wstępujące sprzyjają powstawaniu zachmurzenia i opadów, natomiast ruchy zstępujące powodują ich zanik. Stąd związek niżu z tzw. złą pogodą, wyżu zaś — z dobrą.

Ogólna cyrkulacja atmosfery

Ogólną cyrkulację atmosfery możemy rozpatrywać jako nakładające się na siebie ruchy różnych skal, tworzące całą hierarchię od skali ogólnoplanetarnej do mikroturbulencji, w której energia kinetyczna ulega dyssypacji (zamianie na ciepło) w wyniku występowania lepkości molekularnej. W węższym sensie, przez ogólną cyrkulację atmosfery rozumie się pierwsze elementy tej hierarchii, tzn. pola głównych parametrów mechanicznych i termodynamicznych uśrednione w czasie i wzdłuż równoleżników — a więc zależne jedynie od szerokości geograficznej i wysokości. Schemat tak rozumianej ogólnej cyrkulacji atmosfery w troposferze przedstawia rys. 10.

Cechą charakterystyczną tej cyrkulacji jest występowanie na każdej półkuli dwóch komórek cyrkulacji południkowo-pionowej: tzw. komórki Hadleya, w małych szerokościach geograficznych, oraz tzw. komórki Ferrela, która obejmuje szerokości średnie i duże. W obszarze komórek Hadleya południkowe składowe wiatru przy ziemi są skierowane ku ich wspólnej granicy wewnątrz tzw. międzyzwrotnikowej strefy zbieżności, natomiast składowa równoleżnikowa jest wschodnia (wiatry są odchylone na zachód). Są to znane powszechnie pasaty. Mechanizm działania komórki Hadleya polega na tym, że w pobliżu międzyzwrotnikowej strefy zbieżności troposfera jest intensywnie nagrzewana (głównie w wyniku uwalniania ciepła utajonego kondensacji w tym rejonie obfitującym w chmury). Nagrzewanie to wywołuje konwekcję termiczną w skali planetarnej (tzn. wznoszenie się ciepłego, lżejszego powietrza, a osiadanie chłodniejszego) w przekroju południkowo-pionowym i konwekcja ta jest zasadniczym powodem cyrkulacji w tej komórce. Wschodnia składowa równoleżnikowa przy powierzchni Ziemi jest wynikiem działania siły Coriolisa.

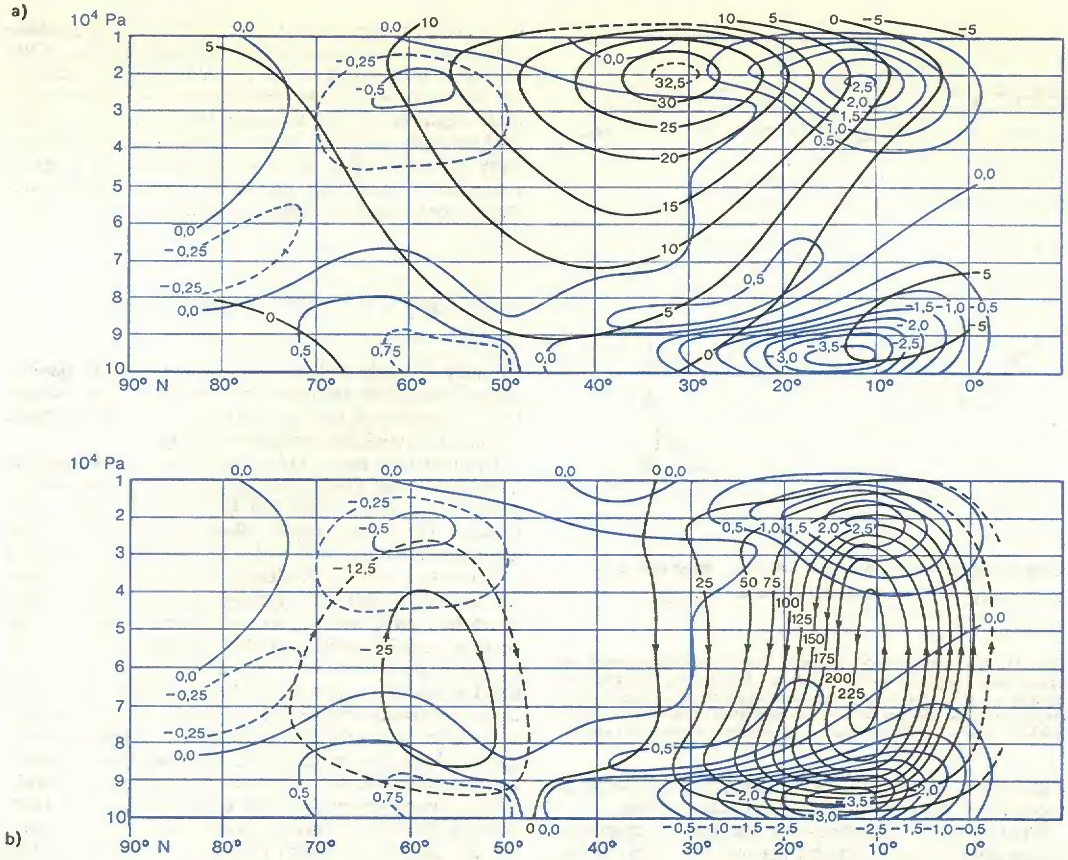
komórka Hadleya

komórka Ferrela

Bardziej skomplikowany jest mechanizm cyrkulacyjny komórki Ferrela. Obecność wiatrów wschodnich w komórce Hadleya wymusza w pewnym sensie istnienie strefy wiatrów zachodnich w innym obszarze, gdyż w przeciwnym razie moment pędu układu Ziemia–atmosfera (połączonego siłami tarcia) nie mógłby być zachowany i dlatego w komórce Ferrela dominują wiatry o zachodniej składowej równoleżnikowej; w związku z tym obszar tej komórki nazywa się także strefą wiatrów zachodnich. Wobec istnienia równowagi kwazigeostroficznej strefę wiatrów zachodnich musi charakteryzować spadek ciśnienia w stronę biegunów, tarcie zaś w pobliżu powierzchni Ziemi powoduje pojawienie się składowej dobiegunowej. Tak więc, w komórce Ferrela cyrkulacja południkowo-pionowa nie jest elementem pierwotnym całej cyrkulacji w tej komórce, lecz wtórnym wynikającym z istniejącej cyrkulacji równoleżnikowej. W odróżnieniu od komórki Hadleya, w której ciepłe powietrze unosi się do góry i następuje przemiana energii cieplnej i potencjalnej w kinetyczną, w komórce Ferrela następuje przemiana odwrotna. W tym sensie komórka Ferrela nie jest silnikiem cieplnym lecz pompą cieplną.

Źródła takich parametrów fizycznych atmosfery jak wilgotność, energia, moment pędu itp., wykazują znaczną zależność od szerokości geograficznej. Na przykład strefa okołorównikowa ma znaczną nad-

przykłady cyrkulacji



Rys. 10. Fragment ogólnej cyrkulacji atmosfery: a) Pionowo-południkowy przekrój przez troposferę półkuli północnej przedstawiający średnią cyrkulację południkową (kolor niebieski) i równoleżnikową (kolor czarny) w zimie. Prędkość cyrkulacji jest scharakteryzowana za pomocą izotach (linii jednakowej prędkości) składowych: zachodniej w m/s (czarny) i południowej (niebieski); b) Intensywność cyrkulacji południkowej jest scharakteryzowana przez zagęszczenie linii prądu przepływu masy (czarny). Odpowiadające jej prędkości przedstawione są w postaci izotach składowej południowej przeniesionej z rys. a. Wyraźnie zaznacza się intensywna cyrkulacja południkowa komórki Hadleya i słaba cyrkulacja komórki Ferrela

wyżkę parowania nad opadami — przeciwnie niż strefa umiarkowana, a moment siły tarcia między Ziemią i atmosferą maleje z szerokością geograficzną. Ponieważ średni bilans tych wielkości w dużej skali czasowej jest wszędzie bliski zeru, musi być wyrównywany południowym transportem przez cyrkulację atmosfery i oceanów.

O ile średnia cyrkulacja południkowo-pionowa w komórce Hadleya jest dość intensywna i w znacznym stopniu zapewnia ten transport w obszarze swego działania, o tyle cyrkulacja południkowo-pionowa w komórce Ferrela jest na to za słaba. Transport ten musi zapewniać nie cyrkulacja pionowa, lecz pozioma realizowana przez zależne od długości geograficznej i czasu zakłócenie cyrkulacji równoleżnikowej (tzw. „falowanie" strefy wiatrów zachodnich). Dotyczy to zwłaszcza momentu pędu, którego stały dopływ z niższych szerokości geograficznych umożliwia utrzymywanie stabilnej cyrkulacji zachodniej. Z tego punktu widzenia bardzo ważny jest rejon styku komórki Hadleya i Ferrela. W obszarze tym występuje duży poziomy gradient temperatury. W wielu miejscach jest on tak duży, że można mówić o skoku temperatury w danej skali — jest to rejon tzw. planetarnej strefy frontalnej, zwanej niekiedy frontem polarnym. Duże poziome gradienty temperatury powodują koncentrację energii osiągalnej (tzn. możliwej do zmiany w energię kinetyczną) w związku z tym jest to strefa szczególnie intensywnych zjawisk atmosferycznych. Jednym z nich jest tzw. prąd strumieniowy (*jet stream*) dość wąski (kilkaset kilometrów szerokości — przy tysiącach kilometrów długości) strumień bardzo silnych wiatrów występujący często w pobliżu tropopauzy. Na froncie polarnym pojawiają się też zafalowania, które często tracą stabilność przekształcając się w cyklony umiarkowanych szerokości (nie mylić z cyklonami tropikalnymi), czyli niże baryczne kształtujące w znacznym stopniu pogodę w naszej strefie geograficznej (il. 190, tabl. 49). Cyklony i towarzyszące im antycyklony (wyże) są głównym mechanizmem transportu południkowego w strefie wiatrów zachodnich.

front polarny

prąd strumieniowy

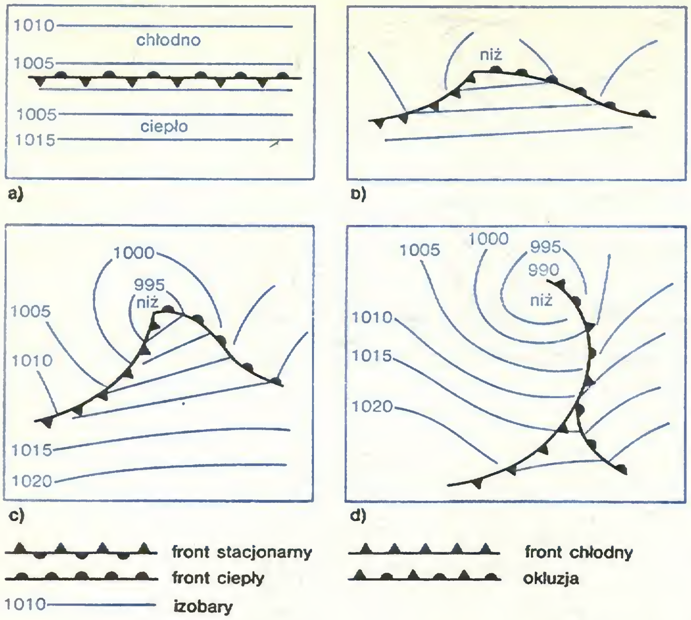
cyklony umiarkowanych szerokości

W klasycznym modelu niżu atmosferycznego wyróżnia się tzw. wycinek chłodny i wycinek ciepły oddzielone od siebie tzw. frontami — ciepłym i chłodnym, które łączą się w centrum cyklonu (rys. 11). Energię swą cyklon czerpie głównie z obniżania się środka ciężkości w wyniku wypierania ku górze lżejszego powietrza wycinka ciepłego przez cięższe powietrze wycinka chłodnego. Na powierzchni Ziemi proces ten prowadzi do połączenia się frontów — zjawisko to nazywa się okluzją. Intensywne ruchy pionowe w rejonie frontów i występujące na nich nieciągłości parametrów meteorologicznych są przyczyną wielu spektakularnych zjawisk meteorologicznych mniejszej skali — głównie chmur i opadów.

fronty atmosferyczne

Niezależnie od wędrownych niżów i wyżów, symetryczną równoleżnikową strukturę komórek Hadleya i Ferrela zakłócają mniej więcej stacjonarne zaburzenia związane z rozmieszczeniem kontynentów i oceanów oraz wielkich łańcuchów górskich.

Szczegóły omówionej wyżej struktury cyrkulacyjnej mają wyraźny bieg roczny — międzyzwrotnikowa strefa zbieżności przemieszcza się wraz z rocznym, pozornym ruchem Słońca na sferze niebieskiej; ponadto w związku z większymi poziomymi gradientami temperatury intensywność zjawisk cyrkulacyjnych jest większa na półkuli zimowej niż na letniej.



Rys. 11. Klasyczny model rozwoju niżu w umiarkowanych szerokościach geograficznych, a) fala na froncie, b) fala traci stabilność — jej amplituda rośnie, zaznacza się wyraźnie wycinek ciepły, c) niż w pełnym rozwoju — początek okludowania, d) niż zokludowany — cieplejsze powietrze zostaje wyparte do góry

Szczególne osobliwości wykazuje też cyrkulacja atmosfery w obszarach objętych nocą polarną.

Brak miejsca uniemożliwia bardziej szczegółowe omawianie całego łańcucha procesów o coraz mniejszych skalach — cyrkulacji lokalnych, bryz, zespołów chmur burzowych czy wreszcie drobnoskalowej turbulencji, które uczestnicząc w wyrównywaniu lokalnych bilansów różnych wielkości fizycznych, tworzą łącznie to, co się potocznie nazywa pogodą.

wpływ oceanów

Jakkolwiek cyrkulacja w troposferze i dolnej stratosferze stanowi zasadniczy mechanizm transportu (zwłaszcza poziomego) energii i innych wielkości fizycznych, nie należy pomijać jej związków z cyrkulacją oceaniczną (→ Fizyka morza) oraz cyrkulacją w wyższych warstwach atmosfery. Związek cyrkulacji troposferycznej z oceaniczną ma charakter sprzężenia zwrotnego — prądy morskie transportujące znaczne ilości ciepła, które — oddawane atmosferze — modyfikuje jej cyrkulację, same są w znacznym stopniu wynikiem działania wiatru. Ze względu na wielką bezwładność mechaniczną i cieplną wody, wpływ oceanu na temperaturę atmosfery jest znacznie większy niż wpływy atmosferyczne na ocean. W związku z tym wpływ oceanu daje się stwierdzić w szerokim przedziale skal ruchów atmosferycznych, podczas gdy wpływ atmosfery na ruchy oceaniczne — przede wszystkim w procesach o bardzo długim czasie trwania, liczonym w miesiącach i latach (jeżeli pominąć wzbudzenie fal wiatrowych i lokalne spiętrzenia wody).

cyrkulacja górnej stratosfery i mezosfery

Cyrkulacja górnej stratosfery i mezosfery jest stosunkowo słabo zbadana, ponieważ sondowania rakietowe — jedyne kompletne źródło informacji o niej — są bardzo kosztowne. Wpływ tej cyrkulacji na zjawiska troposferyczne jest kwestią dyskusyjną; wobec bardzo małej masy powietrza zawartej w tych częściach atmosfery, bezpośrednie oddziaływanie energetyczne na troposferę jest znikome. Górna stratosfera stwarza jednak warunki brzegowe cyrkulacji w niżej leżących warstwach atmosfery i może w ten sposób wpływać na procesy troposferyczne.

Cyrkulacja wysokich warstw atmosfery wykazuje silny wpływ rytmów słonecznych — dobowego, rocznego, jedenastoletniego — oraz nieregularnych zmian aktywności słonecznej, co wiąże się z tym, że głównym źródłem energii w tych warstwach jest pochłanianie promieniowania krótkofalowego Słońca. Cyrkulacje te odznaczają się też pewnymi osobliwościami, np. kierunki wiatru w stratosferze tropikalnej mają nieco zagadkowy cykl trwający ok. 26 miesięcy. Inną osobliwością są tzw. nagłe ogrzania dolnej stratosfery (o kilkanaście K w ciągu kilkunastu godzin) związane z kompresją powietrza przy falowych ruchach pionowych o dużej amplitudzie.

Chmury i opady atmosferyczne

Chmury i opady należą do najważniejszych zjawisk atmosferycznych zarówno ze względu na swe bezpośrednie znaczenie dla człowieka jak i na powiązania z innymi zjawiskami meteorologicznymi.

Chmura (lub mgła, która jest chmurą stykającą się z powierzchnią Ziemi) jest zbiorowiskiem mikroskopijnych kropelek wody lub kryształków lodu o typowych średnicach rzędu 10–30 μm, których prędkość spadania jest tak mała, że zawsze występująca w powietrzu mikroturbulencja nie pozwala im opaść na ziemię. Powstanie chmury wymaga przesycenia atmosfery parą wodną, co następuje najczęściej jako rezultat ochłodzenia wskutek adiabatycznego rozprężania powietrza w ruchach wstępujących — rzadziej w wyniku wypromieniowania ciepła lub w wyniku mieszania się dwóch nienasyconych mas powietrza o różnych temperaturach i wilgotnościach. Ze względu na mechanizm powstawania, chmury można podzielić na: warstwowe, kłębiaste (konwekcyjne), warstwowo-kłębiaste oraz orograficzne (górskie), jakkolwiek wobec wielkiej różnorodności chmur pewne rzadkie ich formy mogą się znaleźć poza tym podziałem.

Używana w służbie meteorologicznej międzynarodowa klasyfikacja chmur przyjmuje oparty na morfologii podział na rodzaje, gatunki i odmiany. Definicje podstawowych rodzajów chmur wg Międzynarodowego Atlasu Chmur podane wraz z nazwami łacińskimi, polskimi oraz skrótami międzynarodowymi są w tabeli oraz ukazane na il. 193 (tabl. 50).

chmury warstwowe

Pionowe wstępujące ruchy powietrza o wielkiej skali (setki — tysiące kilometrów) występujące przeważnie w niżach i w pobliżu frontów atmosferycznych są główną, choć nie jedyną, przyczyną powstawania chmur warstwowych, których rozciągłość pozioma jest dziesiątki i setki razy większa niż ich grubość. Do chmur warstwowych należą rodzaje St, Ns, As, Cs.

chmury konwekcyjne

Chmury konwekcyjne (rodzaje: Cu i Cb), których rozciągłość pionowa i pozioma są zbliżone, powstają w wyniku konwekcji o skali od setek metrów do dziesiątków kilometrów. Konwekcja ta jest wywołana przede wszystkim chwiejnością równowagi hydrostatycznej atmosfery, która się przejawia wówczas, gdy wznosząca się masa powietrza pozostaje lżejsza niż jej względnie nieruchome otoczenie. Ponieważ głównym czynnikiem, który o tym decyduje, jest różnica temperatur (powietrze cieplejsze jest zazwyczaj lżejsze od chłodniejszego), a temperatura powietrza wznoszącego się maleje wraz z wysokością na skutek rozprężania adiabatycznego o ok. 10 K/km, istnienie w powietrzu gradientu pionowego temperatury większego niż 10 K/km oznacza stan równowagi chwiejnej. W razie wystąpienia kondensacji wydzielane ciepło utajone zmniejsza tę krytyczną wartość niekiedy nawet do 5 K/km. Konwekcja może się także rozwijać w wyniku chwiejności dynamicznej spowodowanej niejednorodnością przestrzenną pola wiatru, w wyniku niejednorodności pola wilgotności oraz wzajemnego współdziałania wszystkich tych efektów. Silnie rozbudowane chmury konwekcyjne mogą sięgać nawet do stratosfery. Prędkość występujących w nich ruchów pionowych może przekraczać 50 m/s. Chmu-

Definicje podstawowych rodzajów chmur

Nazwa[a]	Opis
Cirrus (*Ci*, pierzaste)	chmury w kształcie oddzielnych, białych, delikatnych włókien bądź białych lub przeważnie białych ławic, czy też wąskich pasm. Chmury te mają włóknisty wygląd lub jedwabisty połysk albo obie te cechy jednocześnie
Cirrocumulus (*Cc*, kłębiasto-pierzaste)	cienka biała ławica, płat lub warstwa chmur bez cieni, złożona z bardzo małych członów w kształcie ziaren, zmarszczek itp., połączonych lub oddzielonych od siebie i ułożonych mniej lub bardziej regularnie. Większość członów ma pozorną szerokość mniejszą od jednego stopnia
Cirrostratus (*Cs*, warstwowo-pierzaste)	przejrzysta, biaława zasłona z chmur o włóknistym lub gładkim wyglądzie, pokrywająca niebo całkowicie lub częściowo i zwykle powodująca występowanie zjawisk halo
Altocumulus (*Ac*, średnie kłębiaste)	biała, szara bądź częściowo biała, częściowo szara ławica lub warstwa chmur, wykazująca na ogół cienie i złożona z płatów, zaokrąglonych brył, walców itp., połączonych ze sobą lub oddzielonych od siebie, niekiedy o wyglądzie częściowo włóknistym lub rozmytym. Pozorna szerokość większości regularnie ułożonych małych członów chmury zawiera się zwykle w granicach od jednego do pięciu stopni.
Altostratus (*As*, średnie warstwowe)	płat lub warstwa chmur szarawych bądź niebieskawych, o wyglądzie prążkowanym, włóknistym lub jednolitym, pokrywająca niebo całkowicie lub częściowo i miejscami tak cienka, że słońce jest widoczne jak przez matowe szkło. Chmura Altostratus nie powoduje występowania zjawisk halo
Nimbostratus (*Ns*, warstwowe deszczowe)	szara warstwa chmur, często ciemna o wyglądzie rozmytym wskutek mniej lub bardziej ciągłego opadu deszczu lub śniegu, w większości przypadków dochodzącego do ziemi. Chmura ta jest wszędzie tak gruba, że całkowicie przesłania Słońce. Poniżej tej warstwy często występują niskie, postrzępione chmury, które mogą być z nią połączone lub od niej oddzielone
Stratocumulus (*Sc*, kłębiasto-warstwowe)	Szara, biaława bądź częściowo szara, częściowo biaława ławica, płat lub warstwa chmur, posiadająca prawie zawsze ciemne części; złożona z zaokrąglonych brył, walców itp., połączonych ze sobą lub oddzielonych od siebie i nie posiadających wyglądu włóknistego. Większość regularnie ułożonych małych członów chmury ma pozorną szerokość większą od pięciu stopni
Stratus (*St*, niskie warstwowe)	na ogół szara warstwa chmur o dość jednolitej podstawie, mogąca dać opad mżawki, słupków lodowych lub śniegu ziarnistego. Jeżeli Słońce jest widoczne przez chmurę, jego zarys jest wyraźny. Chmura Stratus nie powoduje występowania zjawisk halo, z wyjątkiem być może przypadków, gdy temperatura powietrza jest bardzo niska. Chmura Stratus występuje niekiedy w postaci postrzępionych ławic
Cumulus (*Cu*, kłębiaste)	oddzielne na ogół gęste chmury, o ostrych zarysach, rozwijające się w kierunku pionowym w kształcie pagórków, kopuł lub wież, których górna, pączkująca część przypomina często kalafior. Oświetlone przez Słońce części tych chmur są przeważnie lśniąco białe. Podstawa ich jest stosunkowo ciemna i prawie pozioma. Czasami chmury Cumulus są postrzępione
Cumulonibus (*Cb*, kłębiaste deszczowe)	potężna, gęsta chmura o dużej pionowej rozciągłości w kształcie góry lub wielkich wież. Przynajmniej część jej wierzchołka jest zazwyczaj gładka, włóknista lub prążkowana i prawie zawsze spłaszczona. Część ta rozpościera się często w kształcie kowadła lub rozległego pióropusza.

[a] Podana jest nazwa łacińska, skrót międzynarodowy i nazwa polska.

rom takim towarzyszą zwykle intensywne zjawiska elektryczne prowadzące do wyładowań atmosferycznych (chmury burzowe).

Warstwy o małym pionowym gradiencie temperatury (spadku mniejszym niż 5 K/km), a zwłaszcza warstwy inwersji, w których temperatura rośnie wraz z wysokością, hamują pionowy rozwój konwekcji i chmur konwekcyjnych. Chmury konwekcyjne zanikają zazwyczaj podczas pojawienia się w nich prądów zstępujących, w których powietrze ulega ogrzaniu i względnemu wysuszeniu. Czasem w wyniku rozpadu chmur konwekcyjnych pozostają mniej lub bardziej rozległe pola chmur warstwowych lub warstwowo-kłębiastych.

chmury warstwowo-kłębiaste

Chmury warstwowo-kłębiaste (rodzaje: Sc, Ac i Cc) powstają głównie na skutek ruchów falowych lub konwekcyjnych wewnątrz chmur warstwowych.

chmury pierzaste

Chmury tworzące się w bardzo niskich temperaturach ok. −40°C są zbudowane z kryształków lodu, co nadaje im charakterystyczny włóknisty wygląd. Nazywa się je niekiedy chmurami pierzastymi (rodzaje: Ci, Cs, Cc).

chmury orograficzne

Chmury orograficzne tworzą się w wyniku ruchów pionowych powietrza, wymuszanych przez opływ góry. Ich cechą charakterystyczną jest zwięzek z terenem — nie są one unoszone przez wiatr, lecz powietrze przepływa przez nie.

Zanik chmur warstwowych i warstwowo-kłębiastych powodowany jest najczęściej wielkoskalowymi ruchami zstępującymi, w wyniku których następuje ogrzewanie powietrza pod wpływem kompresji, lub mieszaniem z otoczeniem bezchmurnym. W pewnym stopniu może też mieć na to wpływ ogrzewanie promieniowaniem słonecznym, które chmury pochłaniają w szerokim przedziale widmowym.

układy chmur

Chmury, zwłaszcza konwekcyjne występują często w postaci układów o wyraźnej strukturze mezoskalowej (skala rzędu dziesiątków i setek kilometrów), których wykrycie stało się możliwe dzięki zdjęciom satelitarnym (il. 192, tabl. 49). Układy te powstają w wyniku stosunkowo powolnych pionowych ruchów mezoskalowych, które mogą lokalnie zmienić stopień chwiejności równowagi atmosfery. Czasem oddziaływanie między rozwijającą się chmurą i otoczeniem tworzy sprzężenie zwrotne, na skutek którego w pobliżu jednej chmury tworzą się intensywnie następne. Mechanizm ten można zaobserwować przy rozwoju chmur burzowych i ich zespołów, a także w procesach takich jak powstawanie cyklonów tropikalnych. Szczegóły mechanizmów wzajemnego oddziaływania chmur na siebie oraz na otoczenie są obecnie przedmiotem intensywnych badań i mają wielkie znaczenie dla wielu gałęzi fizyki atmosfery.

Poza omówionymi wyżej tzw. makrofizycznymi procesami zachodzącymi w chmurach istotną rolę odgrywają procesy mikrofizyczne związane ze zjawiskami zachodzącymi w pojedynczej cząstce chmurowej lub ich niewielkich zespołach.

kondensacyjne powstawanie cząstki chmurowej

Kondensacyjne powstawanie cząstki chmurowej jest procesem skomplikowanym, ponieważ warunkujące ją przesycenie pary wodnej zależy nie tylko od temperatury, ale także od kształtu i rozmiaru powstającej cząstki oraz ilości i rodzaju rozpuszczonych w niej substancji. Kondensacja pary wodnej w czystym swobodnym powietrzu wymaga przesycenia rzędu 500% (w stosunku do prężności pary nasyconej nad płaską powierzchnią czystej wody podawanej w tablicach). Kondensacja w atmosferze zachodzi na jądrach kondensacji, którymi są głównie cząstki aerozolu zawierające substancje rozpuszczalne w wodzie. Dzięki nim kondensacja w chmurach i mgłach może wystąpić przy wilgotności względnej nawet poniżej 100%. Ilość, wielkość i skład chemiczny jąder kondensacji określa w znacznym stopniu liczbę i rozmiary kropli, które mogą na nich powstać drogą kondensacji. Typowe średnice kropli uzyskiwane na drodze kondensacji nie przekraczają 30–35 μm.

koalescencja

Dalszy wzrost kropli zachodzi przede wszystkim wskutek koalescencji. Najbardziej typowa jest koalescencja grawitacyjna. Polega ona na tym, że krople różnych rozmiarów spadając z różnymi prędkościami zderzają się ze sobą i łączą.

Skomplikowane oddziaływania aerodynamiczne między opadającymi kroplami powodują, że krople o średnicach poniżej 40–50 μm nie są zdolne do efektywnej koalescencji z kroplami mniejszymi. Widać więc, że przy typowych rozmiarach kropli powstających w toku kondensacji koalescencja jest mało prawdopodobna. Dlatego wiele rodzajów chmur nie daje opadów. Zapoczątkowanie koalescencji, która

ciepły deszcz

może doprowadzić do powstania opadu, wymaga pojawienia się w chmurze pewnej liczby cząstek dużych o średnicach powyżej 40 µm. Mogą nimi być np. kropelki kondensujące na tzw. jądrach gigantach — stosunkowo dużych kryształach soli morskiej lub innych substancji, które w pewnej liczbie mogą się znaleźć w powietrzu. Taki mechanizm powstawania opadu jest znany pod nazwą mechanizmu „ciepłego deszczu".

mechanizm trójfazowy

Często spotykanym mechanizmem kondensacji jest tzw. mechanizm lub proces trójfazowy, zw. inaczej mechanizmem Findeisena–Bergerona, działający wówczas, gdy chmura jest mieszaniną przechłodzonych kropelek wody i kryształów lodu. Przesycenie pary nad lodem jest większe niż nad wodą, wobec czego w takiej mieszaninie kryształy lodu gwałtownie rosną kondensacyjnie kosztem kropelek wody i mogą szybko osiągać rozmiary umożliwiające efektywną koalescencję. Pojawienie się znaczniejszej liczby kryształków lodu w chmurach następuje zwykle przy temperaturze poniżej −12°C. Wynika to stąd, że małe kropelki wody ulegają silnemu przechłodzeniu (do ok. −40°C), o ile nie zetkną się z kryształem lodu lub jakąś inną substancją, która może zadziałać jako tzw. jądro zamarzania. Obecne w aerozolu atmosferycznym typowe jądra zamarzania zaczynają działać właśnie w temperaturze poniżej −12°C. Powstający w wyniku procesu trójfazowego opad stały może stopnieć i dotrzeć do ziemi jako deszcz lub — gdy temperatura nie jest dość wysoka — jako śnieg, krupa (grad miękki) lub grad. Rodzaj i wielkość cząstek opadu zależy od bilansu ciepła i wody w procesie ich wzrostu oraz od jego długotrwałości. Silne prądy pionowe w dużych chmurach kłębiastych nie pozwalają na wypadanie z chmury cząstek zbyt małych — stąd charakterystyczne dla tych chmur opady przelotne często zawierają cząstki o znacznych rozmiarach.

W umiarkowanych strefach klimatycznych ok. 90% opadów powstaje w procesie trójfazowym. W strefie tropikalnej ok. 50% opadów to tzw. ciepły deszcz. Poza wspomnianymi tu mechanizmami powstawania cząstek dostatecznie dużych, by zapoczątkować skuteczną koalescencję grawitacyjną, działają jeszcze czasem inne mechanizmy, których natura nie jest obecnie dostatecznie wyjaśniona — być może odgrywają w nich pewną rolę efekty elektryczne lub mikroturbulencja.

sztuczna modyfikacja opadów

Zarówno mechanizm ciepłego deszczu jak i trójfazowy może być stosunkowo łatwo modyfikowany przez wprowadzenie do chmur sztucznych jąder kondensacji lub zamarzania. Na tej zasadzie opierają się próby modyfikacji chmur i opadów zmierzające do wywoływania sztucznego deszczu lub zapobiegania szkodom gradowym. Pomimo dużych wysiłków i niewątpliwych sukcesów w indywidualnych przypadkach, powszechne stosowanie tej techniki nie jest jeszcze możliwe; procesy fizyczne w chmurach nie są na tyle dobrze poznane, by z dostateczną pewnością można było liczyć na pożądane wyniki.

Główne problemy i kierunki rozwoju współczesnej fizyki atmosfery

Główne problemy i kierunki współczesnej fizyki atmosfery są wyznaczane przede wszystkim przez społeczne i gospodarcze potrzeby współczesności. Badania w tej dziedzinie są bowiem bardzo kosztowne i wymagają zorganizowanej na szeroką skalę współpracy międzynarodowej. Wydaje się, że na obecnym etapie najważniejszymi są następujące problemy: prognozowanie pogody, ewolucja klimatu oraz modyfikacja pogody i klimatu (celowa lub niezamierzona) w wyniku działalności człowieka.

Prognozy pogody

Prognozy pogody mogą być subiektywne (oparte głównie na subiektywnej ocenie sytuacji i doświadczeniu meteorologa) lub obiektywne (oparte na ustalonych regułach niezależnych od stosującego go meteorologa). Prognozy obiektywne mogą być prognozami statystycznymi, tj. mogą wykorzystywać empirycznie stwierdzone związki między różnymi zjawiskami meteorologicznymi, ustalającymi prawdopodobieństwa występowania ich w pewnych sekwencjach czasowych, lub deterministycznymi (czasem nazywa się je dynamicznymi), tj. określającymi przyszły stan atmosfery na podstawie znajomości stanów przeszłych oraz praw fizycznych rządzących ewolucją zjawisk atmosferycznych. Do połowy XX w. dominowały w praktyce prognozy subiektywne. Rozwój technik obliczeniowych i przetwarzania danych w drugiej połowie XX w. spowodował szybki rozwój metod dynamicznych i statystycznych. Obecnie stosuje się najczęściej metody stanowiące kombinacje tych trzech rodzajów prognoz. Ponieważ z metod fizyki atmosfery korzysta się najbardziej przy prognozach deterministycznych, więc nimi się głównie zajmiemy.

prognozy deterministyczne

Podstawą prognoz deterministycznych jest fakt, że znajomość pól ciśnienia, wiatru (dwie składowe poziome) i prędkości pionowej oraz temperatury i wilgotności pozwala, z wystarczającą dla celów praktycznych dokładnością, określić występujące zjawiska pogodowe. Problem sprowadza się więc do prognozy pól tych sześciu parametrów.

Wyrażone w postaci równań różniczkowych zasady zachowania pędu energii, masy powietrza, i masy wody dostarczają w zasadzie sześciu równań do wyznaczania sześciu niewiadomych pól. Aby je rozwiązać, należy znać warunki początkowe (tzn. stan niewiadomych pól w chwili uznanej za początkową), a następnie zastosować odpowiednią procedurę matematyczną w celu wyznaczenia stanu w chwili końcowej. Przy wykonywaniu tych działań w praktyce napotyka się niemało trudności. Przede wszystkim w równaniach występują siły i źródła energii, których postać matematyczną nie zawsze umiemy ustalić. Następnie rozpatrujemy zwykle pewną skalę zjawisk — oddziaływanie ze skalami leżącymi poniżej naszej zdolności rozdzielczej powoduje pojawienie się w równaniach dodatkowych wyrażeń o charakterze sił lub źródeł energii, które należy traktować jak „nowe" niewiadome i dla których należy ułożyć odrębne równania, wiążące je ze „starymi". Szukanie takich równań nazywa się parametryzacją procesów podskalowych. Wamaga ona przede wszystkim dobrego rozumienia strony fizycznej wchodzących w grę procesów; można stwierdzić, że znaczna część wysiłku badawczego, tak teoretycznego, jak i obserwacyjnego, współczesnej fizyki atmosfery wiąże się mniej lub bardziej bezpośrednio z próbami znalezienia dobrych metod parametryzacji procesów podskalowych w matematycznych modelach zjawisk atmosferycznych. Pomimo pewnych sukcesów na tym polu, problem jest daleki od rozwiązania i nie jest przy tym wcale pewne, czy rozwiązanie go z zadowalającą dokładnością jest w ogóle możliwe.

parametryzacja procesów podskalowych

warunki początkowe

Trudności nastręcza też formułowanie warunków początkowych, Źródłem ich są obserwacje meteorologiczne zawsze obarczone błędami pomiarowymi oraz błędami wynikającymi z interpretacji między rzadko rozmieszczonymi stacjami meteorologicznymi — niektóre rejony globu, np. oceany, okolice podbiegunowe i obszary słabo zamieszkałe, nie mają niemal zupełnie posterunków meteorologicznych, a zwłaszcza posterunków aerologicznych wykonujących pomiary w atmosferze swobodnej do wysokości ok. 30 km. Toteż duże nadzieje wiąże się z wykorzystaniem sztucznych satelitów. Np. satelitarne pomiary promieniowania podczerwonego atmosfery w różnych przedziałach widmowych pozwalają na wyznaczenie prze-

strzennych rozkładów temperatury i wilgotności, z których (korzystając z kwazihydrostatyczności i kwazigeostroficzności atmosfery) można w przybliżeniu wydedukować przestrzenne rozkłady pozostałych interesujących nas parametrów. Jednakże praktyczne wykorzystanie tej techniki wymaga jeszcze pokonania wielu trudności fizycznych i technicznych.

Oprócz powyższych występują też problemy natury matematycznej (luki w teorii równań różniczkowych i ich przybliżonego rozwiązywania) oraz technicznej związanej z szybkim wykonywaniem obliczeń i przetwarzaniem danych, ponieważ moce obliczeniowe nawet największych współczesnych komputerów nie są dostateczne dla rozwiązania trudniejszych zagadnień prognozy pogody z odpowiednią prędkością.

problem przewidywalności stanów atmosfery

Można by sądzić, że trudności prognozy deterministycznej są w gruncie rzeczy natury technicznej, tzn. supergęsta sieć bardzo dokładnych pomiarów pozwoliłaby uwzględniać procesy o bardzo małych skalach minimalizując wpływ efektów podskalowych oraz niedokładności danych początkowych, a supersprawne komputery pozwoliłyby w przyszłości pokonać trudności obliczeniowe. Eksperymenty numeryczne wykonane na różnych modelach atmosfery zdają się jednak sugerować, że istnieje także pewien teoretyczny kres dokładności i wyprzedzenia prognozy typu deterministycznego. Równania hydrodynamiki stosowane w prognozie są bardzo bowiem czułe na warunki początkowe, tzn. dwa rozwiązania wychodzące z mało różniących się stanów początkowych mogą po dość krótkim czasie bardzo się różnić, przy czym narastanie tych różnic jest tym szybsze, im mniejsze skale przestrzenne i czasowe uwzględniają rozwiązania. Efekt ten może bardzo ograniczyć, a nawet zniweczyć korzyści wynikające z wprowadzenia bardzo małych skal do równań i warunków początkowych. Jeżeli nie jest to własność samego modelu ale także realnej atmosfery, to można się obawiać, że działa tu jakby pewna zasada nieoznaczoności: im bardziej szczegółowa w sensie przestrzennym jest prognoza, tym na krótszy okres można ją opracować. Problem istnienia takiego teoretycznego kresu wyprzedzenia prognozy nosi nazwę problemu przewidywalności stanów atmosfery.

Przypuszcza się, że dla prognoz w skali synoptycznej kresem wyprzedzenia prognozy jest kilkanaście dni. Ponieważ obecnie prognozy te formułuje się na 2–3 doby, możliwości postępu, choć ograniczone, są jeszcze znaczne. Stworzenie sieci obserwacyjnej i środków obliczeniowych — wzajemnie dopasowanych, które pozwoliłyby realizować te zadania w sposób optymalny, jest celem wielkiego, wieloletniego, międzynarodowego programu badawczego znanego pod nazwą GARP (*Global Atmospheric Research Project*).

Prognozy długoterminowe — miesięczne czy sezonowe — są z konieczności odpowiednio mniej precyzyjne. Podejmowane obecnie próby opracowywania takich prognoz mają wciąż jeszcze dość słabe podstawy fizyczne.

Ewolucja klimatu i modyfikacja pogody

ewolucja klimatu

Klimat jest czynnikiem niezwykle ważnym w działalności człowieka, zwłaszcza w jego gospodarce żywnościowej i energetycznej. Choć zmiany klimatyczne w skali czasu życia jednego pokolenia są niewielkie, istnieją przekonywające dowody poważnej ewolucji klimatu w dłuższym czasie. Wyjaśnienie mechanizmów tej ewolucji jest jednym z głównych zadań współczesnej fizyki atmosfery. Zasadniczym czynnikiem powodującym przemiany klimatu są zmiany przestrzennego rozkładu bilansu cieplnego Ziemi i atmosfery. Przyczyny tych zmian mogą tkwić bądź w zmianach dopływu promieniowania słonecznego na skutek periodycznych zmian astronomicznych lub nieperiodycznych wahań aktywności słonecznej, bądź w zmianach własności cieplnych Ziemi i atmosfery (zmiany szaty roślinnej, zmiany składu atmosfery, przemieszczenia biegunów itp.).

stabilność klimatu

Szczególnie istotny jest problem stabilności klimatu — wyjaśnienie, czy pojawienie się jakiejś szczególnie silnej jednorazowej fluktuacji pogody nie wywoła trwałej i w dodatku stosunkowo szybkiej zmiany klimatu Ziemi. Występuje tu pewna analogia z równowagą cegły stojącej na płaskim stole. Leżąc na którejkolwiek ze swych ścian, cegła znajduje się w stanie równowagi i na małe wychylenie zareaguje powrotem do położenia pierwotnego. Jednak przy dostatecznie dużym wychyleniu stan równowagi może być osiągnięty już w innym położeniu. Istnieje wiele mechanizmów atmosferycznych, do których teoretycznie ta analogia mogłaby się stosować. Na przykład, znaczne zwiększenie zawartości pary wodnej w powietrzu spowoduje podniesienie temperatury powierzchni Ziemi w wyniku spotęgowania efektu inspektowego, co wywoła dalsze zwiększone parowanie itd. (dodatnie sprzężenie zwrotne). Wzrost temperatury naruszy w końcu równowagę bilansu cieplnego, co doprowadzi do zmiany układu cyrkulacji przywracającej tę równowagę, lecz — być może — przy zmienionych przeciętnych wartościach temperatury, wilgotności i układu wiatrów. Może się jednak zdarzyć, że wzrost wilgotności spowoduje od razu np. wzrost opadów, który przywróci poprzedni stan równowagi (ujemne sprzężenie zwrotne). Powstaje pytanie, który z tych wariantów odpowiada rzeczywistości. Rozstrzygnięcie problemów tego rodzaju jest szczególnie istotne w perspektywie dość szybkich zmian w atmosferze i powierzchni Ziemi pod wpływem działalności człowieka. Jak wykazała tragiczna w skutkach susza w krajach Sahelu w latach 1969–1975, nawet niewielkie przemieszczenie stref klimatycznych może mieć lokalnie katastrofalne skutki.

W chwili obecnej największy niepokój budzą następujące możliwe procesy. W ostatnim stuleciu obserwuje się stały, stopniowy wzrost koncentracji CO_2 w powietrzu (rys. 12) w wyniku spalania paliw zawierających węgiel, oraz być może także na skutek zanieczyszczenia wód oceanicznych ropą naftową (pogorszenie warunków rozpuszczania CO_2 w wodzie oceanicznej) lub zmniejszenia fotosyntezy w skali światowej. Istnieje obawa, że dalszy wzrost koncentracji CO_2, który odgrywa istotną rolę w „efekcie inspektowym" może uruchomić dodatnie sprzężenie zwrotne omówione wyżej.

zwiększenie zawartości CO_2

zwiększenie koncentracji aerozoli

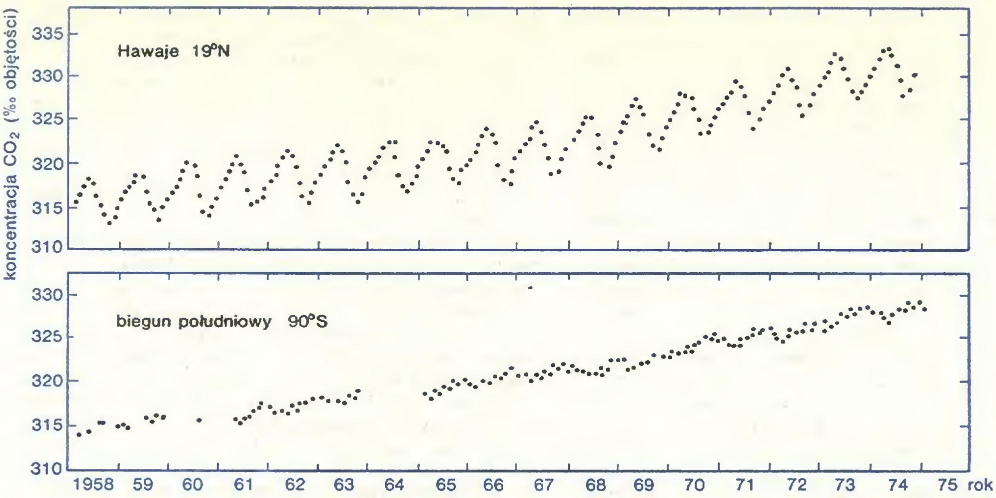
Zwiększona w wyniku procesów spalania koncentracja aerozoli może zmienić albedo Ziemi. Ponadto aerozole są często aktywnymi jądrami kondensacji i zamarzania, co prowadzi do zmian w zachmurzeniu i opadach, i co za tym idzie — do zmiany ilości wydzielanego ciepła kondensacji. Ponadto chmury i pokrywa śnieżna mają istotne znaczenie w bilansie radiacyjnym (wysokie albedo względem promieniowania słonecznego, silna emisja w podczerwieni). W rezultacie może dojść do ogólnej zmiany klimatu.

naruszenie równowagi ozonowej

W wyniku emisji związków chloru i tlenków azotu, które spełniają funkcję katalizatorów w rozpadzie ozonu może nastąpić naruszenie równowagi ozonowej w górnej stratosferze. Ozon ma duże znaczenie w bilansie cieplnym górnej stratosfery, który z kolei może mieć pośredni wpływ na cyrkulację troposferyczną; ponadto ozon pochłaniając aktywne biologicznie promieniowanie nadfioletowe ma znaczenie ochronne dla istot żywych. Jednocześnie jego bezwzględna zawartość w atmosferze jest tak mała, że już obecna aktywność przemysłowa człowieka, a zwłaszcza loty samolotów stratosferycznych (wydzielanie pary wodnej i tlenków azotu w spalinach) mogą spowodować znaczące zmiany jego koncentracji.

modelowanie klimatu

Odpowiedzi na pytania związane z ewolucją i stabilnością klimatu względem różnego rodzaju zakłóceń szuka się przeprowadzając eksperymenty obliczeniowe na modelach matematycznych atmosfery. Są to modele w zasadzie podobne do stosowanych w celach



Rys. 12. Zmiany koncentracji CO_2 w okresie 1958–75 r.

prognozy deterministycznej. Ponieważ czas obliczeń nie ma w tym wypadku znaczenia podstawowego, można stosować znacznie bogatsze układy równań uwzględniające bardziej złożone sprzężenia między procesami atmosferycznymi. Ponadto nie jest już istotne uzyskanie właściwej prognozy na konkretny termin, lecz odtworzenie statystycznych własności pogody w dłuższym przedziale czasu, wobec czego trudności omówione poprzednio w związku z problemem przewidywalności atmosfery odgrywają tu mniejszą rolę. Obecnie istnieje wiele tego rodzaju modeli, na których przeprowadza się liczne eksperymenty obliczeniowe. Nie dały one dotąd definitywnych odpowiedzi na postawione wyżej pytania, ponieważ nie wszystkie sprzężenia, które należałoby uwzględnić w realistycznych modelach atmosfery, są dostatecznie dobrze poznane. Jednakże, wyjaśniając rolę przynajmniej niektórych mechanizmów kształtowania klimatu, niewątpliwie uzyskanie tych odpowiedzi przybliżyły.

lokalne zmiany klimatu

Niezależnie od perspektyw globalnych zmian klimatu, intensywnie bada się zmiany lokalne powstające np. przez zakładanie wielkich zespołów miejskich lub kompleksów przemysłowych. W zmianach warunków klimatycznych ma znaczenie nie tylko pośredni wpływ na naturalne zjawiska atmosferyczne (np. własności cieplne asfaltu i betonu, emisja aerozoli), ale nawet bezpośrednie efekty energetyczne związane z wydzielaniem ciepła i pary wodnej w różnych procesach związanych z życiem i działaniem człowieka.

modyfikacja pogody

Poza omówioną wyżej niezamierzoną modyfikacją pogody, wiele uwagi poświęca się celowemu oddziaływaniu na zjawiska atmosferyczne. Kwestia ta miała wielu entuzjastów zwłaszcza w latach czterdziestych, gdy gwałtowny rozwój różnych gałęzi techniki budził czasem zbyt daleko idące nadzieje. W tym czasie powstało niemało projektów ulepszania klimatu czy też „poprawiania pogody". Najbardziej typowymi były projekty budowy tam na cieśninach morskich celem zmiany cyrkulacji oceanicznej i pośrednio atmosferycznej, topienie lodów polarnych przez zmianę ich albeda lub wielkie prace hydrotechniczno-irygacyjne (np. nawodnienie Sahary lub masowe oddziaływanie na chmury i opady za pomocą sztucznych jąder kondensacji i zamarzania). Proponowano nawet bezpośrednie oddziaływanie energetyczne za pomocą energii atomowej — co było raczej projektem fantastycznym wobec zupełnej dysproporcji między wydajnością dostępnych człowiekowi źródeł energii a energią wielkoskalowych procesów atmosferycznych. (Energia wyzwalana przy wybuchu bomby wodorowej o mocy 1 megatony TNT jest rzędu energii wyzwalanej przez niewielką burzę). Uświadomienie sobie w latach sześćdziesiątych niebezpieczeństw związanych z naruszeniem równowagi środowiska przyrodniczego przez uboczne skutki tego rodzaju poczynań technicznych spowodowało odłożenie takich projektów — nawet technologicznie realnych — jeżeli nie na zawsze, to w każdym razie do czasu znacznie lepszego zrozumienia i przewidywania wszystkich ich konsekwencji.

zapobieganie klęskom

Obecnie intensywne badania w dziedzinie modyfikacji pogody koncentrują się przede wszystkim na lokalnym oddziaływaniu na chmury i opady (metodami wcześniej opisanymi) w celu zapobiegania klęskom takim jak: gradobicie, tornado, huragany związane z cyklonami tropikalnymi lub w celu poprawienia bilansu wodnego na ograniczonych terytoriach. Sama technika oddziaływania na chmury jest obecnie praktycznie rzecz biorąc opanowana — niejasny jednak pozostaje problem, gdzie i kiedy należy ją stosować, aby osiągnąć pożądany rezultat, i w jakich warunkach geograficznych i klimatycznych można spodziewać się sukcesów. Reakcje różnych chmur na takie oddziaływanie są bowiem wciąż dość nieoczekiwane. Jednakże wielki postęp w dziedzinie fizyki chmur osiągnięty w ostatnich latach pozwala oczekiwać pozytywnych rezultatów również w zakresie operacyjnego oddziaływania na chmury i opady.

Obserwacje i pomiary w atmosferze

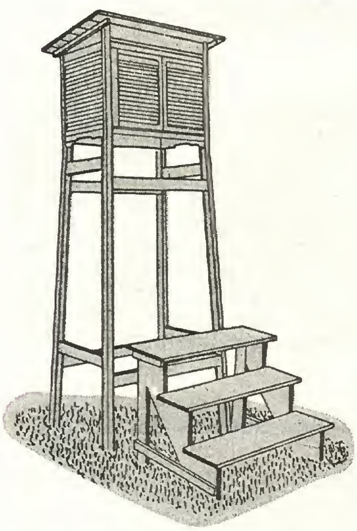
Obserwacje i pomiary zjawisk atmosferycznych — poza zjawiskami najmniejszych skal — muszą być wykonywane z wielu miejsc i następnie syntetyzowane w jeden spójny obraz. Oznacza to, że większość (są bowiem liczne i ważne wyjątki) obserwacji i pomiarów musi być wykonywana w ramach systemu, w skład którego, poza siecią punktów obserwacyjno-pomiarowych, wchodzi organizacja zbierania i przetwarzania wyników. Przyrządy pomiarowe są jedynie „końcówkami" tego systemu i stanowią w nim zazwyczaj najmniej skomplikowane pod względem technicznym elementy. Przy organizacji sieci konieczne jest zapewnienie reprezentatywności obserwacji i pomiarów (tj. by sposób ich wykonywania zapewniał uzyskanie potrzebnej informacji o badanej skali zjawiska) oraz ich porównywalności (tj. aby jednakowym wynikom pomiaru w różnych punktach odpowiadały jednakowe treści fizyczne).

obserwacje i pomiary standardowe

Rozróżniamy obserwacje i pomiary standardowe i specjalne. Obserwacje i pomiary standardowe są to obserwacje wykonywane w ramach międzynarodowej sieci meteorologicznej i są wykorzystywane przede wszystkim przez służby meteorologiczne wszystkich

krajów świata. Sieć ta składa się ze stacji meteorologicznych różnych rzędów o różnym zakresie obserwacji i pomiarów — od stacji mierzących jedynie opad na ziemi — do stacji aerologicznych, które za pomocą tzw. radiosond (il. 191, tabl. 49), wypuszczanych na balonach i transmitujących swe pomiary na ziemię drogą radiową, mierzą pionowe rozkłady temperatury, ciśnienia i wilgotności powietrza oraz wiatru do wysokości 20 km i wyżej. Typowe pomiary standardowe obejmują ciśnienie, temperaturę i wilgotność powietrza, wysokość opadu, zachmurzenie, elementy promieniowania słonecznego oraz ważniejsze zjawiska pogody. Sposób działania tych stacji jest ściśle określony i ujednolicony w ramach danej służby meteorologicznej — podejmowane są też wysiłki uzyskania daleko posuniętej jednolitości w skali światowej. Jest to jedno z zadań Światowej Organizacji Meteorologicznej.

Standardowe pomiary meteorologiczne muszą być przeprowadzane w sposób eliminujący wpływ czynników ubocznych na pomiar i jednakową metodą na wszystkich stacjach w ten sposób, by uzyskać reprezentatywną informację o tej skali zjawisk, która ma być badana (rys. 13).



Rys. 13. Klatka meteorologiczna zabezpiecza umieszczone w niej przyrządy przed przypadkowymi wpływami zewnętrznymi, ułatwiając uzyskanie pomiarów reprezentatywnych i porównywalnych

W wielu rejonach świata występują trudności z zapewnieniem odpowiednio kwalifikowanej obsługi dla stacji meteorologicznych. Można temu częściowo zaradzić instalując automatyczne stacje meteorologiczne (rys. 14), które drogą radiową lub kablową przekazują wyniki pomiarów podstawowych parametrów meteorologicznych. Do zbierania informacji tych wyników ze stacji położonych w obszarach niezamieszkałych, lub instalowanych na oceanach, można wykorzystać sztuczne satelity telekomunikacyjne. Światowa Organizacja Meteorologiczna przywiązuje wielką wagę do rozbudowy sieci stacji tego rodzaju.

obserwacje i pomiary specjalne

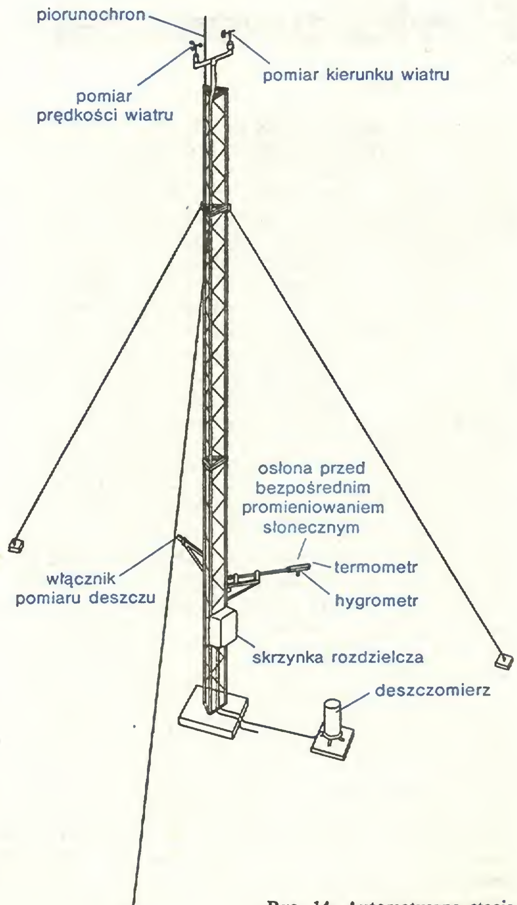
Obserwacje i pomiary specjalne są pomiarami wykonywanymi w celach badawczych i sposób ich przeprowadzania nie jest zazwyczaj regulowany przepisami Światowej Organizacji Meteorologicznej lub służby państwowej, choć sami naukowcy, którzy je prowadzą, współpracują na ogół z innymi zespołami i starają się zapewnić niezbędny stopień porównywalności pomiarów. Niektóre pomiary i obserwacje specjalne, zostają z czasem stopniowo przekształcone w pomiary standardowe.

Pewien przełom w systemie organizacji standardowych obserwacji meteorologicznych zdają się zapowiadać satelity meteorologiczne działające od 1961 r., ponieważ są one w stanie zbierać niektóre informacje o atmosferze w skali globalnej bez pośrednictwa sieci stacji meteorologicznych.

satelity meteorologiczne

Satelity meteorologiczne są albo geostacjonarne (ich prędkość kątowa jest identyczna z prędkością kątową Ziemi, tak że przez cały czas znajdują się nad tym samym punktem równika) albo mogą poruszać się po różnych orbitach. Zasadniczym ich zadaniem jest dostarczenie obrazów zachmurzenia (w świetle widzialnym i w podczerwieni), które jest ważnym źródłem informacji o różnych zjawiskach meteorologicznych. Ilustracja 194 z tabl. 51 przedstawia mozaikę zestawioną ze zdjęć wykonanych przez satelitę Essa 9 poruszającego się na orbicie polarnej. Dokonuje się również pomiarów promieniowania w wybranych wąskich przedziałach widmowych, w których emitują gazy atmosferyczne — przede wszystkim CO_2 i para wodna — na ich podstawie próbuje się wyznaczać pionowe rozkłady temperatury i wilgotności. Satelity mogą też uzyskiwać informacje z obszarów trudno dostępnych (pustynie, oceany), mające sieć meteorologiczną niedostatecznie rozwiniętą.

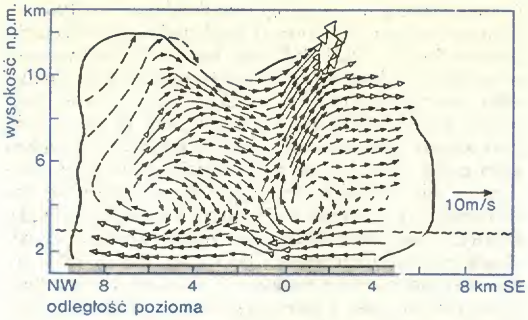
Wysiłki zmierzające do częściowego zastąpienia sieci meteorologicznej (a zwłaszcza bardzo kosztownej w



Rys. 14. Automatyczna stacja meteorologiczna

eksploatacji sieci aerologicznej) obserwacjami satelitarnymi natrafiają jednak na duże trudności. Wszystkie pomiary wykonywane za pomocą satelitów są bowiem pomiarami promieniowania elektromagnetycznego. Pomiary te nie dają się jak dotąd „przetłumaczyć" z dostateczną dokładnością na dane o temperaturze, wietrze itp. potrzebne służbie meteorologicznej. Prace w tej dziedzinie są jednak intensywnie kontynuowane (il. 195, tabl. 52).

Za pomocą satelitów uzyskano szereg ważnych informacji o charakterze globalnym (np. dokładne dane



Rys. 15. Rozkład prądów powietrza w chmurze burzowej wyznaczony za pomocą radaru dopplerowskiego. Strzałki przedstawiają wektory prędkości prądów powietrznych wg skali prądowej w m/s

o bilansie radiacyjnym) jak i mezoskalowym — w przedziale skal 10–1500 km, które są poniżej zdolności rozdzielczej standardowej sieci meteorologicznej.

zastosowanie radaru

Do badań meteorologicznych wykorzystuje się również radar. Bada się chmury a zwłaszcza opady, ponieważ duże cząstki opadowe silnie odbijają fale radarowe. Obserwacje radarowe pozwoliły m.in. zbadać strukturę cyklonów tropikalnych (il. 196, tabl. 52) oraz wewnętrzną budowę chmur burzowych, umożliwiają też stosunkowo dokładną ocenę ilości opadów na niewielkich terytoriach, co ma ogromne znaczenie np. dla odpowiedniej gospodarki zbiornikami wodnymi.

radary dopplerowskie

Wprowadzone niedawno do użytku w meteorologii radary dopplerowskie wyznaczające składowe prędkości namierzanych obiektów w kierunku radaru, pozwoliły zbadać rozkład prądów powietrza w chmurach burzowych. Wyznaczono go mierząc prędkości ruchu cząstek opadowych z dwóch punktów. Komputer przekształca następnie te informacje w obraz graficzny (rys. 16).

D. Blanchard *Od kropel deszczu do wulkanów*, Warszawa 1977; S. P. Chromow *Meteorologia i klimatologia*, Warszawa 1977; G. M. B. Dobson *Badania atmosfery*, Warszawa 1965; R.M. Goody *Atmosfery planet*, Warszawa 1978; T. Kopcewicz *Atmosfera i hydrosfera* w: *Ziemia*, Warszawa 1977.

Fizyka przestrzeni okołoziemskiej

Andrzej Wernik

przestrzeń okołoziemska

Przestrzenią okołoziemską nazywamy obszar wokół Ziemi, w którym dominującą rolę odgrywa pole grawitacyjne i magnetyczne Ziemi i który obejmuje jej gazową otoczkę — atmosferę. Pole grawitacyjne Ziemi podlega powszechnemu prawu grawitacji, a więc siła, z jaką oddziałuje na inne ciała, maleje odwrotnie proporcjonalnie do kwadratu odległości. Teoretycznie więc — sfera oddziaływania Ziemi przez pole grawitacyjne rozciąga się do nieskończoności. Jednak na bardzo dużych odległościach od Ziemi siła, z jaką jej pole grawitacyjne działa na ciało o małej masie, może być znacznie mniejsza od sił grawitacji innych masywnych ciał, takich jak Słońce czy planety. Przyjmujemy więc, że granica grawitacyjnego oddziaływania Ziemi znajduje się w odległości, w której pole grawitacyjne innych ciał niebieskich dominuje nad polem ziemskim. Warunek ten spełniony jest w odległości ok. $8 \cdot 10^6$ km od środka Ziemi. Jeżeli przyjmiemy, że pole magnetyczne Ziemi jest polem dipola, to musimy również przyjąć, że natężenie jego maleje odwrotnie proporcjonalnie do sześcianu odległości, a więc teoretycznie także nie ma granic. Praktycznie, ograniczone jest jednak do obszaru tzw. magnetosfery, która się rozciąga w kierunku Słońca na odległość równą ok. dziesięciu promieniom Ziemi, czyli $6 \cdot 10^4$ km.

magnetosfera

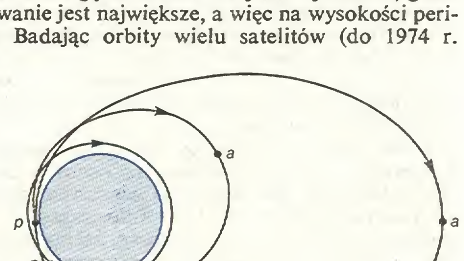
Oprócz pól grawitacyjnego i magnetycznego przestrzeń okołoziemską wypełnia materia — atmosfera, której struktura zależy od wysokości nad powierzchnią Ziemi (tabl. 7, il. 21). Od wysokości zależy jej gęstość, temperatura i skład chemiczny. Pod wpływem ultrafioletowego i rentgenowskiego promieniowania Słońca pewna liczba molekuł i atomów ulega dysocjacji i jonizacji. Na obie składowe: neutralną (molekuły, atomy) i zjonizowaną (jony, elektrony) działają różnorodne siły, które się zmieniają w czasie i przestrzeni, nieraz w bardzo skomplikowany sposób. Jednocześnie w atmosferze, podobnie jak w magnetosferze, występują cząstki naładowane pochodzenia słonecznego, meteory z Układu Słonecznego i promieniowanie kosmiczne powstające w Galaktyce. Zatem w przestrzeni okołoziemskiej zachodzi wiele skomplikowanych procesów fizycznych, warunkujących kompleks zjawisk obserwowanych na powierzchni Ziemi i na pokładach sztucznych satelitów. Wyjaśnienie wielu problemów fizyki przestrzeni okołoziemskiej zawdzięczamy postępom techniki satelitarnej i rakietowej oraz rozwojowi takich działów fizyki jak fizyka plazmy czy fizyka procesów zderzeń cząstek.

atmosfera

W dalszym ciągu przedstawimy pokrótce aktualny stan badań przestrzeni okołoziemskiej. Nie będziemy jednak rozważać struktury warstw atmosfery dostępnych dla pomiarów tradycyjnych z powierzchni Ziemi ani procesów w nich zachodzących. Jako poziom rozgraniczający przyjmiemy wysokość 300–350 km, na której koncentracja cząstek naładowanych jest maksymalna. Czytelników zainteresowanych fizyką warstw atmosfery leżących poniżej tej wysokości odsyłamy do artykułu Fizyka atmosfery. Tam też czytelnik znajdzie wyjaśnienie szeregu pojęć używanych w niniejszym artykule.

Atmosfera na wysokości powyżej 300 km

Gaz otaczający Ziemię, czyli atmosferę, charakteryzuje gęstość, ciśnienie, temperatura i skład chemiczny. Do wyznaczenia wartości tych parametrów na interesujących nas wysokościach najczęściej stosuje się metodę opartą na badaniu zmian kształtu orbit satelitów. Opiszemy pokrótce tę metodę. Satelita porusza się wokół Ziemi nie w próżni, lecz w atmosferze, która stawia pewien opór, hamuje satelitę. Oczywiście hamowanie będzie tym większe, im większy jest satelita i jego prędkość oraz im gęstsza jest atmosfera (wpływ hamowania w atmosferze na orbitę satelity obrazuje rys. 1). Pomiary zmian okresu obrotu satelity oraz wysokości apogeum i perigeum pozwalają na określenie gęstości atmosfery na wysokości, gdzie hamowanie jest największe, a więc na wysokości perigeum. Badając orbity wielu satelitów (do 1974 r.



Rys. 1. Zmiana kształtu orbity satelity na skutek hamującego działania atmosfery (*p* perigeum, *a* apogeum

zbadano przeszło 3000 satelitów) o różnych wysokościach perigeum wyznaczono gęstość na wysokościach od 120 do kilkuset kilometrów, zbadano również zmiany gęstości w ciągu dnia, jej zależność od pory roku i aktywności Słońca. Znając gęstość i skład chemiczny atmosfery na określonej wysokości i zakładając, że atmosfera zachowuje się jak gaz idealny, można wyznaczyć jej temperaturę. Rezultatem takich pomiarów było stwierdzenie, że gęstość atmosfery jest największa po południu, a najmniejsza w drugiej połowie nocy. Różnica między gęstością w dzień i w nocy rośnie ze wzrostem wysokości. Tak np. na wysokości 200 km gęstość w dzień jest 1,5 razy większa niż gęstość w nocy, ale na wysokości 600 km dzienna atmosfera jest ok. 8–10 razy gęstsza niż nocna. Podobnie w ciągu doby zmienia się temperatura, chociaż różnice nie przekraczają 50%. Na interesujących nas wysokościach, tzn. powyżej 300 km, temperatura w niewielkim stopniu zależy od wysokości; waha się ona w granicach od ok. 1000 K w nocy do 1300 K w dzień. Jednocześnie temperatura w znacznym stopniu zależy od aktywności Słońca, wykazuje więc zmiany z okresem rotacji Słońca (27 dni) i cyklem 11-letnim. W okresie dużej aktywności Słońca temperatura wysokich warstw atmosfery może osiągać w ciągu dnia 2000 K.

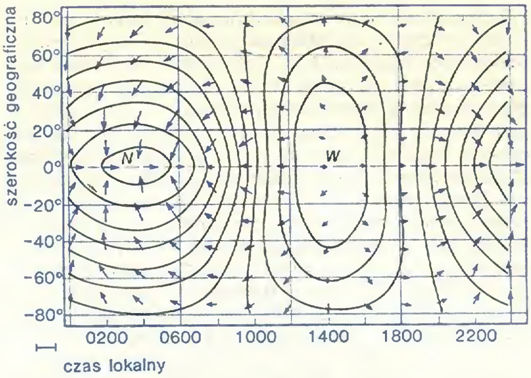
wiatr w atmosferze

Różnice gęstości i temperatury między dniem i nocą powodują powstawanie wiatrów w atmosferze. Jak mówiliśmy, w południe zarówno gęstość, jak i temperatura są największe, zatem duże też jest ciśnienie. Stosując nomenklaturę meteorologów, powiedzielibyśmy, że jest to obszar „wyżu". Jednocześnie, na nie oświetlonej części kuli ziemskiej panuje „niż". Wiemy, że gaz napływa z miejsc o wysokim ciśnieniu i stara się zapełnić miejsca o niskim ciśnieniu. Ten przepływ gazu jest właśnie wiatrem. Tak więc w górnej atmosferze ciągle wieją wiatry. Gdyby oprócz różnicy ciśnień nie istniały żadne inne czynniki wpływające na kierunek i siłę wiatru, wiatr w atmosferze miałby przed południem kierunek ku zachodowi, a po południu — ku wschodowi. W rzeczywistości ruch obrotowy Ziemi ze wschodu na zachód powoduje, że wiatry popołudniowe są silniejsze od przedpołudniowych. Istotnym czynnikiem wpływającym na prędkość i kierunek wiatru jest tzw. unoszenie jonami. Chodzi mianowicie o to, że obie składowe atmosfery, neutralna i zjonizowana, nie mogą się poruszać niezależnie od siebie; zderzają się ze sobą, przekazując sobie nawzajem pewną prędkość. Jednak ruch jonów nie jest swobodny. Mogą się one poruszać jedynie wzdłuż linii sił pola magnetycznego Ziemi. Na ruch cząstek naładowanych wpływa bowiem pole magnetyczne Ziemi, wskutek czego poruszają się one po spirali „nanizanej" na linie sił pola. Zderzając się ze sobą, jony nadają atomom prędkość w kierunku pola magnetycznego. Składowa prędkości jonów prostopadła do pola powoduje jedynie ruch periodyczny pojedynczych atomów wokół linii sił pola, lecz nie przesuwa mas gazu. W zderzeniach z atomami jony nadają atomom prędkość w kierunku pola magnetycznego. W rezultacie wiatr cząstek neutralnych odchyla się w kierunku pola magnetycznego. Układ wiatrów w atmosferze na wysokości 300 km przedstawiony jest na rys. 2. Prędkości wiatrów, zwłaszcza w nocy, są bardzo duże, przekraczają niekiedy 200 m/s. Oznacza to, że element gazu atmosferycznego może przewędrować tysiące kilometrów w ciągu jednego dnia.

unoszenie jonami

ogrzewanie się atmosfery

Co jest przyczyną ogrzewania atmosfery do temperatury wyższej niż 1000 K? Znaczne różnice temperatury między dniem i nocą oraz wyższe temperatury w dzień sugerują, że źródłem ogrzewania atmosfery jest Słońce. Na wysokościach powyżej 300 km głównym mechanizmem ogrzewania jest pochłanianie (absorpcja) ultrafioletowego promieniowania Słońca. Mechanizm przekazywania energii promieniowania cząsteczkom gazu jest w zasadzie prosty. Pochłaniane promieniowanie słoneczne wzbudza molekuły i atomy oraz wywołuje dysocjację molekuł i jonizację atomów.



Rys. 2. Układ wiatrów na wysokości 300 km wraz z układem izobar na tej wysokości (*W* wyż, *N* niż). Kierunek i długość strzałek obrazuje kierunek i wartość prędkości. Długość odcinka (u dołu z lewej strony) odpowiada prędkości 200 m/s

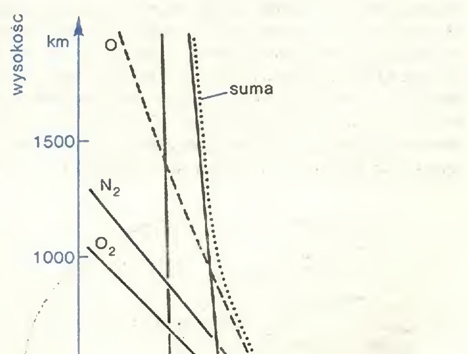
Wzbudzone, zdysocjowane i zjonizowane molekuły oraz atomy zderzają się z otaczającymi cząsteczkami gazu i przekazują im nadwyżkę energii uzyskaną w wyniku absorpcji promieniowania. Zwiększa się energia kinetyczna cząsteczek gazu, co jest równoważne ze wzrostem temperatury gazu.

Na dużych wysokościach gęstość gazu jest mała, a zatem ilość pochłanianego promieniowania jest niewielka i atmosfera powinna nagrzewać się nieznacznie. A przecież, jak mówiliśmy, temperatura na dużej wysokości, powiedzmy 1000 km, prawie się nie różni od temperatury na wysokości 100 km. Otóż okazuje się, że na dużych wysokościach atmosfera jest bardzo dobrym przewodnikiem ciepła. Tak więc ciepło nagromadzone w dolnych jej partiach rozprzestrzenia się szybko w górę i ogrzewa gaz do temperatury równej temperaturze na mniejszych wysokościach.

cząstki neutralne w atmosferze

Jakie cząstki neutralne mogą występować w górnych warstwach atmosfery? Odpowiedź na to pytanie daje rys. 3, na którym pokazano, jak się wraz z wysokością zmienia skład chemiczny atmosfery ziemskiej. Zauważmy, że powyżej 800 km dominują lekkie atomy tlenu, helu i wodoru, podczas gdy powyżej 1400 km zasadniczymi składnikami atmosfery są już tylko hel i wodór. Gęstość ich jest mała, a więc atom helu lub wodoru może przelecieć znaczną odległość, zanim się zderzy z innym atomem. Odległość ta, tzw. średnia droga swobodna, wynosi ok. 180 m na wysokości 400 km, ale na wysokości 1000 km wynosi aż 7 km.

składniki atmosfery



Rys. 3. Rozkład z wysokością ważniejszych składników chemicznych atmosfery (podana jest liczba cząstek w 1 cm³)

Między zderzeniami atom porusza się jedynie w polu ciężkości Ziemi. Zależnie od prędkości tory atomów mogą mieć różny kształt. Gdy prędkość atomu jest mała, porusza się on po elipsie i jest jakby mikroskopijnym satelitą Ziemi. Gdy prędkość atomu jest duża, jego energia kinetyczna może się okazać większa niż energia potencjalna, którą ma w polu grawitacyjnym Ziemi, a wówczas atom może uciec ze sfery przyciągania Ziemi i może się stać jednym z atomów przestrzeni międzyplanetarnej. Prędkość, przy której obie energie, kinetyczna i potencjalna, są równe, nazywa się prędkością ucieczki lub II prędkością kosmiczną i na wysokości 1000 km wynosi 10,4 km/s. Co najmniej taką właśnie prędkość musi mieć nie tylko atom, ale i rakieta kosmiczna, aby dotrzeć do innych planet.

II prędkość kosmiczna

Cząstki uciekają z całej atmosfery, nie tylko z wysokich jej warstw. Jednak na małych wysokościach bardzo małe jest prawdopodobieństwo znalezienia cząstki o wystarczająco dużej energii, ponieważ temperatura atmosfery jest zbyt niska. Warunki sprzyjające ucieczce atomów istnieją jedynie na wysokościach powyżej 600 km. Tę część atmosfery nazywamy egzosferą. Wskutek ciągłej ucieczki lekkich atomów z egzosfery mogłoby po pewnym czasie dojść do tego, że w atmosferze w ogóle nie byłoby atomów wodoru i helu. Istnieją jednak bogate źródła tych atomów. Na przykład helu dostarcza radioaktywny rozpad ciężkich atomów, takich jak uran, tor i rad, występujących w pewnej obfitości w skorupie ziemskiej. Tak utworzony hel dyfunduje w górne partie atmosfery. Źródłem wodoru w atmosferze jest para wodna i metan. Mimo iż występują one głównie w troposferze i stratosferze, mogą dyfundować do wysokości, na których ultrafioletowe promieniowanie Słońca powoduje ich dysocjację. Powstałe czyste atomy wodoru wobec małego ciężaru łatwo się rozprzestrzeniają do bardzo dużych wysokości.

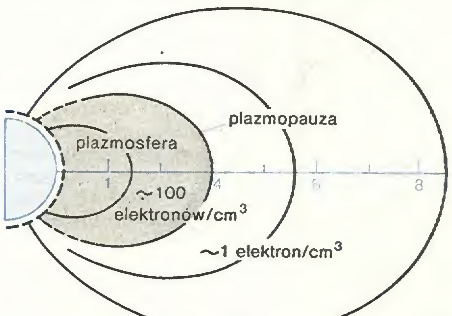
egzosfera

Znaczną część egzosfery stanowią cząstki naładowane — jony i elektrony. Ich koncentracja w egzosferze zmniejsza się znacznie wolniej niż koncentracja cząstek neutralnych. Mniej więcej powyżej 1000 km jest już w 1 cm^3 więcej jonów i elektronów niż cząstek neutralnych. Na tych wysokościach dominującymi cząstkami neutralnymi są atomy helu i wodoru, a większość jonów stanowią jony helu He^+ i wodoru H^+ (protony). Od pewnej wysokości koncentracja H^+ jest większa od koncentracji He^+. Obszar atmosfery powyżej tej wysokości (ok. 1200 km) nazywamy protonosferą. Koncentracja cząstek zjonizowanych w protonosferze maleje stopniowo, aby w pewnej odległości od Ziemi gwałtownie spaść z liczby 100 elektronów do 1 elektronu w 1 cm^3. Miejsce, w którym następuje skok koncentracji, nazywa się plazmopauzą. O przyczynie skoku koncentracji mówić będziemy w następującym rozdziale, poświęconym magnetosferze. Odległość do plazmopauzy jest taka jak odległość do linii siły ziemskiego pola magnetycznego, która przecina płaszczyznę równika w odległości równej czterem promieniom Ziemi ($2 \cdot 10^4$ km). Mówimy, że plazmopauza leży w odległości $L = 4$. Sytuację tę obrazuje rys. 4. Naszkicowano na nim kilka linii sił pola i zaznaczono obszar poniżej plazmopauzy, tzw. plazmosferę. Plazmosfera obejmuje oczywiście jonosferę, egzosferę i protonosferę. Jak widać z rysunku, wysokość plazmopauzy jest mniejsza na dużych szerokościach geomagnetycznych.

cząstki naładowane w atmosferze

protonosfera

plazmopauza



Rys. 4. Schematyczny przekrój przez magnetosferę w płaszczyźnie prostopadłej do równika

plazmosfera

Procesem fizycznym o bardzo istotnym znaczeniu w fizyce atmosfery jest dyfuzja cząstek naładowanych. Elektrony jako cząstki najlżejsze mają tendencję do „wypływania" na „powierzchnię" atmosfery. Jednak działa na nie przyciągająca siła elektrostatyczna dodatnio naładowanych jonów. Tak więc elektrony, „wypływając", jednocześnie wyciągają jony na większe wysokości. Powoduje to zmianę rozkładu wysokościowego cząstek naładowanych. Na dużych wysokościach jest ich więcej niż wówczas, gdyby proces tej dyfuzji nie odgrywał roli (np. poniżej 300 km). Dyfuzja cząstek naładowanych może się odbywać tylko wzdłuż linii pola magnetycznego. Stąd wnioskujemy, że w pobliżu równika magnetycznego, gdzie linie pola są równoległe do powierzchni Ziemi, dyfuzja nie zmienia rozkładu wysokościowego cząstek. Proces dyfuzji może powodować przepływ elektronów i jonów między jonosferą i protonosferą.

dyfuzja cząstek naładowanych

Magnetosfera

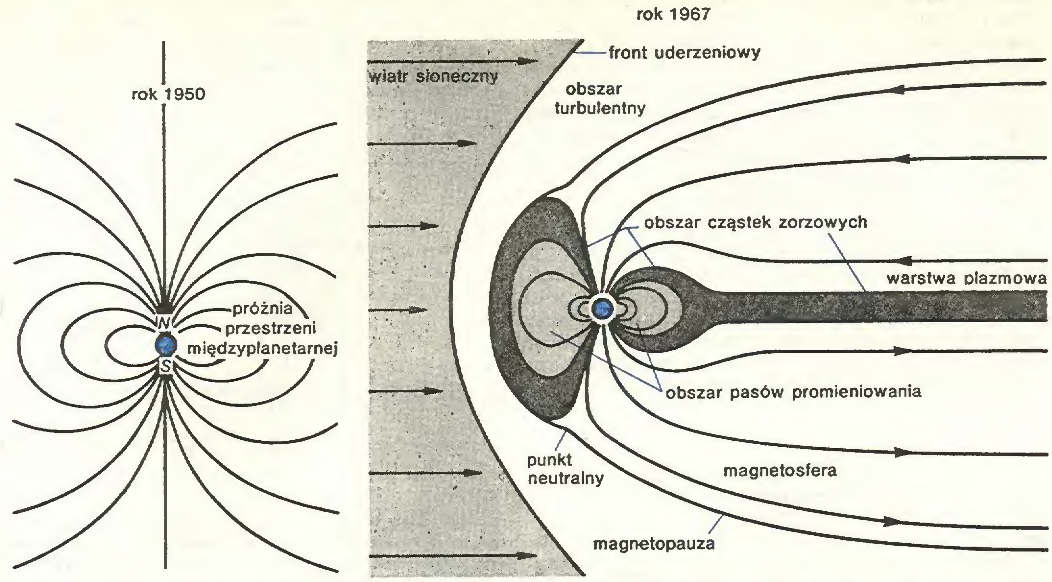
Do 1960 r. sądzono, że ziemskie pole magnetyczne mało się różni od pola dipola magnetycznego (→ Magnetyzm ziemski), jedynie w okresie dużych rozbłysków na Słońcu jego kształt ulega zmianie. Rozważania teoretyczne i pomiary wykonywane za pomocą aparatury umieszczonej na pokładach satelitów ziemskich zmusiły do zrewidowania pojęć o kształcie pola magnetycznego Ziemi. W rzeczywistości struktura pola magnetycznego wygląda najprawdopodobniej tak, jak to przedstawia prawa strona rys. 5. Do odległości równej kilku promieniom Ziemi pole geomagnetyczne ma rzeczywiście kształt taki, jak pole dipola. W większych odległościach na pole magnetyczne oddziałuje strumień cząstek (elektronów, protonów i jonów helu) wypływający ze Słońca. Strumień ten, napotykając przeszkodę w postaci pola magnetycznego Ziemi, „naciska" na nie i deformuje je. Ciśnienie wiatru słonecznego jest bardzo małe i wynosi ok. $2 \cdot 10^{-9}$ Pa. Polu magnetycznemu można również przypisać pewne ciśnienie związane z naprężeniem linii pola. Maleje ono szybko wraz z odległością. Odległość, na której jest ono równe ciśnieniu wiatru słonecznego, określa granicę pola magnetycznego Ziemi, tzw. magnetopauzę. Magnetopauza nie styka się bezpośrednio z wiatrem słonecznym, lecz oddzielona jest od niego warstwą, w której przepływ cząstek wiatru ma charakter turbulentny. Obszar wewnątrz magnetopauzy nazywamy magnetosferą. Od strony Słońca magnetopauza leży w odległości ok. 10 promieni Ziemi ($6 \cdot 10^4$ km). Po nocnej stronie Ziemi odległość do magnetopauzy jest rzędu kilkuset promieni ziemskich. Jest to tzw. ogon magnetosfery (rys. 5). W obszarze ogona, w płaszczyźnie równika magnetycznego, obserwuje się podwyższenie koncentracji cząstek naładowanych — warstwę plazmową. Pole magnetyczne w warstwie plazmowej jest bardzo słabe i na stronie północnej ma kierunek ku Ziemi, a na stronie południowej — od Ziemi. W płaszczyźnie równika natężenie pola jest równe zeru. Po dziennej stronie magnetopauzy, w tzw. punktach neutralnych, następuje rozdzielenie linii sił pola magnetycznego: część z nich biegnie przez płaszczyznę równika ku przeciwnemu biegunowi, część zaś ku ogonowi magnetosfery. Wewnątrz magnetosfery znajdują się pasy promieniowania, zwane także pasami Van Allena (od nazwiska uczonego, który je odkrył, mierząc na satelitach strumienie cząstek o dużych energiach). Umownie rozróżniamy dwa pasy: wewnętrzny, leżący na linii siły pola magnetycznego, która przecina równik w odległości 1,5 pro-

magnetopauza

ogon magnetosfery

pasy promieniowania

mienia Ziemi ($L \approx 1{,}5$) i zewnętrzny — w odległości $L \approx 3$–4). Należy jednak zauważyć, że granica między pasami nie jest ostro zaznaczona. Wewnętrzny pas



Rys. 5. Porównanie kształtu pola magnetycznego Ziemi bez wiatru słonecznego (z lewej strony) i z wiatrem słonecznym (z prawej strony) (przekrój płaszczyzną prostopadłą do równika geomagnetycznego, przechodzącą przez linię Słońce–Ziemia)

promieniowania składa się z protonów o energiach równych dziesiątkom i setkom MeV. W zewnętrznym pasie dominują elektrony o energiach od dziesiątków keV do kilku MeV. Koncentracja cząstek w zewnętrznym pasie jest bardzo mała, całkowita ich masa nie przekracza 15 kg. Strumienie elektronów w zewnętrznym pasie promieniowania silnie fluktuują, wskutek czego zmienia się natężenie pola magnetycznego na powierzchni Ziemi. Cząstki uwięzione są w pasach promieniowania i poruszają się po skomplikowanych torach (rys. 6), przy tym ruch jonów i ruch elektronów odbywają się we wzajemnie przeciwnych kierunkach.



Rys. 6. Tory cząstek w magnetosferze. Protony dryfują na zachód, elektrony na wschód

Przypomnijmy znany fakt, że ruch przewodnika w polu magnetycznym prowadzi do powstania pola elektrycznego prostopadłego do kierunku ruchu i kierunku pola. Natomiast przewodnik umieszczony w skrzyżowanych polach elektrycznym i magnetycznym zostaje wprawiony w ruch w kierunku prostopadłym do obu pól. Plazma w przestrzeni okołoziemskiej jest przewodnikiem w polu magnetycznym Ziemi. Jej ruchowi towarzyszyć więc będzie pole elektryczne i odwrotnie — pole elektryczne spowoduje ruch plazmy.

Po tej uwadze omówimy zjawisko konwekcji magnetosferycznej i przyczyny powstania skoku koncentracji elektronów na granicy plazmosfery.

pole elektryczne konwekcji

Wiatr słoneczny opływający magnetosferę wywołuje pole elektryczne skierowane z zachodu na wschód i zw. polem elektrycznym konwekcji. Ponieważ w różnych modelach magnetosfery mechanizm generacji tego pola jest różny, nie będziemy go szerzej omawiali. Pole elektryczne konwekcji przenosi się bez strat wzdłuż linii sił pola geomagnetycznego do wewnątrz magnetosfery i plazmosfery. Jego natężenie jest duże w pobliżu magnetopauzy, a więc tam, gdzie parametr L jest duży. W warunkach spokojnych magnetycznie osiąga ono wartość 20 mV/m, a w czasie burz magnetycznych nawet 120 mV/m, lecz szybko maleje w miarę oddalania się od magnetopauzy. Pole konwekcji wraz z polem geomagnetycznym powodują ciągły ruch plazmy wewnątrz magnetosfery w kierunku od ogona, czyli od strony nocnej do dziennej. Można więc mówić o cyrkulacji plazmy w magnetosferze: w pobliżu magnetopauzy wiatr słoneczny unosi plazmę do ogona magnetosfery, skąd dzięki polu konwekcji przedostaje się ona do wnętrza magnetosfery.

pole elektryczne korotacji

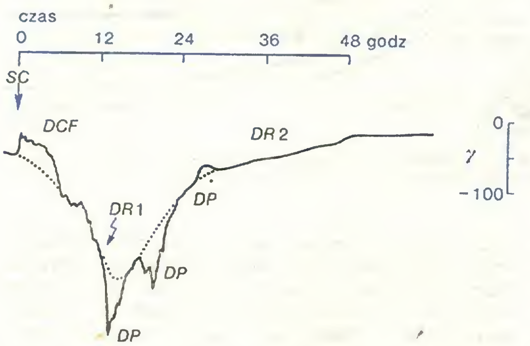
W przestrzeni okołoziemskiej oprócz pola elektrycznego konwekcji występuje pole elektryczne korotacji. Wskutek lepkości plazma w przestrzeni okołoziemskiej rotuje razem z Ziemią. Ruch plazmy w polu geomagnetycznym powoduje powstanie pola elektrycznego. Charakter zależności zmian natężenia tego pola od parametru L jest zupełnie inny niż dla pola konwekcji. Natężenie pola korotacji mało zależy od L (aż do $L \approx 4$) i wynosi ok. 15 mV/m. Ale jeśli L jest większe od 4, natężenie pola korotacji szybko maleje.

Istnienie pól konwekcji i korotacji warunkuje powstanie skoku koncentracji elektronów na granicy plazmosfery. Tam, gdzie natężenie pola korotacji jest większe od natężenia pola konwekcji, gęstość plazmy powinna być większa niż w obszarze, w którym dominuje pole konwekcji. Istotnie, pole konwekcji jest przyczyną cyrkulacji magnetosferycznej, która powoduje, że cząstki uciekają do ogona magnetosfery, a następnie w przestrzeń międzyplanetarną. Ich koncentracja jest więc mniejsza niż koncentracja cząstek w obszarze, w którym natężenie pola korotacji jest wystarczająco duże. A zatem możemy przyjąć, że plazmopauza jest powierzchnią, na której pola elektryczne konwekcji i korotacji są równe. Poniżej plazmopauzy dominuje pole korotacji, powyżej zaś — pole konwekcji.

Burze magnetyczne i zorze polarne

Pole magnetyczne Ziemi często jest zaburzone. Przejawia się to w postaci zmian amplitudy składowych pola magnetycznego rejestrowanych na magnetogramach na powierzchni Ziemi (→ Magnetyzm ziemski). Jeżeli amplituda zmian jest bardzo duża, mówimy o burzy magnetycznej. Typowym przykładem magnetogramu otrzymanego w czasie burzy magnetycznej jest rys. 7. Obrazuje on jedynie zmiany składowej poziomej (horyzontalnej) *H*. Burza często zaczyna się skokiem natężenia pola magnetycznego. Mówimy wówczas o „nagłym początku burzy" (*SC* na rys. 7). Po nagłym początku, przez kilka godzin natężenie pola jest duże. Jest to początkowa faza burzy (*DCF*). Po kilku godzinach rozpoczyna się główna faza burzy, którą charakteryzuje znaczny spadek natężenia pola (*DR*1). Z chwilą osiągnięcia minimum natężenie pola powraca powoli do wartości, jaką miało przed burzą (*DR*2). Niekiedy natężenie pola w czasie głównej fazy burzy silnie maleje, dając efekty oznaczone na rysunku literami *DP*.

fazy burzy magnetycznej

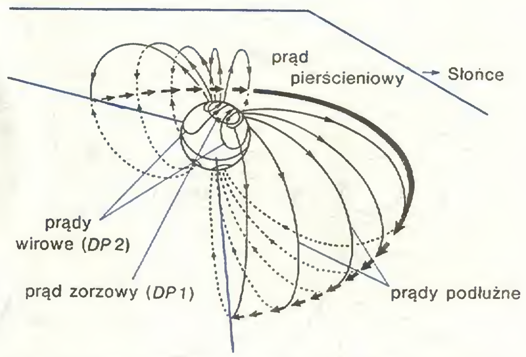


Rys. 7. Przykład magnetogramu uzyskanego w czasie burzy magnetycznej (pokazano tylko składową poziomą *H* pola magnetycznego). Czas liczony jest od momentu „nagłego początku burzy" (*SC*). Skala zmian uwidoczniona jest z prawej strony ($1\ \gamma = 10^{-9}$T)

Jak wytłumaczyć te zjawiska? Przyczyną burz magnetycznych są zaburzenia na Słońcu. W wyniku rozbłysku na Słońcu odrywa się od niego obłok plazmy, która jest gęstsza i porusza się szybciej niż plazma w „spokojnym" wietrze słonecznym. W ciągu 1–2 dni obłok taki dociera w pobliże Ziemi. W momencie zetknięcia się jego z magnetosferą zostaje ona w sposób nagły, impulsowy zaburzona. Powstaje fala uderzeniowa (→ Fale uderzeniowe), która rozchodzi się w magnetosferze, czego efektem jest nagły początek burzy (*SC*). Następnie plazma obłoku opływa magnetosferę i powstają prądy płynące po powierzchni magnetopauzy. Prądy te zmieniają natężenie pola magnetycznego wewnątrz magnetosfery, co się przejawia w postaci początkowej fazy burzy (*DCF*). Aby wyjaśnić główną fazę burzy, należy założyć, że w magnetosferze płyną prądy pierścieniowe. Powinny one płynąć ku zachodowi, by się zmniejszyła składowa horyzontalna pola magnetycznego (zgodnie z „regułą lewej ręki"). W rzeczywistości musimy przyjąć, że istnieją dwa prądy pierścieniowe: silny, odpowiedzialny za zmiany *DR*1, którego natężenie szybko wzrasta i który znika w ciągu jednego dnia, oraz słaby (*DR*2), którego natężenie maleje powoli. Przypuszcza się, że prądy pierścieniowe są wynikiem wstrzyknięcia do magnetosfery protonów o małych energiach, które następnie zmuszone są do ruchu po torach przedstawionych na rys. 6. Miejscem, w którym cząstki wstrzykiwane są do magnetosfery, jest prawdopodobnie jej ogon, ale problem ten nie jest w pełni wyjaśniony. Na rys. 8 prąd pierścieniowy przedstawiony jest w postaci grubych strzałek. Rozpływa się on ku obszarom polarnym, wzdłuż linii sił pola geomagnetycznego. W obszarze polarnym prądy wzdłuż linii pola zamykają się na wysokości jonosferycznej warstwy *E*, tworząc prąd zorzowy (*DP*1), zaznaczony (rys. 8) grubą strzałką na powierzchni kuli wyobrażającej Ziemię. Płynie on w wąskim (rzędu 400 km) i cienkim (kilka km) pasie wokół biegunów magnetycznych. Po stronie dziennej prąd płynie ku zachodowi, po stronie nocnej — ku wschodowi, a zatem w dzień powoduje zmniejszenie się składowej *H*, w nocy zaś — jej zwiększenie.

prądy pierścieniowe

prąd zorzowy



Rys. 8. Rozkład prądów w magnetosferze typowy dla burzy magnetycznej. Strzałki pokazują prąd pierścieniowy, który rozpływa się wzdłuż linii sił pola magnetycznego do jonosfery polarnej, tworząc tam prąd zorzowy

Drugi rodzaj prądu (*DP*2) — to układy prądów wirowych w jonosferze płynących po stronie wschodniej i zachodniej, głównie na dużych szerokościach geomagnetycznych. Efekt ich widoczny jest również na średnich szerokościach. Powstanie tych prądów związane jest z konwekcją magnetosferyczną.

prądy wirowe

Należy jeszcze wspomnieć o efekcie, który spowodowany jest dużym przewodnictwem elektrycznym plazmy, a polega na tym, że linie sił pola magnetycznego poruszają się razem z plazmą, są w nią „wmrożone". Tak więc z ruchem konwekcyjnym plazmy magnetosferycznej związany jest ruch linii sił pola. Bierze w nim udział cała linia siły łącznie z częścią wnikającą do jonosfery. „Przyczepione" są do niej elektrony i jony. Jednak praktycznie biorąc, tylko elektrony mogą uczestniczyć w ruchu linii sił. Jony bowiem są hamowane przez zderzenia z cząstkami neutralnymi. W rezultacie elektrony, unoszone liniami sił pola magnetycznego wprawionymi w ruch ruchami konwekcyjnymi plazmy magnetosferycznej, poruszają się w jonosferze względem prawie nieruchomych jonów. Powstaje więc prąd płynący w kierunku przeciwnym do ruchu linii sił pola. Ponieważ w obszarach polarnych konwekcja ma kierunek od Słońca, przeto linie sił przemieszczają się od Słońca, a prąd płynie ku Słońcu. W miarę przechodzenia do coraz mniejszych szerokości geomagnetycznych kierunek ruchu plazmy w magnetosferze coraz bardziej odchyla się od kierunku od Słońca i na średnich szerokościach przybiera już kierunek ku Słońcu. W efekcie również prąd w jonosferze zmienia kierunek. Powstają dwa układy prądów wirowych. Na rys. 8 są one zaznaczone cienkimi strzałkami na powierzchni Ziemi. Układy prądów *DP*1 i *DP*2 wywołują zmiany pola magnetycznego oznaczone na rys. 7 literami *DP*.

Bardzo spektakularnym zjawiskiem, występującym w okresie burz magnetycznych na dużych szerokościach geomagnetycznych, są zorze polarne (tabl. 6, il. 20). Zielone i czerwone obłoki, promienie lub draperie pobudzały wyobraźnię ludzką od najdawniejszych czasów. Jednak na wyjaśnienie mechanizmu powstawania zórz musieliśmy czekać aż do dwudziestego wieku. Jest rzeczą pewną, że zorze to świecenie atmosfery wywołane bombardowaniem elektronami o energii rzędu kilkudziesięciu keV. Cząstki te wzbudzają atomy, które — wracając do stanu podstawowego — wypromieniowują energię w postaci kwantów

zorze polarne

światła. Powstaje pytanie, skąd się biorą cząstki o takiej energii w atmosferze. Na rys. 5 zaznaczono obszar leżący poza pasami promieniowania i łączący się z ogonem magnetosfery. Obszar ten, zrzutowany wzdłuż linii sił pola magnetycznego na powierzchnię Ziemi, daje dwa pasy otaczające bieguny magnetyczne. Nazywamy je „owalami zorzowymi". Owale przybliżają się do bieguna (do ok. 78° szerokości) w dzień, a oddalają (do ok. 68° szerokości) w nocy. Położenie owalu względem Słońca nie zmienia się. Gdy Ziemia rotuje, owal przechodzi przez różne obszary jej powierzchni. Na przykład nocna część owalu zatacza okrąg położony na szerokości ok. 68°. Ten pas na Ziemi nazywa się strefą zorzową, tu zorze występują najczęściej. Nocna część owalu łączy się z ogonem magnetosfery. Z ogona, albo nawet spoza magnetosfery (ale przez ogon), cząstki przedostają się do bliższych Ziemi obszarów magnetosfery. W czasie wędrówki w ogonie są one przyspieszane i gdy dochodzą wreszcie do granic atmosfery, mogą mieć energie wystarczająco duże na to, aby wywołać zjawisko zorzy polarnej.

strefa zorzowa

Na zakończenie możemy tylko wspomnieć, że z burzami magnetycznymi i zorzami polarnymi wiąże się wiele zjawisk, np. wzrost pochłaniania fal radiowych, powstawanie nieregularności koncentracji jonów i elektronów w jonosferze, okresowe, o małej amplitudzie pulsacje pola magnetycznego, spadek natężenia promieniowania kosmicznego (efekt Forbusha → Promieniowanie kosmiczne).

A. D. Daniłow *Chemia, atmosfera, kosmos*, Warszawa 1976; F. Delobeau *The Environment of the Earth*, Dordrecht–Holland 1971; P. Oberc *Magnetosfera*, Acta Geophys. Pol. 18, 103 (1970); H. Rishbeth, O. K. Garriott *Introduction to Ionospheric Physics*, New York 1969.

Magnetyzm ziemski

Magdalena Kądziałko-Hofmokl

Kilkaset lat przed naszą erą zauważono zjawiska, które dzisiaj określa się mianem zjawisk magnetycznych. W miarę upływu czasu i przybywania danych obserwacyjnych powstała nauka o magnetyzmie ziemskim, z której z kolei wyodrębniły się bardziej szczegółowe dyscypliny, zajmujące się takimi zagadnieniami jak stałe pole magnetyczne, paleomagnetyzm, zmienne pole magnetyczne, indukcja elektromagnetyczna we wnętrzu Ziemi.

pierwsze obserwacje i instrumenty

Pierwszych odkryć związanych ze zjawiskami magnetycznymi dokonali Chińczycy; już w starożytnych Chinach znane były przyciągające własności skały zawierającej magnetyt. Przypuszcza się również, że Chińczycy znali kierunkowe składowe pola magnetycznego Ziemi. Najwcześniejszy, z 720 r. n.e., znany opis wyznaczania deklinacji magnetycznej pochodzi również z Chin. Pierwszym instrumentem wykorzystującym zjawisko deklinacji był kompas, znany prawdopodobnie od czasów starożytnych. Najdawniejszy opis tego instrumentu, sporządzony w Europie, pochodzi dopiero z 1190 r., a jego autorem jest angielski mnich, Aleksander Neckam.

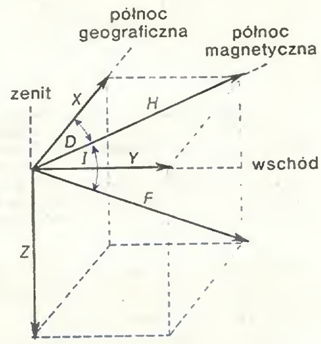
Przez długie lata uważano, że kąt deklinacji zależy od używanego instrumentu. Dopiero dokładne pomiary tego kąta, dokonywane w połowie XVI w. w różnych punktach kuli ziemskiej przez portugalskiego żeglarza Joao da Castro, wykazały ostatecznie, że deklinacja zależy od miejsca, w którym się dokonuje pomiaru. W XVI wieku zauważono (G. Hartmann, R. Norman), że igła kompasu tworzy pewien kąt z płaszczyzną horyzontu — było to odkrycie inklinacji magnetycznej.

traktat Gilberta

Z upływem czasu przybywało informacji o zjawiskach związanych z magnetyzmem ziemskim. W 1600 r. ukazało się dzieło Wiliama Gilberta *De Magnete* ... (O magnesie, ciałach magnetycznych i największym magnesie — Ziemi), w którym Gilbert wykazał, że pole magnetyczne Ziemi ma takie cechy jak pole jednorodnie namagnesowanej kuli, przy czym położenie biegunów magnetycznych Ziemi pokrywa się z położeniem jej biegunów geograficznych. Wnioski Gilberta miały ogromne znaczenie dla rozwoju nauki o polu geomagnetycznym. Oprócz wyjaśnienia deklinacji i inklinacji Gilbert wykazał, że źródeł pola magnetycznego Ziemi należy szukać w jej wnętrzu, a nie poza nią. Z czasem okazało się, że wniosek Gilberta jest słuszny nie w odniesieniu do całego pola geomagnetycznego, lecz do jego przeważającej części, zwanej stałym polem geomagnetycznym. Na stałe pole magnetyczne, pochodzące od źródeł znajdujących się we wnętrzu Ziemi, nakłada się bowiem zmienne pole magnetyczne, którego źródła znajdują się w przestrzeni pozaziemskiej.

składowe pola magnetycznego Ziemi

Natężenie pola magnetycznego $\vec{F}$ mierzone na powierzchni Ziemi jest wielkością wektorową (właściwie wielkością mierzoną jest indukcja magnetyczna, lecz przyjęto stosować nazwę pole magnetyczne). Można je przedstawić (rys. 1) za pomocą trzech wielkości: 2 składowych — składowej pionowej Z (mierzonej pionowo w dół), składowej poziomej H — oraz deklinacji D (deklinacja magnetyczna jest to kąt pomiędzy



Rys. 1. Składowe całkowitego wektora pola magnetycznego Ziemi $\vec{F}$

południkiem geograficznym a magnetycznym w miejscu obserwacji; mierzy się ją w kierunku wschodnim). Innym sposobem przedstawienia pola może być podanie wartości całego natężenia pola F, deklinacji D i inklinacji I (inklinacja jest to kąt, który tworzy wektor $\vec{F}$ z poziomem; inklinację mierzoną w dół od poziomu uważamy za dodatnią). Rzadziej stosuje się składowe X, Y, Z skierowane odpowiednio: X — w kierunku północnym, Y — w kierunku wschodnim, Z — pionowo w dół.

Stałe pole geomagnetyczne

Ukazanie się traktatu W. Gilberta było przełomowym wydarzeniem w historii magnetyzmu ziemskiego. Coraz większa liczba prac o zjawiskach magnetyzmu oraz większa liczba danych obserwacyjnych pozwoliły w 1839 r. uczonemu niemieckiemu Fryderykowi Gaussowi ująć matematycznie główny wniosek Gilberta, czyli przedstawić stałe pole magnetyczne Ziemi w pierwszym przybliżeniu jako pole dipola umieszczonego w środku Ziemi i skierowanego wzdłuż jej osi obrotu.

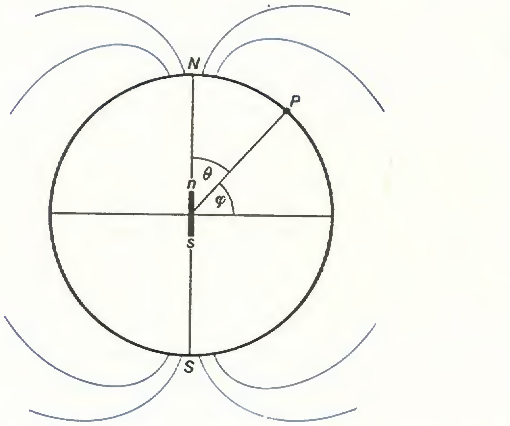
Metoda analizy Gaussa — tzw. analiza sferyczno-harmoniczna pola magnetycznego Ziemi — polega na przedstawieniu natężenia

metoda analizy Gaussa

tego pola w każdym punkcie na powierzchni Ziemi w funkcji współrzędnych geograficznych tego punktu. Podstawą analizy są wartości natężenia pola otrzymane w wielu punktach, rozmieszczonych możliwie regularnie na powierzchni Ziemi. Z obliczeń otrzymuje się natężenie całkowitego wektora F lub jego składowych X, Y, Z w postaci nieskończonych szeregów współczynników stałych pomnożonych przez funkcje długości i szerokości geograficznej. Ograniczając się do kilku pierwszych wyrazów tych szeregów, otrzymuje się dla poszczególnych składowych pola w punkcie P (rys. 2) wzory: $X = -g_1^0 \sin\theta + (g_1^1 \cos\lambda + h_1^1 \sin\lambda)\cos\theta$,

$$Y = g_1^1 \sin\lambda - h_1^1 \cos\lambda, \quad Z = 2[g_1^0 \cos\theta + (g_1^1 \cos\lambda + h_1^1 \sin\lambda)\sin\theta],$$

gdzie dla epoki 1955, 0 współczynniki liczbowe wynosiły: $g_1^0 = -3{,}05 \cdot 10^5$ T, $g_1^1 = -0{,}23 \cdot 10^5$ T, $h_1^1 = 0{,}58 \cdot 10^5$ T. Wielkość θ jest dopełnieniem szerokości geograficznej punktu P, λ — długością geograficzną punktu P. Wyrażenia te opisują składowe pola jednorodnie namagnesowanej kuli, w tym wypadku kuli ziemskiej, i zarazem składowe pola dipola umieszczonego w środku Ziemi i skierowanego wzdłuż jej osi obrotu (dipol osiowy). Linie sił pola takiego dipola układają się jak na rys. 2 — „wchodzą" w głąb Ziemi i „wychodzą" z niej w punktach zwanych biegunami. Dalsze człony szeregów Gaussa opisują składowe pola niedipolowego.



Rys. 2. Linie sił pola dipola magnetycznego (n–s) umieszczonego w środku Ziemi i skierowanego wzdłuż jej osi obrotu. N biegun północny, S biegun południowy, P punkt na powierzchni Ziemi o szerokości geograficznej φ i dopełnieniu szerokości geograficznej θ. W przypadku dipola osiowego współrzędne geograficzne pokrywają się z geomagnetycznymi, a inklinacja I wiąże się z dopełnieniem szerokości θ następującym wzorem: $\operatorname{tg} I = 2 \operatorname{ctg} \theta$

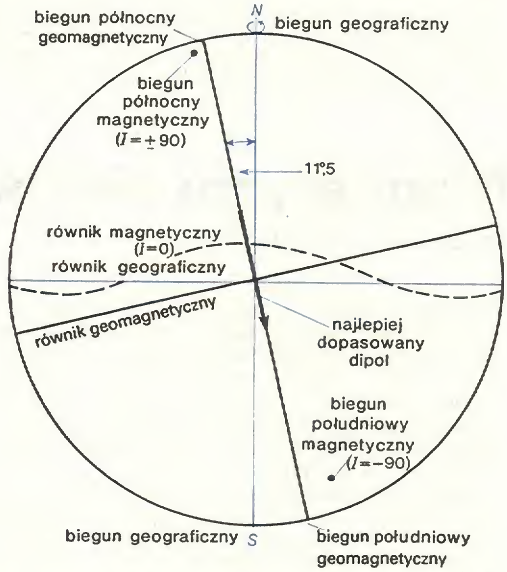
Natężenie stałego pola geomagnetycznego stanowi ponad 99% natężenia całego pola obserwowanego na powierzchni Ziemi. Lepiej niż dipol osiowy opisuje je dipol umieszczony również w środku Ziemi, ale tworzący kąt 11°,5 z jej osią obrotu (rys. 3). Moment magnetyczny takiego dipola wynosi $8 \cdot 10^{15}$ Tm^3. (Moment magnetyczny kuli o powierzchni 1 m^2 wykonanej z miękkiego żelaza wynosi ok. $1 \cdot 10^{-4}$ Tm^3). Pole, które on wytwarza, opisuje ok. 90% pola stałego. Pozostałe 10% to wspomniane już pole niedipolowe, o nieregularnym rozkładzie na powierzchni Ziemi.

pole dipolowe

Stałe pole magnetyczne, mimo że nosi nazwę stałego, ulega powolnym zmianom. Obecnie istnieje na świecie ponad 150 obserwatoriów magnetycznych prowadzących ciągłą rejestrację poszczególnych składowych. Poza tym co kilka lat przeprowadza się pomiary pola geomagnetycznego w określonych punktach, zwanych wiekowymi. Materiał obserwacyjny uzupełniają pomiary dokonywane ze statków na oceanach i morzach oraz z samolotów na różnych wysokościach nad powierzchnią Ziemi. Dane obserwacyjne są podstawą do sporządzania co kilka lat światowych map poszczególnych składowych pola magnetycznego. Wyniki uzyskane w obserwatoriach w postaci ciągłego zapisu poszczególnych składowych pola geomagnetycznego — magnetogramu (rys. 25) — zawierają zarówno część stałą pola, jak i zmienną. Aby otrzymać składowe tylko pola stałego, uśrednia się wyniki obserwacji dotyczące jakiegoś okresu — najczęściej jest to okres jednego roku. Tę wartość średnią, zwaną w tym wypadku średnią roczną, otrzymuje się obliczając kolejno średnie godzinne, dobowe, miesięczne i — na końcu — roczne, przy czym odrzuca się wyniki obserwacji dokonanych podczas dni szczególnie zaburzonych magnetycznie. Otrzymana w ten sposób średnia roczna wartość pola nie zawiera zmian godzinnych, dobowych ani sezonowych. Ponieważ, o czym będzie mowa później, na wymienione tu zmiany nakładają się zmiany związane z okresowym występowaniem plam na Słońcu (okres ok. 11 lat), czasem oblicza się średnie wartości pola dla 10-letnich przedziałów czasu. Średnie wartości otrzymane dla jakiegoś przedziału czasu przypisuje się momentowi wypadającemu w środku tego przedzia-

punkty wiekowe

otrzymywanie wartości średnich

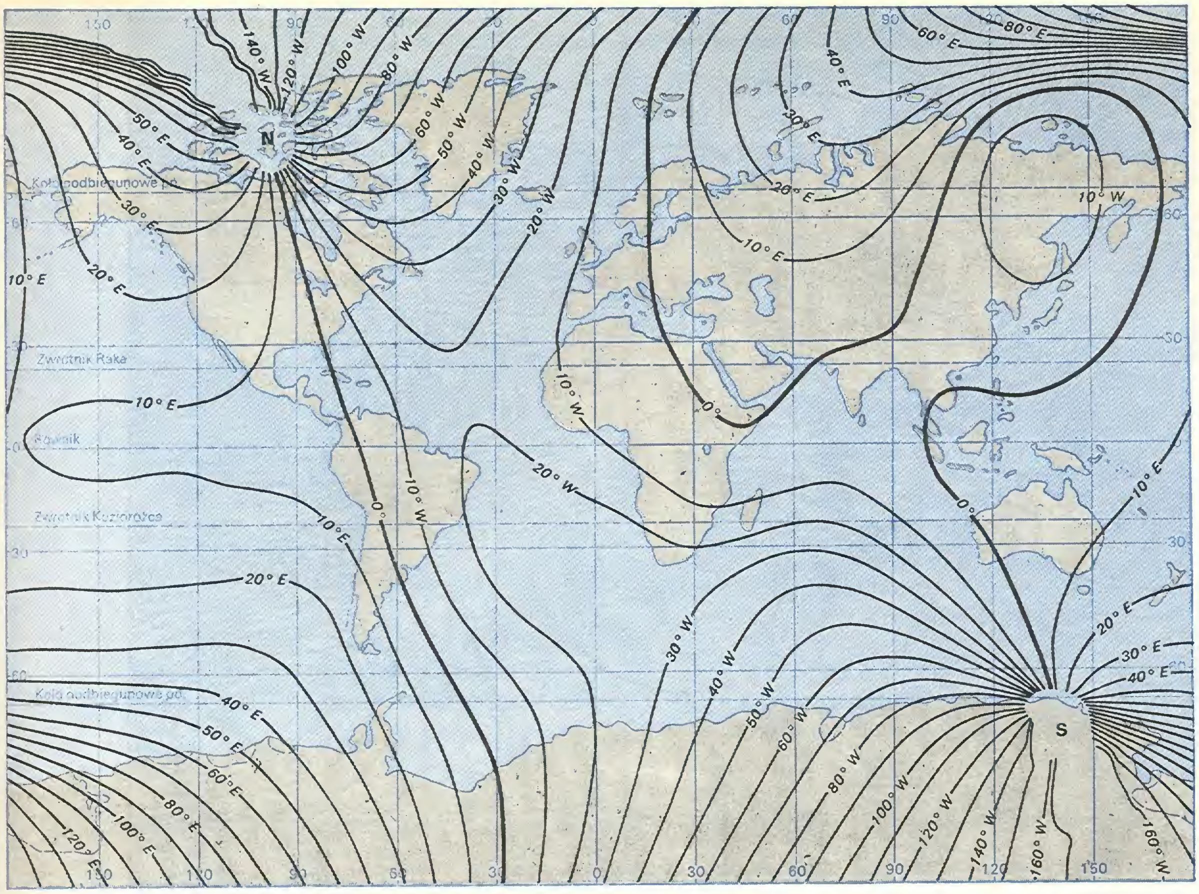


Rys. 3. Dipol magnetyczny umieszczony w środku Ziemi, nachylony pod kątem 11°,5 względem jej osi obrotu. Na rysunku są zaznaczone bieguny geograficzne północny i południowy (N i S), bieguny geomagnetyczne, czyli miejsca przecięcia osi dipola z powierzchnią Ziemi, oraz bieguny magnetyczne, czyli miejsca, w których inklinacja magnetyczna I wynosi +90° (biegun północny) i −90° (biegun południowy). Południowy biegun dipola jest skierowany w stronę półkuli południowej, a biegun północny — w stronę północnej. A zatem z punktu widzenia fizyki na półkuli geograficznej południowej znajduje się północny biegun magnetyczny, a na półkuli północnej — południowy. Jednak ze względu na tradycję i wygodę przyjęto nazwy i oznaczenia zgodne z nazwami geograficznymi. Na rysunku zaznaczony jest również równik geograficzny, geomagnetyczny i magnetyczny. Analogicznie do współrzędnych geograficznych wprowadzono współrzędne geomagnetyczne: szerokość geomagnetyczna punktu jest to kąt, który tworzy wektor wodzący poprowadzony ze środka Ziemi do tego punktu z płaszczyzną równika geomagnetycznego, a długość geomagnetyczna jest to kąt, który tworzy południk geomagnetyczny przechodzący przez dany punkt z południkiem przechodzącym przez biegun geograficzny. Obecnie bieguny geomagnetyczne zajmują następujące położenia: biegun geomagnetyczny północny 78°,5 *N* 70°*W*, biegun geomagnetyczny południowy 78°,5 *S* 110°*E* (bieguny geomagnetyczne leżą naprzeciwko siebie). Położenia biegunów magnetycznych były w tym czasie następujące: biegun magnetyczny północny 75°,5 *N*, 101°*W*, biegun magnetyczny południowy 66°*S* 140°*E* (bieguny magnetyczne nie leżą naprzeciwko siebie) (wg M. W. Mc Elhinny *Paleomagnetism and plate tectonics*, Cambridge 1973)

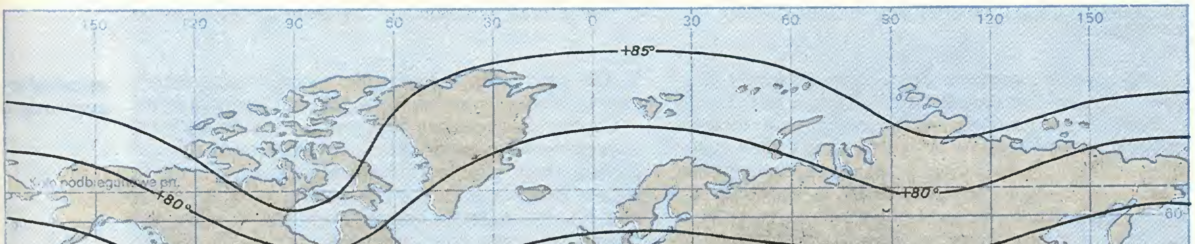
łu; moment taki nazywa się epoką. Przeprowadzając analizę sferyczno-harmoniczną pola dla jakiejś epoki, wykorzystuje się również, jak wspomniano, wyniki jednorazowych pomiarów składowych pola. I tu również uwzględnia się poprawki na pole zmienne, korzystając w tym celu z rezultatów uzyskanych w najbliższych obserwatoriach. Na rys. 4a, b, c, przedstawiono mapę izogon (linie jednakowej deklinacji), izoklin (linie jednakowej inklinacji) i izodynam (linie jednakowego natężenia składowej H) dla epoki 1965,0. Widoczne na rys. 4a punkty, w których się zbiegają izogony, to bieguny magnetyczne, północny N i południowy S. Na biegunie północnym inklinacja wynosi +90°, a na południowym −90°, składowa pozioma H w miejscu biegunów wynosi zero. Charakterystyczną linią widoczną na mapie izoklin jest linia

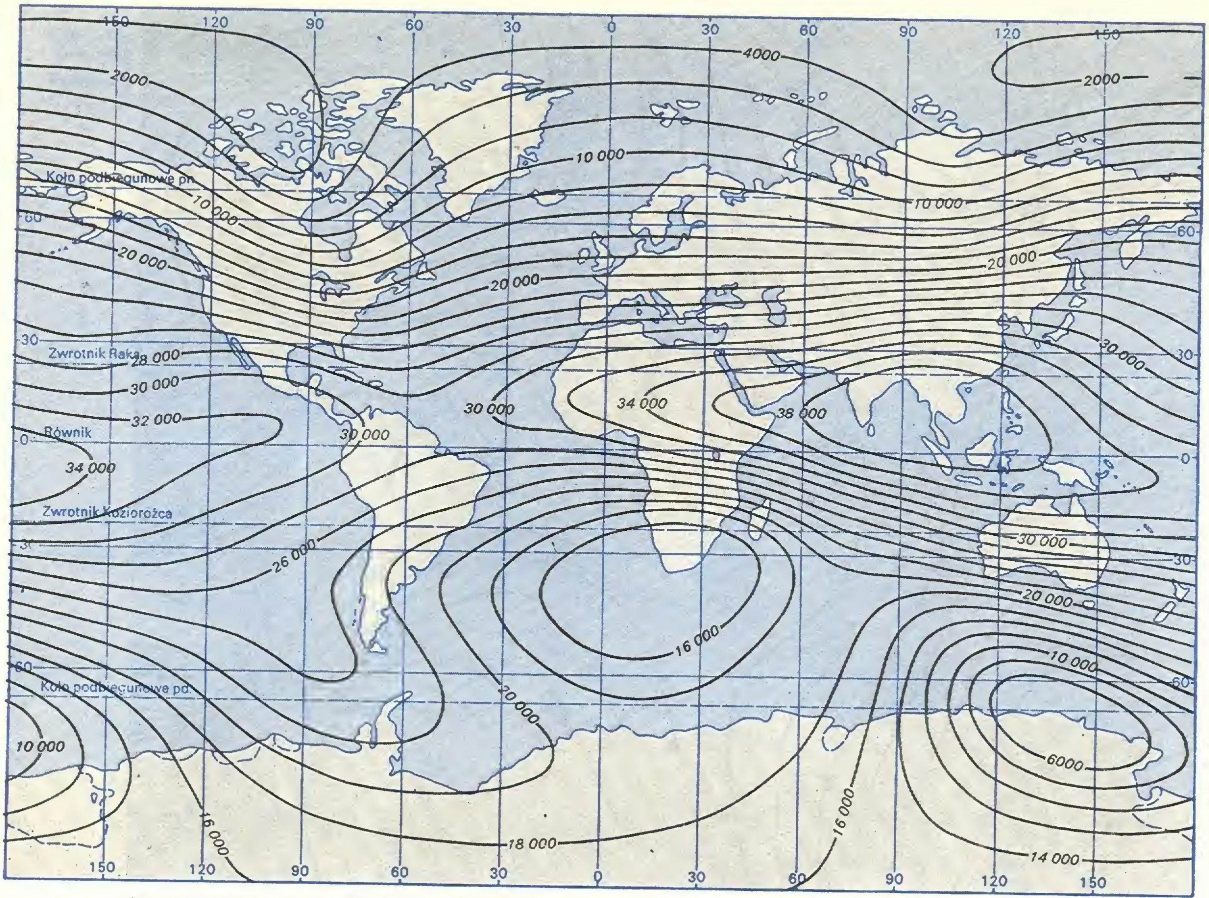
epoka

a)


N
S
Koło podbiegunowe pn.
Zwrotnik Raka
Równik
Zwrotnik Koziorożca
Koło podbiegunowe pd.
140° W
120° W
100° W
80° W
60° W
50° W
40° W
30° W
20° W
10° W
0°
10° E
20° E
30° E
40° E
50° E
60° E
80° E
100° E
120° E
160° W

b)


+85°
+80°
+70°
+60°
+50°
+25°
0°
-25°
-50°
-60°
-70°
-75°
-80°
Koło podbiegunowe pn.
Zwrotnik Raka
Równik
Zwrotnik Koziorożca



c)

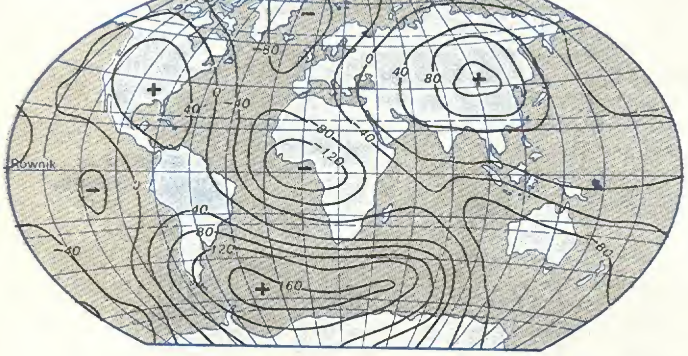
Rys. 4. Pole magnetyczne Ziemi dla epoki 1965,0: a) deklinacja magnetyczna w stopniach, b) inklinacja magnetyczna w stopniach, c) natężenie składowej poziomej pola w nanoteslach, nT (wg K. A. Wienert *Notes on geomagnetic observatory and survey practice*, UNESCO 1970)

równika magnetycznego (oznaczona kolorem czarnym — grubo), wzdłuż której inklinacja jest 0.

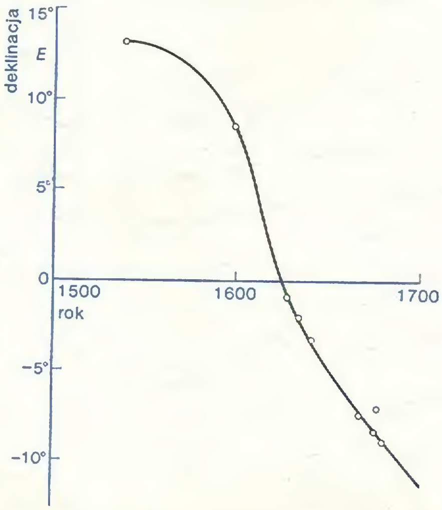
W 1947 r. ukazało się zestawienie danych obserwacyjnych za lata 1905–1945. Autorzy przedstawili wyniki analizy numerycznej, przeprowadzonej metodą Gaussa, dla wielu epok na zebranym materiale. Od tej pory zaczęto wykonywać podobne analizy co kilka lat. Jednym z najciekawszych i chyba najistotniejszych rezultatów tych prac jest możliwość rozdzielenia stałego pola na części dipolową i niedipolową. Rysunek 5 przedstawia rozkład pola niedipolowego na powierzchni Ziemi w epoce 1945,0. Cechą tego pola jest widoczna na rysunku obecność centrów, wokół których układają się izolinie. Pole niedipolowe osiąga w centrach maksymalne amplitudy, sięgające $1{,}5 \cdot 10^{-5}$ T. Obszary o średnicach wielu tysięcy kilometrów wokół centrów nazwano anomaliami regionalnymi.

anomalie regionalne

Szczegółowe pomiary pola geomagnetycznego w poszczególnych rejonach kuli ziemskiej wykazują, że istnieją obszary, w których wartość pola jest znacznie wyższa (lub niższa) niż w obszarach sąsiednich. Są to tzw. lokalne anomalie magnetyczne związane z występowaniem skał o dużej zawartości minerałów magne-



Rys. 5. Pole niedipolowe dla epoki 1945,0 (w odstępach $4 \cdot 10^{-7}$T). Znaki + i — oznaczają maksymalne wartości ujemne i dodatnie pola niedipolowego (wg D. H. Tarling *Principles and Applications of Paleomagnetism*, London 1971)



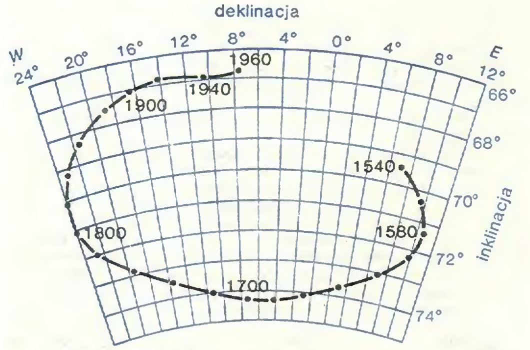
Rys. 6. Zmiany wiekowe deklinacji magnetycznej w Gdańsku wg obserwacji J. Retyka (1539), Anonima (1600), J. Heweliusza (1628, 1635, 1642, 1670, 1667?, 1682) i E. Halleya (1679) (wg T. Olczak *Jan Heweliusz i magnetyzm ziemski*, Postępy Astronomii, 3 65, 1955)

tycznych. Przykładami takich anomalii są anomalia kurska oraz anomalie w północnej Szwecji. Głębokości źródeł wywołujących anomalie lokalne nie przekraczają 25 km, podczas gdy źródła pól niedipolowego i dipolowego znajdują się nieporównanie głębiej, bo poniżej 2900 km. O pochodzeniu pola stałego będzie mowa przy końcu następnego paragrafu.

zmiany wiekowe

Średnie roczne wartości poszczególnych składowych pola stałego, otrzymywane w opisany wyżej sposób, zmieniają się w miarę upływu czasu. Długookresowe zmiany pola nazwano zmianami wiekowymi. W 1634 r. angielski uczony H. Gellibrand zauważył, że deklinacja mierzona w Londynie zmienia się wraz z czasem. Przed Gellibrandem zmiany deklinacji magnetycznej obserwował już Jan Heweliusz (1611–1687), wybitny astronom z Gdańska. Na rys. 6 przedstawiona jest krzywa zmian wiekowych deklinacji magnetycznej w Gdańsku zestawiona przez T. Olczaka na podstawie obserwacji z lat 1539–1679. Wiele cennych danych o deklinacji magnetycznej dostarczają zegary słoneczne i instrumenty astronomiczne. Zarówno na jednych, jak i na drugich twórcy bądź uczeni, którzy z nich korzystali, zaznaczali kąt deklinacji właściwy dla miejsca budowy lub użytkowania instrumentu. Z wykresów zmian deklinacji czy inklinacji w długich okresach czasu widać, że zmiany te są cykliczne. Na rys. 7 przedstawiona jest krzywa zmian wiekowych deklinacji i inklinacji dla Londynu od 1540 do 1960 r. Przebieg krzywej wskazuje, że okres zmian wynosi ok. 400 lat. Obecnie wyróżnia się kilka typów zmian wiekowych, o okresach rzędu kilku tysięcy, kilkuset i kilkudziesięciu lat. Ze szczegółowych badań części dipolowej i niedipolowej stałego pola wynika, że ulegają one zmianom wiekowym. W ciągu ostatnich 100 lat natężenie pola dipolowego malało o ok. 0,04% rocznie i pole przesuwało się w kierunku zachodnim ok. 0°,05 w ciągu roku. Zjawisko to nazwano dryfem zachodnim pola niedipolowego. Pole niedipolowe zmieniało swoje natężenie mniej systematycznie, w jednych okresach rosło, w innych znów malało. W 1950 r. E. Bullard i in. zauważyli, że pole niedipolowe przesuwa się na zachód z prędkością 0°,18 rocznie.

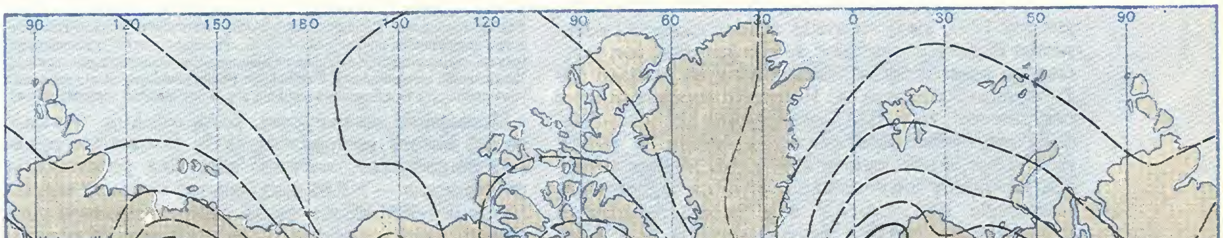
dryf zachodni



Rys. 7. Zmiany wiekowe deklinacji i inklinacji w Londynie od 1540 r. do 1960 r. (wg M. H. P. Bott, *The Interior of the Earth*, London 1971)

Zmiany wiekowe poszczególnych składowych, podobnie jak same składowe, wygodnie jest przedstawić w postaci map izolinii, zwanych izoporami. Mapy izopor różnych epok zestawili po raz pierwszy E. Vestine i in. w 1947 r. Jedną z wykonanych przez nich map przedstawia rys. 8. I tu, podobnie jak na mapie pola niedipolowego, zwracają uwagę charakterystyczne centra, w środku których zmiany wiekowe przybierają szczególnie duże wartości. Centra te nazwano ogniskami izopor. W pięciu z nich zmiany składowej pionowej przekraczają 10^{-7} T/rok. Po-

izopory



Rys. 8. Zmiany wiekowe składowej pionowej dla epoki 1922,5 w nanoteslach na rok (wg M. H. P. Bott *The Interior of the Earth*, London 1971)

równanie map izopor dla różnych epok wykazało, że ogniska przesuwają się w kierunku zachodnim z prędkością ok. 0°,2 rocznie. A zatem słuszny był wyrażony już w 1692 r. przez Edmunda Halleya pogląd, że zmiany wiekowe są wywołane dryfem zachodnim pola niedipolowego.

W tym miejscu należy zauważyć, że przedstawiony powyżej podział pola geomagnetycznego na części dipolową i niedipolową jest umowny i nie ma sensu fizycznego, pozwala jednak na łatwiejsze przedstawienie własności pola stałego.

Paleomagnetyzm

paleomagnetyzm i archeomagnetyzm

Kilkadziesiąt lat temu zaczęła się gwałtownie rozwijać nowa dziedzina magnetyzmu ziemskiego — paleomagnetyzm, nauka zajmująca się polem magnetycznym Ziemi w przeszłości geologicznej. Przedmiotem badań są skały, które w pewnych warunkach mogą przechowywać informacje o natężeniu i kierunku pola geomagnetycznego panującego w okresie ich powstawania. Dziedziną pokrewną paleomagnetyzmowi jest archeomagnetyzm — nauka zajmująca się ziemskim polem magnetycznym w przeszłości historycznej i prehistorycznej, wykorzystująca wyniki badań wypalanych glin.

Dzięki paleomagnetyzmowi można poznawać nie tylko przeszłość pola magnetycznego Ziemi, ale i samej Ziemi. Wyniki badań są niejednokrotnie podstawą do przyjęcia lub odrzucenia jakiejś hipotezy geofizycznej czy geologicznej. Chociaż paleomagnetyzm jako nauka zaczął się rozwijać stosunkowo niedawno, to jednak początki jego sięgają czasów starożytnych. We wstępie była już mowa o odkryciu w starożytnych Chinach skały magnetytowej i jej własności. Jedną z najważniejszych dat w historii paleomagnetyzmu jest rok 1797, kiedy to wielki uczony niemiecki Aleksander Humboldt zauważył, że igła kompasu jest przyciągana przez skały w wierzchołkowych partiach gór Palatynatu Reńskiego. A. Humboldt zupełnie słusznie przypuszczał, że te silnie magnetyczne skały zawdzięczają swoje właściwości uderzeniu piorunu. W 1849 r. A. Delesse, a następnie w 1853 r. M. Melloni stwierdzili, że młode lawy wulkaniczne są namagnesowane równolegle do kierunku pola geomagnetycznego. Systematyczne badania namagnesowania skał zapoczątkował w dwudziestych latach naszego wieku R. Chevallier swoimi pracami nad lawami Etny wylanymi w czasach historycznych.

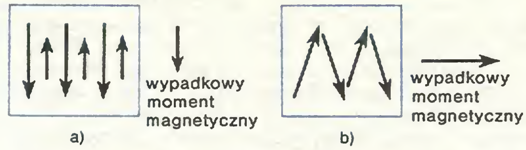
Pozostałość magnetyczna skał

pozostałość magnetyczna

Zjawiskiem będącym podstawą paleomagnetyzmu jest zdolność pewnych skał do uzyskiwania pozostałości magnetycznej o kierunku pola działającego na skałę w okresie jej powstawania. Mierząc wektor pozostałości magnetycznej, można otrzymać wartości kątów deklinacji i inklinacji dawnego pola w miejscu, w którym się znajduje badana skała. Następnie, przy założeniu, że pole geomagnetyczne było zawsze polem dipola umieszczonego w środku Ziemi i skierowanego wzdłuż jej osi obrotu, można znaleźć położenie bieguna magnetycznego w okresie powstawania skały.

właściwości magnetyczne skał

Jak wyżej wspomniano, informacje o polu geomagnetycznym w dawnych epokach geologicznych są zachowane przez skały. Stwierdzenie to należy jednak uściślić: do badań paleomagnetycznych stosuje się skały zawierające minerały magnetyczne. Właściwości magnetyczne skały zależą od ilości i rodzaju zawartych w niej minerałów tego typu. W zależności od pochodzenia skały można podzielić na: skały magmowe (wylewne, np. bazalt, i głębinowe, np. granit), skały osadowe (np. piaskowce, wapienie) i skały zmetamorfizowane (np. gnejsy, marmury). Większość skał zawiera pewną ilość minerałów magnetycznych. Mianem tym określa się minerały o różnej budowie, zarówno ferrimagnetyki, jak i nieskompensowane antyferromagnetyki. Minerały ferrimagnetyczne zawierają jony żelaza dwu- i trójwartościowego ułożone w dwóch regularnych podsieciach w taki sposób, że wektory momentów magnetycznych obu podsieci są antyrównoległe. Wypadkowy moment magnetyczny ferrimagnetyków jest różny od zera, ponieważ momenty magnetyczne tych podsieci nie są równe. W nieskompensowanych antyferromagnetykach wartości momentów są takie same, ale wektory momentów nie są wzajemnie równoległe, lecz nachylone względem siebie pod pewnym kątem. Wzajemne ustawienie momentów magnetycznych w obu typach minerałów jest przedstawione na rys. 9.



Rys. 9. Wzajemne ustawienie momentów magnetycznych: a) w ferrimagnetykach, b) w nieskompensowanych antyferromagnetykach

minerały magnetyczne

W przyrodzie najczęściej spotyka się następujące minerały magnetyczne:

tytanomagnetyty (ferrimagnetyki) — minerały z szeregu magnetyt (Fe_3O_4)-ulwospinel (ulvit) (Fe_2TiO_4),

hemoilmenity — minerały z szeregu hematyt (α-Fe_2O_3)-ilmenit ($FeTiO_4$), o strukturze romboedralnej,

tytanomaghemity — tytanomagnetyty z defektami sieci krystalicznej,

wodorotlenki żelaza, np. getyt, α FeOOH, pirotyny FeS_{1+x} ($0 \leqslant x \leqslant 0,14$).

Spośród wymienionych minerałów największą wartość namagnesowania nasycenia ma magnetyt — ok. $4,5 \cdot 10^{-2}$ T. Wartości namagnesowania nasycenia tytanomagnetytów są tym mniejsze, im więcej tytanu zawiera minerał. Maghemit charakteryzuje nieco mniejsze niż magnetyt namagnesowanie nasycenia $4 \cdot 10^{-2}$ T. W miarę wzrostu zawartości tytanu (czyli w minerałach szeregu tytanomaghemitowego) wartość tego momentu maleje. Hematyt, jako antyferromagnetyk z pasożytniczym ferromagnetyzmem, ma znacznie niższe namagnesowanie nasycenia, $0,5$–$2,5 \cdot 10^{-4}$ T. Najsilniej magnetyczne są te minerały z szeregu hemoilmenitowego, które zawierają 55–75% ilmenitu, tzw. ferriilmenity (ferrimagnetyki). Minerały magnetyczne charakteryzuje, podobnie jak wszystkie materiały o własnościach magnetycznych, pewna krytyczna temperatura, tzw. temperatura Curie w wypadku ferrimagnetyków lub temperatura Néela w wypadku antyferromagnetyków. Powyżej tej temperatury materiały stają się paramagnetykami. Temperatura Curie czystego magnetytu wynosi ok. 575°C i maleje ze wzrostem zawartości tytanu do ok. −200°C dla ulwospinelu. Temperatura Néela hematytu oraz Curie czystego maghemitu wynosi ok. 675°C. W miarę wzrostu zawartości tytanu odpowiednie temperatury minerałów obu szeregów maleją. Znajomość temperatury Curie (lub Néela) badanej skały pozwala stwierdzić, jakie minerały znajdują się w jej frakcji magnetycznej.

temperatura Curie i Néela

naturalna pozostałość magnetyczna

Pozostałość magnetyczną, którą uzyskują w czasie swojej historii geologicznej skały zawierające minerały magnetyczne, nazwano naturalną pozostałością magnetyczną — NRM (ang. *natural remanent magnetization*). NRM składa się na ogół ze składowej pierwotnej, uzyskanej przez skałę w okresie jej powstawania, i ze składowych wtórnych, powstałych w czasie historii geologicznej skały. W skałach magmowych składowa pierwotna NRM jest przeważnie pochodzenia termicznego; ten typ pozostałości powstaje podczas stygnięcia skały w ziemskim polu magnetycznym, od temperatur wyższych niż temperatura Curie (lub Néela) minerałów magnetycznych do temperatur niższych od tej wartości. Pierwotna pozostałość magnetyczna w skałach osadowych jest wynikiem osadzania w ziemskim polu magnetycznym ziaren minerałów magnetycznych. Każde ziarno ma już jakąś pozostałość magnetyczną, ziemskie pole jest tylko czynnikiem porządkującym.

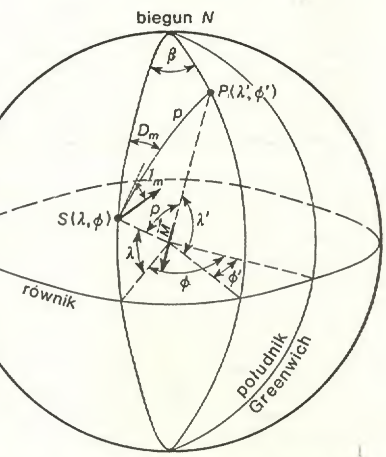
W wypadku skał przeobrażonych (zmetamorfizowanych) trudno mówić o określonym typie pozostałości. Historia tych skał jest zbyt skomplikowana, by można ją było odtworzyć na podstawie badań ich własności magnetycznych. Dlatego przy badaniach paleomagnetycznych na ogół nie bierze się ich pod uwagę.

Na pierwotną część naturalnej pozostałości magnetycznej nakładają się prawie zawsze składowe wtórne, które powstają w wyniku działania różnych czynników fizykochemicznych, jak np. gorące roztwory, długotrwałe podwyższenie temperatury, utlenianie, magnesujące działanie obecnego pola geomagnetycznego. Uzyskane w ten sposób wtórne pozostałości mają inne kierunki niż kierunek składowej pierwotnej i pierwszym krokiem w badaniach paleomagnetycznych jest

„oczyszczenie" NRM z tych składowych. W 1958 r. J. A. As i J.D.A. Zijderveld zastosowali po raz pierwszy metodę usuwania wtórnych składowych działając na próbki skał zmiennym polem magnetycznym o malejącej amplitudzie. W przeważającej większości



Rys. 10. Rozkład kierunków NRM po oczyszczeniu dla próbek skał bazaltowych z Ligoty Tułowickiej koło Opola. Położenie bieguna dla Ligoty Tułowickiej $\varphi' = 66°N$, $\lambda' = 185°E$. Każde kółeczko przedstawia kierunek trwałej pozostałości magnetycznej dla jednej próbki. Średni kierunek otrzymany dla całej kolekcji wynosi D_m : 183°, I_m : −44°. Krzyżykiem zaznaczono kierunek obecnego pola geomagnetycznego w miejscu pobrania próbek (wg J. Kruczyk, M. Kądziałko-Hofmokl, M. Jeleńska *Origin of Magnetic Minerals and Natural Remanent Magnetization of Basaltic Rocks from Lower Silesia, Poland*, Pure and Applied Geophysics, vol. 110, IX, 2031, 1973)



Rys. 11. Wyznaczanie współrzędnych geograficznych bieguna paleomagnetycznego P na podstawie średniego kierunku pozostałości magnetycznej D_m, I_m skał oraz współrzędnych geograficznych miejsca pobrania próbek S. λ długość geograficzna w punkcie S, φ szerokość geograficzna w punkcie S, λ' długość geograficzna bieguna, φ' szerokość geograficzna bieguna. $\vec{M}$ dipol magnetyczny Ziemi w okresie powstania badanych skał. Jak widać z rysunku współrzędne geograficzne bieguna — długość λ' i geograficzna szerokość φ' wiążą się z kierunkiem pola D_m, I_m czyli średnim kierunkiem pozostałości magnetycznej całej kolekcji oraz współrzędnymi geograficznymi miejsca pobrania próbek λ i φ następującymi wzorami:

$$\sin\lambda' = \sin\lambda\cos p + \cos\lambda\sin p\cos D \quad (-90° \leqslant \lambda' \leqslant +90°)$$

$$\varphi' = \varphi + \beta \text{ gdy } \cos p \geqslant \sin\lambda\sin\lambda'$$

lub

$$\varphi' = \varphi + 180° - \beta \text{ gdy } \cos p < \sin\lambda\sin\lambda',$$

gdzie

$$\sin\beta = \sin p\sin D/\cos\lambda' \quad (-90° \leqslant \beta \leqslant +90°),$$

p dopełnienie szerokości geograficznej miejsca pobrania próbek w okresie powstawania skały. p wiąże się z inklinacją I wzorem $\operatorname{tg} I = 2\operatorname{ctg} p$ (wg M. W. Mc Elhinny *Paleomagnetism and plate tectonics*, Cambridge 1973)

metody usuwania składowych wtórnych

wypadków okazało się, że składowe te są znacznie mniej stabilne niż składowa pierwotna. Metoda rozmagnesowywania zmiennym polem magnetycznym jest obecnie powszechnie stosowana. Jednocześnie korzysta się również z metody oczyszczania termicznego, która polega na grzaniu próbek do coraz wyższej temperatury i chłodzeniu do temperatury pokojowej. Metoda ta również prowadzi do usunięcia wtórnych składowych NRM.

Pozostała po oczyszczeniu najbardziej stabilna składowa NRM jest już na ogół składową pierwotną, a jej kierunek jest zgodny z kierunkiem pola geomagnetycznego w okresie, w którym powstała badana skała. Aby uzyskać możliwie wiarygodną informację o kierunku pola w epoce powstawania badanego kompleksu skalnego, trzeba dysponować kolekcją próbek liczącą co najmniej 10 okazów. Otrzymane po oczyszczeniu kierunki pozostałości magnetycznej próbek badanej kolekcji, czyli wartości kątów deklinacji i inklinacji, przedstawia się na kołowym wykresie, który jest rzutem jednej z półkul na płaszczyznę równika. Deklinacja zmienia się wzdłuż obwodu koła od N w kierunku ruchu wskazówek zegara od 0° do 360°, a inklinacja — wzdłuż promienia od 0° na obwodzie do 90° w środku koła. Rysunek 10 przedstawia otrzymaną w ten sposób grupę kierunków pozostałości magnetycznej kolekcji próbek skał bazaltowych z Ligoty Tułowickiej pod Opolem. Średni kierunek pierwotnej składowej NRM, obliczony ze wszystkich próbek kolekcji, jest przeważnie szukanym kierunkiem dawnego pola ziemskiego. Znając go, można odtworzyć położenie bieguna geomagnetycznego w epoce powstawania badanego kompleksu skalnego. Rysunek 11 przedstawia, jak znaleźć położenie bieguna paleomagnetycznego ówczesnego pola, znając średni kierunek pozostałości magnetycznej skał w danym miejscu, czyli kierunek pola geomagnetycznego, w którym skała uzyskała tę pozostałość.

Inwersja pola geomagnetycznego

Odkrycie skał namagnesowanych w kierunku przeciwnym niż obecny kierunek pola geomagnetycznego było bardzo ważne dla poznania przeszłości Ziemi oraz rozwoju teorii pochodzenia stałego pola ziemskiego. Zjawisko to zauważył po raz pierwszy B. Brunhes w 1906 r. Z czasem okazało się, że skały o odwrotnej naturalnej pozostałości magnetycznej występują równie często jak te, które mają normalną pozostałość (czyli zgodną z kierunkiem obecnego pola). Oba rodzaje NRM występują zarówno wśród skał magmowych, jak i osadowych. A zatem powstało pytanie, czy pozostałość magnetyczna skały może się odwrócić wskutek jakichś procesów fizycznych (samoodwrócenie), czy też pole magnetyczne zmieniało w przeszłości swój kierunek (inwersja pola).

samoodwrócenie pozostałości magnetycznej

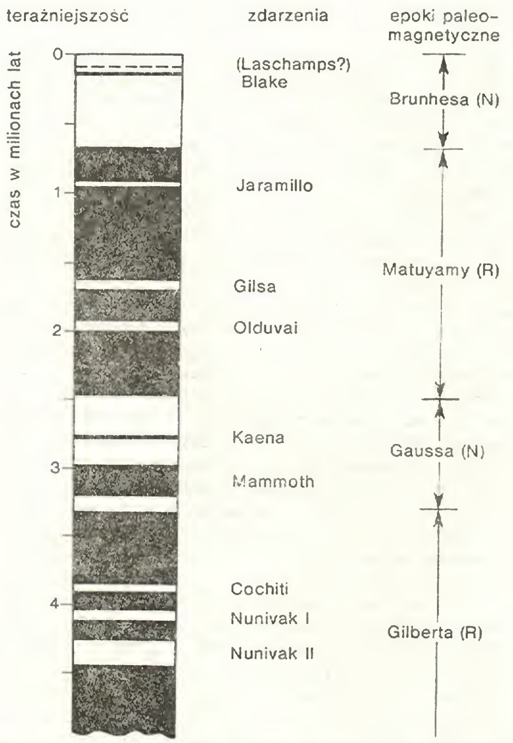
Pierwszą skałą, przy badaniu której zaobserwowano samoodwrócenie pozostałości magnetycznej, był dacyt z góry Haruna (Japonia). Stwierdzono, że skała ta uzyskała odwrotną NRM w normalnym polu ziemskim i że obserwuje się odwrócenie pozostałości magnetycznej tej skały w warunkach laboratoryjnych. L. Néel podał kilka możliwych wyjaśnień zjawiska samoodwrócenia. Istotę tego zjawiska bardzo dobrze tłumaczy najprostszy model. L. Néel zakłada, że w skale występują przerastające się wzajemnie fazy magnetyczne o różnych punktach Curie. Faza o wyższym punkcie Curie, stygnąc, ma w normalnym polu magnetycznym pozostałość o normalnym kierunku. Faza ta wytwarza w obszarze zajętym przez fazę drugą pole o kierunku przeciwnym i niekiedy natężenie tego pola może przewyższyć natężenie pola geomagnetycznego. Zatem druga faza, stygnąc do temperatur niższych niż jej temperatura Curie, uzyskuje pozostałość odwrotną. Jeżeli ta ostatnia dominuje, mamy do czynienia ze skałą o odwrotnej NRM. Jest to wypadek oddziaływania magnetostatycznego. Pozostałe me-

chanizmy Néela są bardziej skomplikowane i nie będą tu omawiane. Zjawisko samoodwrócenia zostało niejednokrotnie zaobserwowane, ale najczęściej przyczyn obserwowanej odwrotnej pozostałości trzeba szukać gdzie indziej. Przyjęto więc hipotezę, zgodnie z którą większość skał o odwrotnej naturalnej pozostałości magnetycznej uzyskała ją w polu ziemskim o biegunowości przeciwnej niż obecnie, czyli w polu, którego biegun północny znajdował się na geograficznej półkuli południowej, a południowy — na północnej.

Aby możliwie dokładnie poznać historię inwersji pola geomagnetycznego, korzysta się obecnie z wyników trzech różnych metod: z badań paleomagnetycznych dobrze datowanych potoków lawowych o dużej miąższości (grubości) oraz skał osadowych, z badań skał z rdzeni podmorskich otrzymanych metodą wierceń i z badań oceanicznych anomalii magnetycznych. Najlepiej znany jest najmłodszy okres życia Ziemi, czyli ostatnie 4,5 mln lat (wiek skał tego okresu można określić z największą dokładnością). Dzięki tym trzem metodom ustalono skalę geochronologiczną inwersji w tym okresie (rys. 12). Jak widać, oprócz długotrwałych okresów (epok paleomagnetycznych) o biegunowości normalnej (*N*) i odwrotnej (*R*) zaobserwowano szereg okresów krótszych, trwających około 10^5 lat, tzw. zdarzeń. Ostatnio stwierdzono, że w czasie trwającej obecnie normalnej epoki Brunhesa wystąpiły jeszcze krócej trwające inwersje, które nazwano wycieczkami bieguna geomagnetycznego. Wystąpiły one 18, 30 i 49 tysięcy lat temu.

wycieczki bieguna geomagnetycznego

Ze statystycznych opracowań danych paleomagnetycznych wynika, że w czasie inwersji biegunowość pola zmieniała się o 180° (w granicach błędu) i że okresy normalny i odwrotny są jednakowo prawdopodobne. Próbowano korelować okresy inwersji z innymi zjawiskami geofizycznymi, np. z gwałtownym wzrostem natężenia promieniowania kosmicznego, co z kolei powinno się odbić w zmianach fauny, ale badania nie potwierdziły tej hipotezy.



Rys. 12. Geochronologiczna skala inwersji dla ostatnich 4,5 milionów lat. Kolorem białym zaznaczono okresy o normalnej biegunowości, czarnym — o odwrotnej. Nazwy epok pochodzą od nazwisk, nazwy zdarzeń — od miejsc, w których stwierdzono okres o określonej biegunowości (wg D. H. Tarling *Principles and Applications of Paleomagnetism*, London 1971)

Interesujący jest okres przejściowy, kiedy to następuje zmiana kierunku pola. Obecnie nie ma jednolitej teorii co do przebiegu inwersji, ale najsłuszniejsze wydają się sugestie A. Chramowa. Według niego kierunek pola oscyluje kilkakrotnie między obydwoma stabilnymi położeniami, aż wreszcie ustala się w kierunku przeciwnym do poprzedniego. Wiele analiz wskazuje również na to, że natężenie pola w okresie przejściowym maleje do 0,1–0,3 swojej normalnej wartości, a długość trwania okresu przejściowego wynosi prawdopodobnie kilka–kilkanaście tysięcy lat.

Dryf kontynentów

Jeżeli byśmy wszystkie istniejące dane dotyczące położenia biegunów paleomagnetycznych potraktowali tak, jakby pole nie zmieniało zwrotu, to by się okazało, że bieguny w najmłodszej epoce życia Ziemi, czwartorzędzie (podział na epoki i okresy geologiczne znajduje się w tabeli), grupowały się wokół obecnego bieguna geomagnetycznego. Inaczej wyglądała sytuacja we wcześniejszych epokach geologicznych. Położenia biegunów paleomagnetycznych nie tylko ulegały przesunięciu w stosunku do położenia obecnego bieguna, ale i wyniki otrzymane z badań skał z różnych

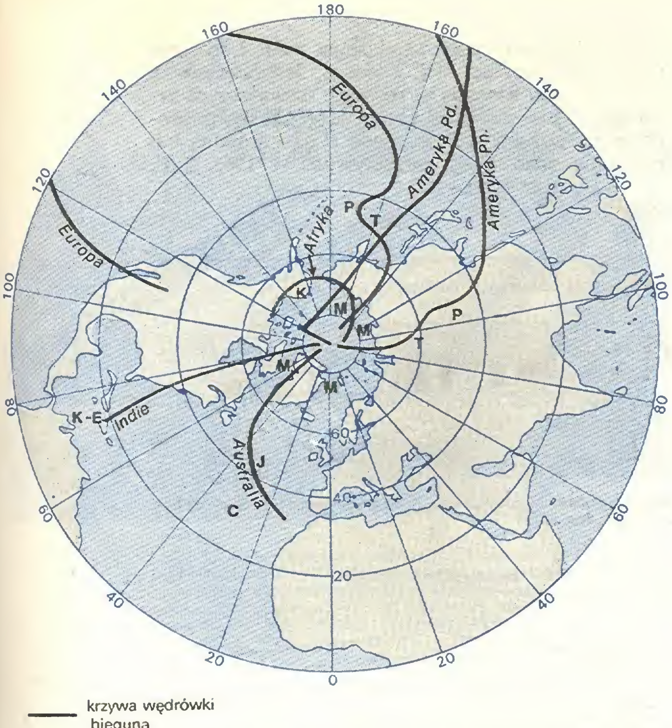
Skala geochronologiczna

Czas, mln lat	Epoka	Okres
0–1	kenozoik	czwartorzęd
		trzeciorzęd
1–13		pliocen
13–25		miocen
25–36		oligocen
36–58		eocen
58–63		paleocen
63–135	mezozoik	kreda
135–180		jura
180–230		trias
230–280	paleozoik	perm
280–345		karbon
345–405		dewon
405–425		sylur
425–500		ordowik
500–600?		kambr

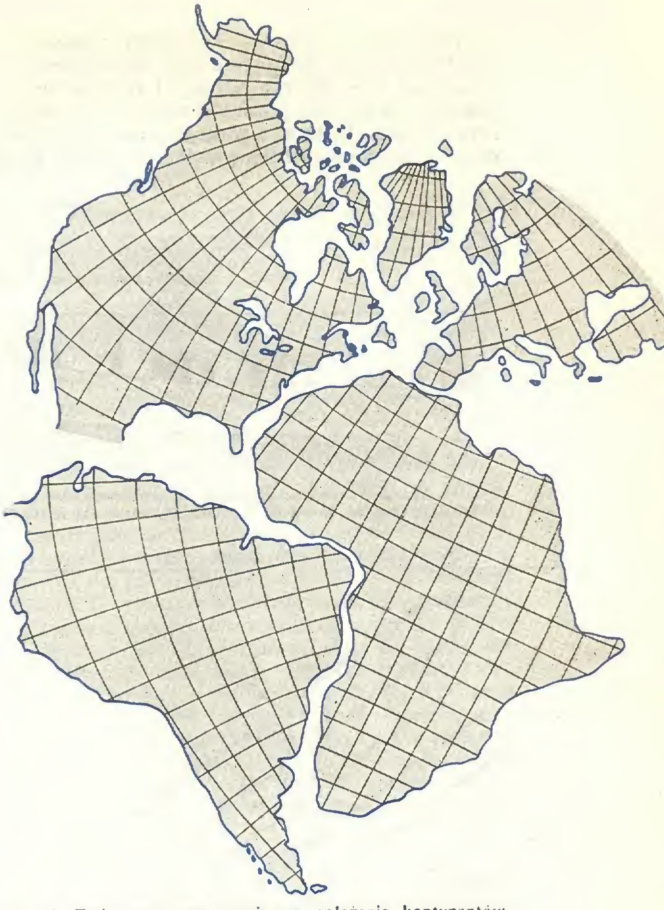
kontynentów różnią się między sobą. Łącząc ze sobą uśrednione położenia bieguna w kolejnych okresach geologicznych otrzymuje się tzw. krzywą wędrówki bieguna geomagnetycznego. Krzywe takie dla siedmiu obszarów (Ameryka Pn., Ameryka Pd., Afryka, Europa i pn. część Azji, Australia, Indie) przedstawia rys. 13. Należy dodać, że dane, na podstawie których sporządza się podobne syntezy, nie są ostateczne. W miarę zwiększania się liczby wyników i dokładności ich opracowania wprowadza się poprawki, czasem dość istotne. Mimo to prawdziwe pozostaje zjawisko przesuwania się w czasie bieguna geomagnetycznego względem skorupy ziemskiej oraz fakt, że ruch ten jest różny z różnych kontynentów. Próby takiego zestawienia danych, by bieguny z różnych obszarów układały się na jednej, wspólnej krzywej, zakończyły się przypomnieniem dawnej, liczącej sobie kilkadziesiąt lat hipotezy Wegenera, zgodnie z którą, w górnym paleozoiku istniał jeden ląd — Pangea, który w erze mezozoicznej i triasie podzielił się na kilka części. Obecnie, w wyniku szczegółowych badań geofizycznych i geologicznych, przyjmuje się raczej hipotezę du Toit wzajemnego ruchu (dryfu) kontynentów. Hipoteza du Toit mówi o istnieniu w paleozoiku nie jednego, jak u A. Wegenera, lecz dwóch prałądów. Jeden z nich, Gondwana, grupował obszary dzisiejszej Ameryki Pd., Afryki, Madagaskaru, Indii, Australii i Antarktydy; drugi, Laurazja, składał się z dzisiejszej Ameryki Pn., Grenlandii, Europy i Azji (bez Indii). Dane paleomagnetyczne wydają się potwierdzać hipotezę dwóch prałądów, które w mezozoiku zaczęły się dzielić na części odpowiadające wy-

krzywe wędrówki bieguna

hipoteza du Toit



Rys. 13. Krzywe wędrówki bieguna paleomagnetycznego otrzymane na podstawie wyników badań paleomagnetycznych skał z Ameryki Północnej, Ameryki Południowej, Europy i północnej części Azji, Indii, Australii, Afryki. Współcześnie wszystkie krzywe zbiegają się we wspólnym punkcie-biegunie, im starsza epoka geologiczna, tym bardziej są od siebie odległe (*K* kambr, *P* paleozoik, *J* jura, *C* kreda, *M* mezozoik, *E* eocen, *T* trzeciorzęd) (wg H. A. Cook *Physics of the Ecarth and Planets*, Londyn 1973)



Rys. 14. Zrekonstruowane wzajemne położenie kontynentów Ameryki Południowej i Afryki oraz Europy i Ameryki Północnej przed mezozoikiem (wg A. H. Cook *Physics of the Earth and Planets*, London 1973)

mienionym kontynentom, a te z kolei w miarę upływu czasu przesuwały się w kierunku swoich obecnych położeń. Jedną z prób rekonstrukcji dawnego wzajemnego ułożenia kontynentów przedstawia rys.14. Taki obraz uzyskano dopasowując numerycznie odpowiednie linie brzegowe. Dzięki przyjęciu hipotezy dryfu kontynentów udało się J.C. Bridenowi i A. Hallamowi w 1970 r. dopasować kontynenty w podobny sposób i uzyskać wspólną dla nich wszystkich krzywą wędrówki bieguna (rys. 15).

Z badań paleomagnetycznych wynika jeszcze jedna informacja dotycząca dryfu kontynentów: ruchy dryfowe nie odbywały się stale z tą samą prędkością. Dłuższe okresy kwazistatyczne (późny kambr–ordowik, dewon–dolny karbon, górny karbon–perm) były przerywane epizodami dryfowymi. Przypuszcza się, że nasileniu dryfu towarzyszyły ruchy górotwórcze. Na podstawie współcześnie prowadzonych badań astronomicznych dryfu Eurazji i Ameryki M. Feissel obliczył, że obecna prędkość rozsuwania się tych kontynentów wynosi $15\cdot10^{-7}$ stopni/rok (błąd oznaczenia $7\cdot10^{-7}$ stopni/rok).

Spekulacje dotyczące dryfu kontynentów pozostawały w sferze hipotez do ok. 1950 r. Dzięki wynikom badań paleomagnetycznych, w latach 1950–60 zaczęto tę hipotezę nazywać teorią, ale wciąż nie był znany mechanizm, który mógłby powodować ruchy tak wielkich bloków, jakimi są kontynenty. Dopiero od 1961 r., gdy przyjęto hipotezę rozsuwania się den oceanicznych, zaczęto wyjaśniać zjawiska dryfu poprzez zjawisko rozsuwania się den.

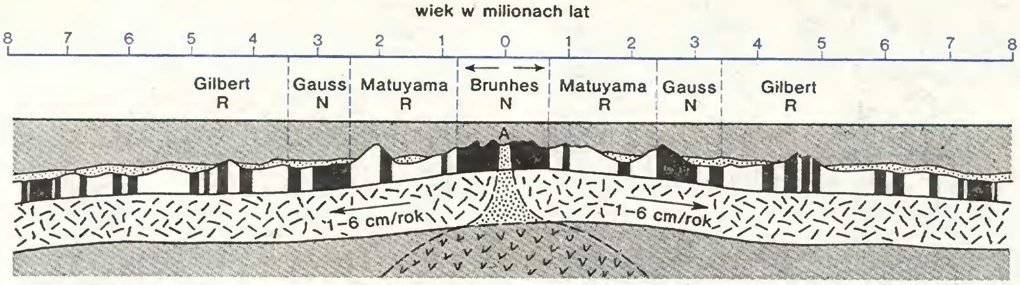
rozsuwanie się den oceanicznych

W 1958 r. po raz pierwszy zauważono, że na obszarze Oceanu Spokojnego, a następnie na oceanach Atlantyckim i Indyjskim, występują anomalie magnetyczne w postaci pasów anomalii na przemian ujemnych i dodatnich, o amplitudach przeważnie $2\cdot10^{-7}$ T–$4\cdot10^{-7}$ T i szerokości kilkudziesięciu kilometrów. Pasy tworzące „zebrowaty" obraz układają się równolegle do grzbietów oceanicznych, symetrycznie po obu ich stronach. Taki obraz anomalii zaobserwowanych

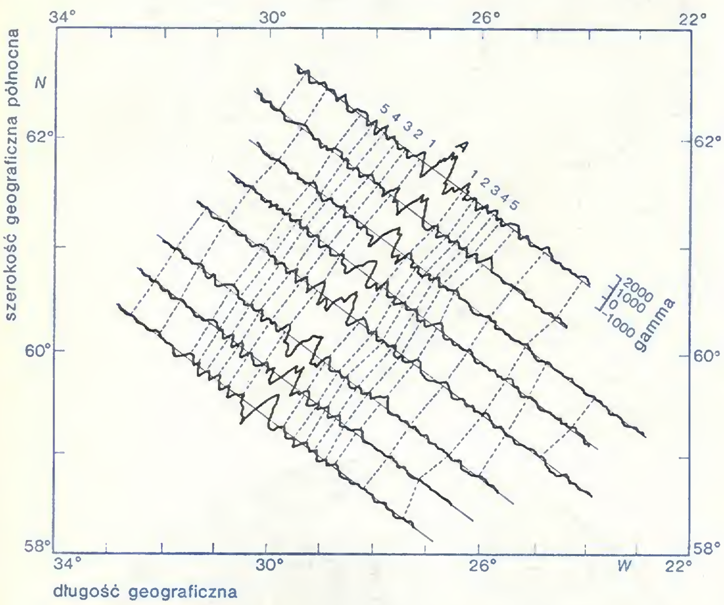


Rys. 15. Wspólna krzywa wędrówki bieguna południowego dla okresu od syluru *S* do mezozoiku *M* poprzez perm *P* i karbon *C* (gruba krzywa przerywana oznacza przypuszczalne drogi bieguna dla Gondwany i Laurazji przed ich połączeniem się w sylurze. Cienkie linie przerywane — kierunki rozdzielania się kontynentów począwszy od mezozoiku) (wg M. W. Mc Elhinny *Paleomagnetism and plate tectonics*, Cambridge 1973)

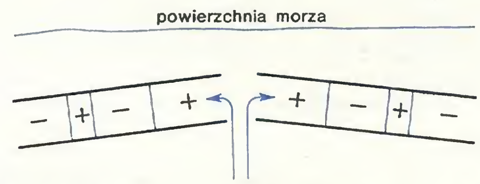
po obu stronach grzbietu Reykjanes przedstawiają rys. 16, 17, 18, 19. Po upływie kilku lat od pierwszych obserwacji R.S. Dietz, a później H.H. Hess wyrazili pogląd, że w strefach grzbietów tworzy się nowa skorupa ziemska. Materiał z górnego płaszcza Ziemi wydobywa się na powierzchnię wzdłuż grzbietów. Wskutek nacisku dotychczasowa skorupa pęka i rozsuwa się na boki, robiąc miejsce nowej. Jednocześnie ze stwierdzeniem istnienia w rejonach oceanicznych zebrowatych anomalii zauważono, że próbki skał pochodzące z obszarów tych anomalii mają normalną lub odwrotną pozostałość magnetyczną. W 1963 r. F.



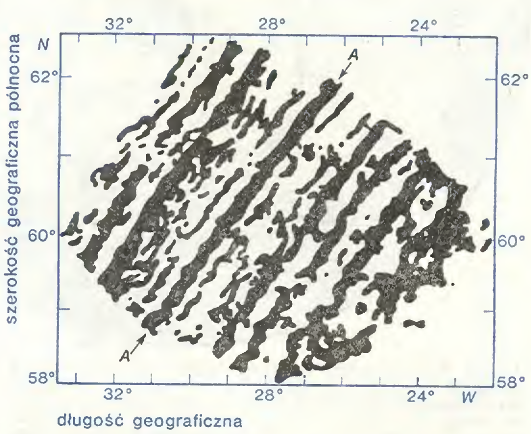
Rys. 16. Morskie anomalie magnetyczne zaobserwowane po obu stronach grzbietu oceanicznego Reykjanes (na południowy zachód od Islandii) (rys. 16–19 wg A. H. Cook *Physics of the Earth and Planets*, London 1973)



Rys. 17. Zmiany natężenia pola geomagnetycznego zarejestrowane wzdłuż kilku profili biegnących prostopadle do linii grzbietu. Grzbiet jest zaznaczony literą *A*. Numery *1...5* oznaczają kolejne anomalie dodatnie (powyżej linii zerowej odpowiadającej średniej wartości pola w badanym rejonie) i ujemne (poniżej tej linii). Z boku przedstawiono skalę wartości natężenia pola wyrażonych w gammach (1 γ = 1 nT). Linie kreskowane (niebieskie) łączą te same anomalie występujące na różnych profilach



Rys. 19. Schemat powstawania zebrowatego obrazu anomali dodatnich (+) i ujemnych (–) po obu stronach grzbietu oceanicznego. Wzdłuż grzbietu wydobywa się nowy materiał skorupy oceanicznej (zaznaczony strzałkami)



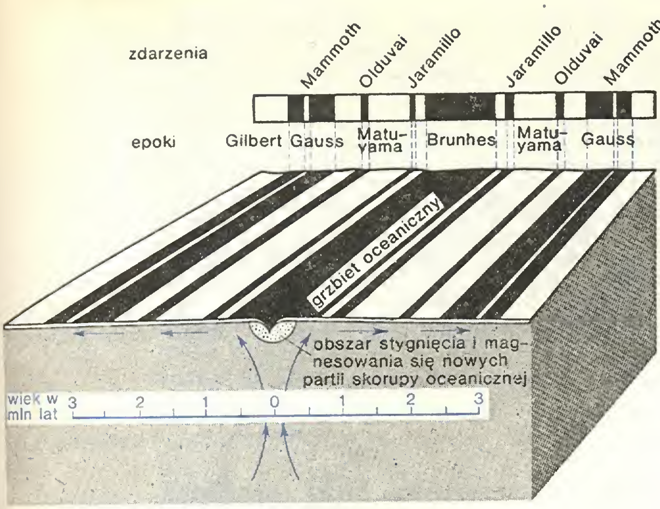
Rys. 18. Anomalie przedstawione w postaci pasów czarnych (dodatnie) i białych (anomalie ujemne), *A* grzbiet Reykjanes

Vine i D.M. Matthews uzupełnili hipotezę Dietza i Hessa sugerując, że obserwowane zjawisko łączy się z inwersjami pola geomagnetycznego. Ta część skorupy, która powstała w okresie, gdy pole miało normalny kierunek, jest źródłem dodatniej anomalii i próbki skał z tego obszaru mają normalną pozostałość magnetyczną. Odpowiednio, anomalie ujemne są dowodem na to, że w okresie powstawania tych części skorupy pole geomagnetyczne miało kierunek odwrotny. W wyniku szczegółowych badań anomalii w rejonach grzbietów oceanicznych, próbek z rdzeni podmorskich i próbek skał kontynentalnych udało się stwierdzić korelację pomiędzy anomaliami i poszczególnymi epokami paleomagnetycznymi (rys. 20). Oszacowano również prędkości rozsuwania się dna w różnych obszarach kuli ziemskiej (1–6 cm/rok), a zatem — poznano historię dryfu kontynentów do ok. 200 milionów lat wstecz.

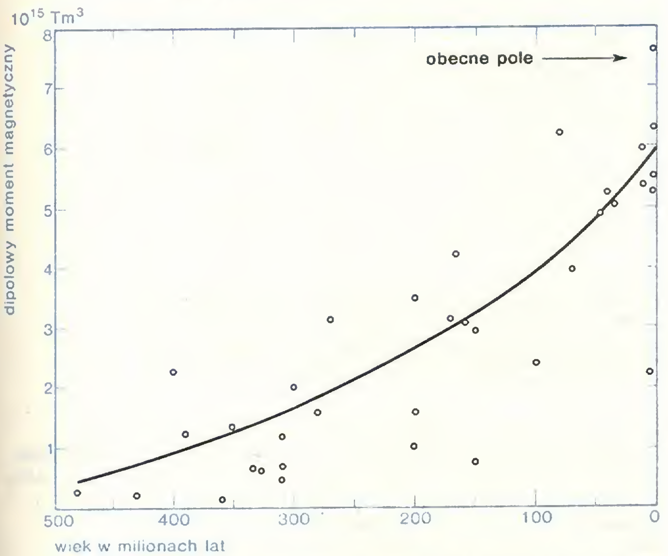
Omówione w skrócie hipotezy Dietza–Hessa i Vine–Matthewsa są traktowane jako potwierdzenie teorii dryfu kontynentów; zwolennicy tej teorii widzą mechanizm wywołujący dryf w procesach rozsuwania się den oceanicznych. Koncepcja dryfu kontynentów ma wśród geofizyków wielu przeciwników, ale jak dotychczas nie ma lepszej teorii wyjaśniającej dane dostarczane przez paleomagnetyzm. Wnioski wypływające z badań paleomagnetycznych znajdują potwierdzenie w wynikach badania klimatów w różnych epokach geologicznych (paleoklimatów).

hipotezy wyjaśniające wędrówkę bieguna

Pozostaje do wyjaśnienia, co oznacza krzywa wędrówki bieguna. I ten problem nie został jeszcze ostatecznie rozwiązany. Istnieje kilka hipotez próbujących tłumaczyć to zjawisko. Wydaje się, że najbardziej prawdopodobna jest hipoteza wychodząca z założenia, że Ziemia nie zachowuje się jak ciało sztywne, lecz płynie. W stanie równowagi oś obrotu Ziemi pokrywa się z osią jej głównego momentu bezwładności. Gdy wskutek ruchu kontynentów zmienia się rozkład mas w skorupie ziemskiej, wówczas zmienia się również moment bezwładności. Ziemia powraca do stanu równowagi wówczas, gdy się przesuwa jej oś obrotu, a zatem położenia biegunów geograficznych i geoma-



Rys. 20. Schematyczna korelacja morskich anomalii magnetycznych z epokami paleomagnetycznymi. Anomaliom dodatnim odpowiadają epoki o normalnym kierunku pola (pasy czarne), anomaliom ujemnym — epoki o kierunku odwrotnym (pasy białe) (wg M. W. Mc Elhinny *Paleomagnetism and plate tectonics*, Cambridge 1973)



Rys. 21. Zmiany momentu magnetycznego Ziemi w ciągu ostatnich 500 milionów lat (wg D. H. Tarling *Princples and Applications of Paleomagnetism*, London 1971)

gnetycznych. Szybkość tego ruchu zależy od lepkich własności płaszcza, które nie są jeszcze w pełni zbadane. Omówione tu pokrótce zjawiska, sugerujące przesuwanie się kontynentów bądź ich części względem siebie i osi obrotu Ziemi, zostały ujęte w jedną gałąź wiedzy o Ziemi, nazwaną tektoniką płyt lub tektoniką kier. Dane paleomagnetyczne w ich obecnej postaci, podobieństwo linii brzegowych niektórych kontynentów, szereg danych geologicznych wskazują na słuszność hipotez stanowiących podstawę tej dziedziny. Wciąż jednak brak opisu mechanizmów, które by mogły wywoływać procesy wymagane przez tektonikę płyt. Być może nowe dane geofizyczne zmuszą do skorygowania dzisiejszych teorii i wskażą na zupełnie odmienną interpretację obserwowanych dotychczas zjawisk.

tektonika płyt

Badania paleomagnetyczne pozwalają poznać nie tylko zmiany kierunku pola geomagnetycznego w odległych epokach geologicznych, ale i zmiany jego natężenia.

Na podstawie dostępnych obecnie danych można sądzić, że moment magnetyczny Ziemi miał wartość podobną do dzisiejszej 2,6 miliarda lat temu, następnie malał i osiągnął minimum 500 milionów lat temu; w ciągu ostatnich 400 milionów lat wartość momentu wzrastała, jak to widać na rysunku 21.

Archeomagnetyzm

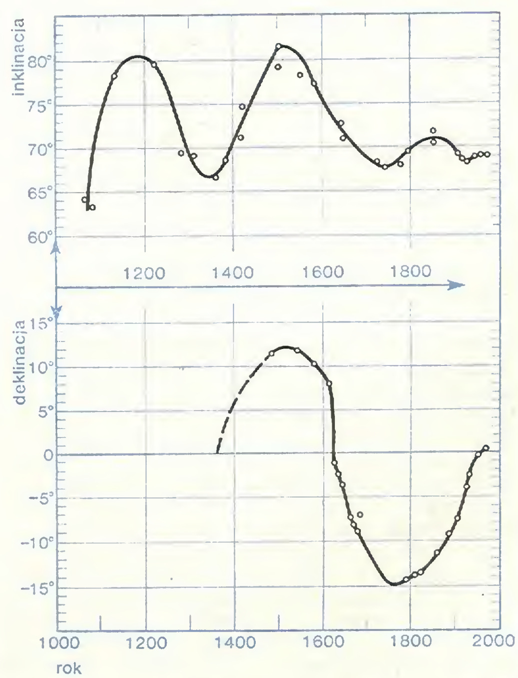
Jak już wspomniano, lukę pomiędzy obecnymi obserwacjami pola geomagnetycznego a danymi paleomagnetycznymi wypełnia archeomagnetyzm. Archeomagnetyzm zajmuje się badaniem zmian pola ziemskiego w przeszłości prehistorycznej i historycznej. Przedmiotem badań są wypalane gliny (cegły, paleniska, wyroby ceramiczne), które podczas procesu wypalania uzyskują stabilną pozostałość magnetyczną w kierunku istniejącego ówcześnie pola geomagnetycznego.

badanie zmian pola

Pomiary wartości i kierunku NRM dobrze datowanych próbek pozwoliły na odtworzenie zmian wiekowych pola geomagnetycznego do 5500 lat wstecz. Zgromadzone dotychczas dane wskazują, że zmiany natężenia pola w tym okresie nie przekraczają 15% jego średniej wartości, amplituda zmian kierunku pola wynosiła ok. 20°, a okres tych zmian — ok. 1000 lat. Rysunek 22 przedstawia zmiany inklinacji i deklinacji otrzymane z badań archeomagnetycznych dla Gdańska. Zmiany inklinacji obejmują okres od 1080 r. do chwili obecnej. Linią ciągłą zaznaczono zmiany deklinacji otrzymane z pomiarów obejmujących okres od 1360 r. do chwili obecnej, a linią przerywaną — hipotetyczną zmianę ekstrapolowaną tak, by była podobna do krzywej doświadczalnej. Wyniki pozwalają sądzić, że zmiany wiekowe pola geomagnetycznego w Gdańsku w czasie objętym badaniami przebiegały z okresem ok. 600 lat.

zastosowanie w archeologii

Badania archeomagnetyczne znajdują praktyczne zastosowanie w archeologii; pozwalają na datowanie fragmentów wypalanych glin, których wiek nie jest dokładnie znany, z dokładnością co najmniej 25 lat, a stosowane metody nie powodują zniszczenia badanych obiektów.



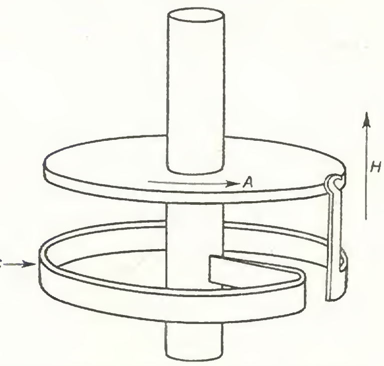
Rys. 22. Zmiany inklinacji w Gdańsku od 1080 r. i zmiany deklinacji w Gdańsku od 1360 r. do chwili obecnej (wykorzystano tu także dane przytoczone na rys. 6). Krzywa ciągła — wyniki pomiarów, krzywa przerywana — hipotetyczna ekstrapolacja (wg W. Czyszek *Archaeomagnetism: A preliminary report*, Materiały i Prace Inst. Geof. PAN, vol. 76, 59, 1974)

Teoria pochodzenia stałego pola geomagnetycznego i zmian wiekowych

Źródeł stałego pola geomagnetycznego obserwowanego obecnie, w czasach historycznych i w dawnych epokach geologicznych należy szukać we wnętrzu Ziemi. Źródłem stałego pola dipolowego mogłaby być np. jednorodnie namagnesowana kula ziemska o namagnesowaniu $75 \cdot 10^{-7}$ T lub jednorodnie namagnesowane jądro Ziemi o namagnesowaniu $49 \cdot 10^{-6}$ T. Mógłby to być również pewien stacjonarny układ prądów elektrycznych płynących we wnętrzu Ziemi. Jednak żaden z tych mechanizmów nie mógłby spowodować ani zjawiska zmian wiekowych, ani inwersji pola geomagnetycznego. Jedyną istniejącą obecnie teorią, która, jak się wydaje, potrafi wyjaśnić wszystkie cechy stałego pola, jest zaproponowana w latach czterdziestych tzw. teoria samowzbudnego dynamo Elsassera–Bullarda. Aby ją opisać, trzeba powiedzieć parę słów o budowie wnętrza Ziemi. Badania fal sejsmicznych przechodzących przez głębokie (poniżej 3000 km) warstwy Ziemi pozwoliły stwierdzić, że wewnętrzna część kuli ziemskiej, o promieniu ok. 3500 km, jądro, zachowuje się inaczej niż jej partie zewnętrzne. Ustalono, że jądro składa się z dwóch warstw, z których jedna, zewnętrzna, jest ciekła, natomiast wewnętrzna jest prawdopodobnie stała (→ Fizyka skorupy i wnętrza Ziemi). Dane dotyczące gęstości i składu chemicznego jądra wskazują, że dominuje tam żelazo i nikiel, a zatem powinno być ono dobrym przewodnikiem elektryczności. Te własności jądra leżą u podstaw działania wewnątrz Ziemi mechanizmu Elsassera–Bullarda. Rysunek 23 przedstawia schematycznie zasadę działania samowzbudnego

budowa wnętrza Ziemi

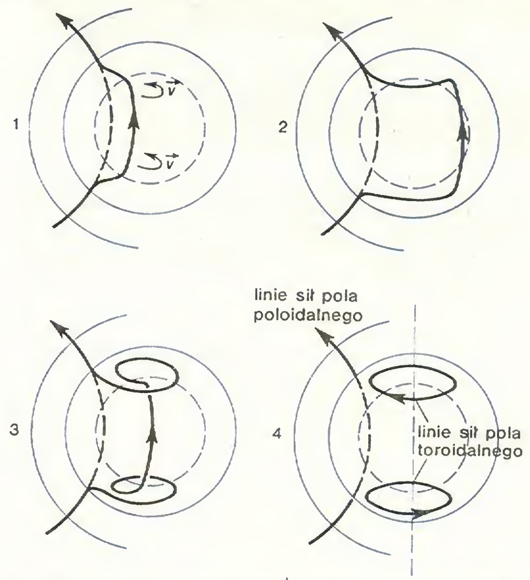
teoria samowzbudnego dynamo



Rys. 23. Schemat działania samowzbudnego dynamo Elsassera–Bullarda

dynamo. Tarcza *A* wykonana z dobrego przewodnika elektryczności obraca się w kierunku wskazanym strzałką. W otoczeniu układu istnieje niewielkie pole magnetyczne skierowane wzdłuż osi obrotu tarczy. Tarcza jest połączona z osią przez cewkę, przedstawioną na rysunku jako jeden zwój *C*. Podczas obrotu tarczy w zewnętrznym polu magnetycznym indukuje się w niej siła elektromotoryczna skierowana wzdłuż jej promienia. Przez cewkę *C* zaczyna płynąć prąd, który z kolei jest źródłem dodatkowego pola magnetycznego, równoległego do osi obrotu tarczy. Przy odpowiednio dobranym kierunku ruchu to nowo powstałe pole dodaje się do istniejącego już pola *H*. Wzrost natężenia pola powoduje wzmocnienie siły elektromotorycznej w tarczy, a więc i natężenia prądu płynącego przez cewkę — i tak dalej. A zatem taki układ może spowodować znaczne wzmocnienie początkowego pola magnetycznego o bardzo małym natężeniu.

We wnętrzu Ziemi rolę obracającej się tarczy przewodzącej i cewki odgrywają prądy konwekcyjne płynące w ciekłym, przewodzącym jądrze. Jednakże same tylko ruchy konwekcyjne nie wystarczyłyby do wytworzenia układu dynamo; poruszające się do góry i na dół cząstki cieczy tworzyłyby zamknięte pętle tylko w płaszczyźnie południkowej. Dzięki sile Coriolisa, wywołanej ruchem obrotowym Ziemi, „rurki" prądów konwekcyjnych wyciągają się wzdłuż płaszczyzn równoleżnikowych, tworząc odpowiednik tarczy. Przebieg procesu przedstawia schematycznie rys. 24. Niezbędne do funkcjonowania dynamo początkowe pole magnetyczne (w wypadku Ziemi jest to składowa słabego pola panującego w przestrzeni okołoziemskiej) jest skierowane wzdłuż osi obrotu Ziemi.



Rys. 24. Schemat powstawania toroidalnego pola w obracającej się przewodzącej kuli. Cyfry od 1 do 4 oznaczają kolejne etapy tego procesu, $\vec{v}$ wektor prędkości obrotu Ziemi, kierunek obrotu jest zaznaczony strzałkami (rys. 23, 24 wg M. W. Mc Elhinny, *Paleomagnetism and plate tectonics*, Cambridge 1973)

Jest to tzw. pole poloidalne, które nie ma składowej w płaszczyźnie równoleżnikowej. Linie sił tego pola poruszają się razem z ciekłym materiałem jądra, są w nim niejako zamrożone. W ten sposób przez „wyciągnięcie" linii sił pola poloidalnego powstaje składowa równoleżnikowa pola, tzw. pole toroidalne, wzmacniające prądy elektryczne, co z kolei prowadzi do wzmocnienia pola poloidalnego obserwowanego na powierzchni Ziemi.

pole poloidalne i toroidalne

Z przedstawionej tu schematycznie teorii dynamo samowzbudnego wynika, że stałe pole geomagnetyczne powstaje w wyniku wzajemnego oddziaływania ruchów cieczy przewodzącej i pola elektromagnetycznego. Dziedzina zajmująca się takimi zjawiskami nosi nazwę magnetohydrodynamiki, a fale, których opis zawiera zarówno elementy hydrodynamiki, jak i elektrodynamiki, nazywają się falami magnetohydrodynamicznymi. Równania opisujące działanie ziemskiego dynamo są bardzo skomplikowane i nie będziemy ich tu przytaczać. Próby ich rozwiązania, przy różnych założeniach dotyczących warunków termodynamicznych panujących we wnętrzu Ziemi, doprowadziły do wielu ciekawych wniosków. Stwierdzono np., że prędkość obrotu zewnętrznych partii jądra jest mniejsza niż jego partii wewnętrznych. Zjawisko to może być odpowiedzialne za dryf zachodni pola niedipolowego i centrów zmian wiekowych. Istnienie tych centrów wiąże się z występowaniem na powierzchni jądra układów prądów elektrycznych.

Aby wyjaśnić możliwość zmiany kierunku pola, T. Rikitake zaproponował w 1958 r. przyjęcie dwóch sprzężonych układów dynamo. W wyniku prac nad

układem równań opisujących ten model okazało się, że przy pewnych założeniach (co do warunków panujących we wnętrzu Ziemi) prądy elektryczne występujące w równaniach oscylują wokół stanu stacjonarnego przez jakiś czas, przy czym w miarę upływu czasu rośnie amplituda tych oscylacji. W pewnym momencie następuje gwałtowny przeskok do oscylacji wokół innego stanu stacjonarnego. A zatem prąd elektryczny, a w konsekwencji pole magnetyczne wytwarzane przez układ Rikitake, może w pewnych warunkach zmienić swój kierunek. Fizyczna natura impulsów, które by mogły spowodować taki proces, nie jest jeszcze znana, ale możliwość oscylacji prądów pozostaje w zgodzie z obserwowanymi na powierzchni Ziemi oscylacjami pola geomagnetycznego w okresach przejściowych pomiędzy normalnym i odwrotnym kierunkiem pola.

model dwóch układów dynamo

Zmienne pole geomagnetyczne

Zmienne pole magnetyczne, które stanowi tylko niewielki procent pola obserwowanego na powierzchni Ziemi, pochodzi od źródeł znajdujących się poza Ziemią. Ta część pola, mimo że tak mała, jest przedmiotem ciągłych obserwacji oraz analiz i przynosi informacje dotyczące nie tylko pola magnetycznego Ziemi, ale i własności przestrzeni okołoziemskiej. Wyniki ciągłych obserwacji gromadzone od szeregu lat w postaci magnetogramów pozwoliły na rozróżnienie kilku typów zmian pola. Na rys. 25 przedstawiony jest magnetogram otrzymany w Centralnym Obserwatorium Geofizycznym Polskiej Akademii Nauk w Belsku Dużym.

W 1722 r. londyńczyk G. Graham, obserwując przez mikroskop igłę kompasu, zauważył, że wykonuje ona pewne ruchy w płaszczyźnie poziomej. Są to pierwsze obserwacje zmian deklinacji; zmiany te były czasem regularne i niezbyt szybkie, kiedy indziej znów nieregularne i gwałtowne. Nieco później wspólnie z A. Celsiusem, który podobne obserwacje prowadził w Uppsali, stwierdzili, że często w obu miastach jednocześnie panował magnetyczny spokój lub pojawiało się zaburzenie.

Początkowo obserwowano tylko zmiany deklinacji magnetycznej. Dopiero po upływie ok. 100 lat od odkrycia G. Grahama F. Gauss wykonał w Getyndze pierwsze pomiary inklinacji magnetycznej i natężenia całkowitego wektora pola. Wkrótce potem pod patronatem F. Gaussa rozpoczęła działalność Unia Magnetyczna, powołana w celu dokonywania w latach 1836–1841 jednoczesnych pomiarów magnetycznych w istniejących ówcześnie 50 obserwatoriach. Następne okresy intensywnych badań zjawisk magnetycznych na powierzchni Ziemi to Pierwszy Rok Polarny (1882–1883), Drugi Rok Polarny (1932–1933) i Międzynarodowy Rok Geofizyczny (1957–1958).

W Polsce pierwsze obserwatorium magnetyczne założył w 1910 r. Stanisław Kalinowski w Świdrze pod Warszawą. Działalność tej placówki została w ostatnich latach zakłócona przez biegnącą w pobliżu linię kolei elektrycznej. Obecnie ciągłą rejestrację składowych pola geomagnetycznego na terenie Polski prowadzą: Obserwatorium Geofizyczne Polskiej Akademii Nauk na Półwyspie Helskim w Helu oraz Centralne Obserwatorium Geofizyczne Polskiej Akademii Nauk w Belsku Dużym (woj. radomskie). Poza tym rejestrację magnetyczną o charakterze uzupełniającym prowadzi Obserwatorium Sejsmologiczne Polskiej Akademii Nauk w Raciborzu.

obserwatoria magnetyczne w Polsce

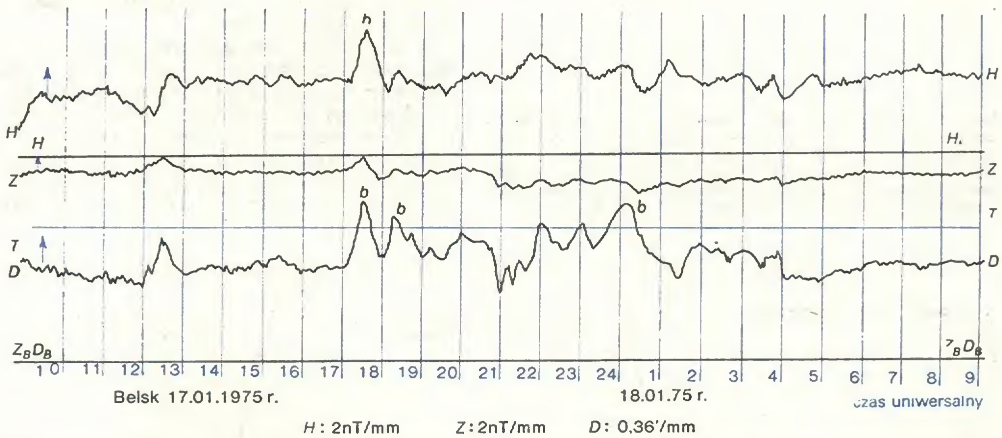
Analiza danych obserwacyjnych, gromadzonych dzięki pracy obserwatoriów magnetycznych, pozwoliła na rozdzielenie pola zmiennego na część zmieniającą się spokojnie (gładko) i na część zawierającą zmiany zachodzące w sposób nieuporządkowany. Zmiany płynne — to tzw. zmiany spokojne lub niezaburzone, zmiany nieuporządkowane — to tzw. zmiany zaburzone.

Na magnetogramie zmiany spokojne charakteryzują gładkie odcinki krzywych przebiegających poniżej lub powyżej średniej wartości danej składowej pola w miejscu obserwacji, zaznaczonej na rys. 25 jako linia bazy. Wśród zmian spokojnych poszczególnych składowych można wydzielić zmiany o okresach: dobowym, sezonowym, rocznym, jedenastoletnim. Zmiany spokojne o okresie dobowym oznacza się symbolem S_q i przypisuje się je zmieniającemu się z okresem dobowym wpływowi Słońca. Zmiany S_q są różne, zależnie od szerokości geograficznej miejsca obserwacji i pory roku. Na rys. 26 przedstawione są te zmiany na różnych szerokościach geograficznych w trzech sezonach dla składowej poziomej, deklinacji i inklinacji. W naszych szerokościach geograficznych zmiana spokojna deklinacji przebiega w następujący sposób: W pierwszej ćwiartce doby liczonej od godziny 0 czasu miejscowego południk magnetyczny przesuwa się powoli na wschód i osiąga maksimum wschodnie w godzinach 7–9 czasu miejscowego. Następnie zawraca szybko ku zachodowi, maksymalne położenie zachodnie osiąga ok. godziny 13[30]. Potem kieruje się znów na wschód i ok. godziny 22[30] osiąga drugie wschodnie maksimum.

zmiany spokojne

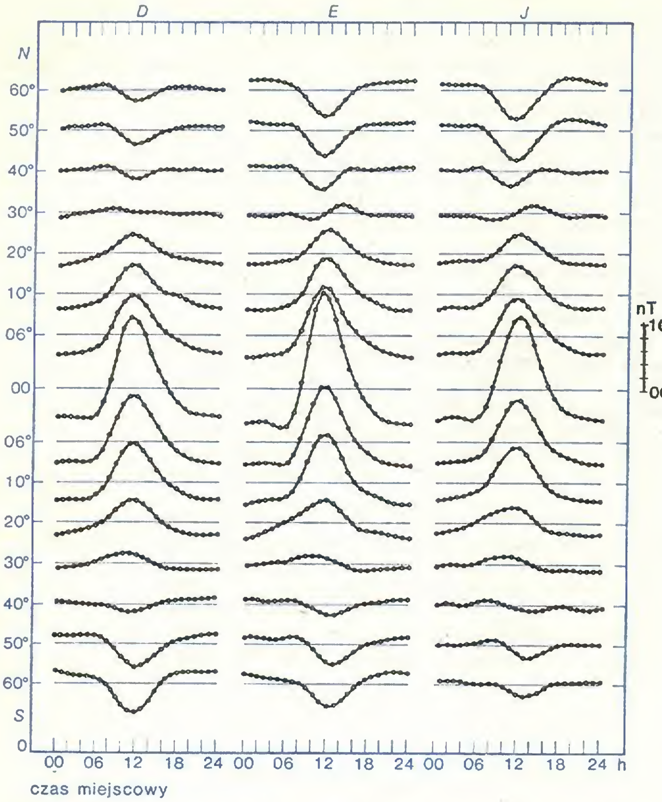
Oprócz spokojnej zmiany słonecznej istnieje również spokojna zmiana księżycowa, wywołana wpły-

przykład magnetogramu



Rys. 25. Magnetogram przedstawiający bieg dobowy składowej poziomej H, składowej pionowej Z i deklinacji D otrzymany w Belsku 17.I.–18.I.1975 r. Linie H_B, Z_B i D_B linie baz poszczególnych składowych, linia T bieg dobowy temperatury w pomieszczeniu, w którym jest prowadzona rejestracja. Na wykresach H i D widoczne są zmiany zatokowe b

wem Księżyca, o okresie doby księżycowej, oznaczana symbolem L i zaobserwowana po raz pierwszy przez K. Kreila w 1850 r. Szczegółowe badania zmian L



Rys. 26. Uśrednione dla całej kuli ziemskiej zmiany S_q składowej poziomej dla różnych szerokości geograficznych dla trzech sezonów: D (styczeń, luty, listopad, grudzień), E (marzec, kwiecień, wrzesień, październik), J (maj, czerwiec, lipiec, sierpień) (wg S. Matsushita *Solar quiet and lunar daily variations fields*, w *Physics of Geomagnetic Phenomena*, vol. I, New York 1967

wykazały, że zależą one od szerokości geograficznej miejsca obserwacji i od pory roku, podobnie jak zmiany S_q. Maksymalne wartości zmian księżycowych składowych H i Z wynoszą $1\text{–}2 \cdot 10^{-9}$ T, a deklinacji — 40″. Zmiany S_q są kilkadziesiąt razy silniejsze.

Źródeł powstania zmian spokojnych, zarówno słonecznych, jak i księżycowych, upatruje się w występowaniu w warstwie E jonosfery (czyli 100–120 km nad powierzchnią Ziemi) prądów elektrycznych o szczególnej konfiguracji, związanych ze zjawiskami pływowymi, zależnymi odpowiednio od Słońca i Księżyca.

zmiany zaburzone

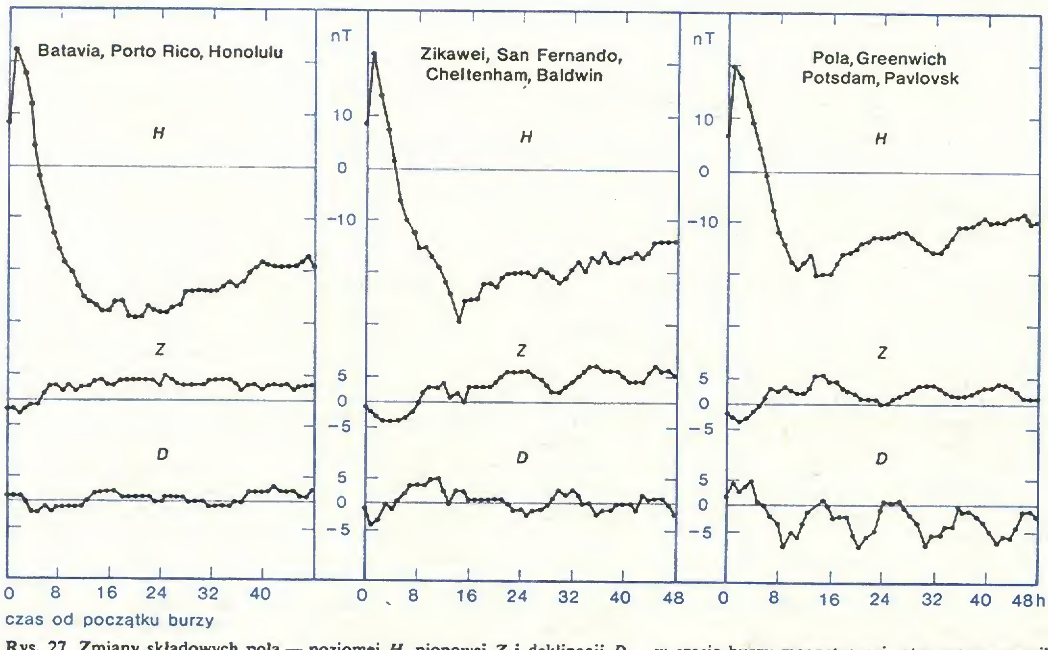
Zmiany zaburzone pola geomagnetycznego składają się z szeregu nakładających się na siebie elementów, które można poklasyfikować w następujący sposób: zmiany periodyczne — zaburzone dobowe zmiany słoneczne o okresie doby słonecznej i pulsacje magnetyczne, których okres wynosi kilka minut, zmiany zatokowe, które na magnetogramie mają kształt zatoki, zmiany aperiodyczne, które występują w czasie burz magnetycznych przede wszystkim na składowej poziomej. Główną część zaburzeń stanowią nieprawidłowe fluktuacje, które się składają z szeregu następujących po sobie zmian o różnych okresach i amplitudach.

zorze polarne

Burzom magnetycznym towarzyszy niejednokrotnie zjawisko zórz polarnych, szczególnie na dużych szerokościach geograficznych. Zjawisko to polega na świeceniu atomów tlenu i cząsteczek azotu znajdujących się w atmosferze. Zorze przybierają różne kształty: spokojnych lub pulsujących łuków, wstęg, welonów, promienistych koron. W tym ostatnim wypadku promienie układają się wzdłuż linii sił pola geomagnetycznego, co zauważył po raz pierwszy J.C. Wilcke w 1770 r. Związek zórz polarnych ze zjawiskami magnetycznymi został stwierdzony właśnie dzięki jego obserwacjom.

burze magnetyczne

A. Humboldt, badając burze magnetyczne, stwierdził, że najsilniejsze z nich zaczynają się niemal równocześnie na całej Ziemi i że powodują zmniejszanie się składowej poziomej pola, która następnie rośnie powoli do poprzedniej wartości. Charakterystyczną cechą większości burz jest nagłość ich pojawiania się. Pierwszy, krótki impuls zwiększa natężenie składowej poziomej. Ten wzrost trwa krótko, od $1/2$ godziny do

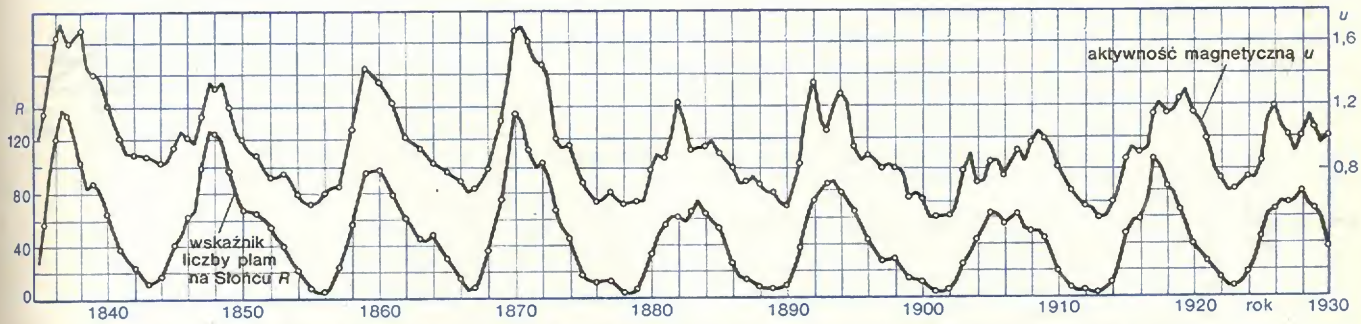


Rys. 27. Zmiany składowych pola — poziomej H, pionowej Z i deklinacji D — w czasie burzy magnetycznej, otrzymane w wyniku uśrednienia zmian zaobserwowanych podczas 40 burz dla trzech grup obserwatoriów: w niskich szerokościach geomagnetycznych (a), w średnich szerokościach (b) i w wysokich szerokościach (c) (wg S. Chapman *Solar emission and magnetic-auroral storms on the Earth* w *Magnetism and Cosmos*, 3, Edinburgh, London 1965)

godziny. Zmniejszanie się natężenia składowej H trwa kilka godzin, a powrót do stanu normalnego może trwać kilka dni. W czasie burzy amplituda zmian pola może osiągnąć kilka tysięcy nanotesli. Na rys. 27 przedstawiono zmiany H, Z i D podczas burz magnetycznych o średnim natężeniu.

aktywność magnetyczna

Miarą zaburzenia magnetycznego jest wprowadzona sztucznie wielkość zwana aktywnością magnetyczną. Jeżeli jakiś czas pole się nie zmienia, to aktywność wynosi 0. Aktywność jest tym większa, im większa jest amplituda zmian pola i im szybsze są jego zmiany. W 1939 r. Międzynarodowe Towarzystwo Magnetyzmu Ziemskiego i Elektryczności wprowadziło dla określenia stopnia zaburzenia tzw. indeks K, wyrażony w ballach od $K = 0$ do $K = 9$. Każdemu ballowi odpowiada amplituda wahań składowych pola geomagnetycznego w trzygodzinnych odstępach czasu — z uwzględnieniem poprawki na spokojne zmiany dobowe. Za amplitudę przyjmuje się tu maksymalną różnicę między największym i najmniejszym odchyleniem jednej ze składowych (H, Z lub D) od normalnej krzywej biegu dobowego w trzygodzinnym przedziale czasu. Ponieważ poziom zaburzeń zależy od szerokości geograficznej miejsca obserwacji, więc w każdym obserwatorium przyjmuje się nieco inne amplitudy zaburzeń przy tych samych wartościach indeksu K, np. w Centralnym Obserwatorium Geofizycznym w Belsku Dużym przyjmuje się indeks $K = 9$ dla amplitudy 500 nT. Wspólny dla całej Ziemi, tzw. planetarny indeks K, otrzymuje się uśredniając indeksy K z obserwatoriów położonych między 50° i 63° szerokości geomagnetycznej. Poza tym istnieje wiele innych miar aktywności magnetycznej. Jedna z nich, tzw. miara u, uwzględnia wpływ burz magnetycznych. Wyznaczenie tego wskaźnika polega na znalezieniu różnicy między średnią wartością składowej poziomej podczas dwóch kolejnych dni, bez uwzględniania znaku składowej. Aktywność magnetyczna pozostaje w ścisłym związku z aktywnością Słońca; widać to wyraźnie na rys. 28, gdzie przedstawione są krzywe zmian obu miar aktywności (aktywność Słońca określa się tzw. liczbą Wolffa, czyli wskaźnikiem liczby plam na Słońcu).



Rys. 28. Aktywność magnetyczna (miara u) i wskaźnik liczby plam na Słońcu R dla okresu od 1834 do 1930 roku (wg S. Chapman *Solar emission and magnetic-auroral storms on the Earth*, w *Magnetism and Cosmos*, 3, Edinburgh, London 1965)

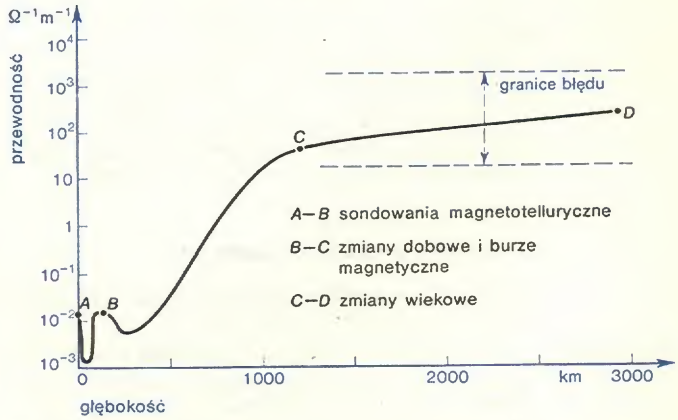
Indukcja elektromagnetyczna we wnętrzu Ziemi

prądy telluryczne

Zmiany pola geomagnetycznego o okresach od kilku sekund do kilku tysięcy lat indukują prądy elektryczne w przewodzących warstwach Ziemi. Prądy te, zwane prądami tellurycznymi, są źródłem wtórnych składowych pola, kóre się dodają do składowych pochodzących od źródeł zewnętrznych. Rozkład i natężenie prądów tellurycznych zależą od częstości zmian oraz od rozkładu przewodności we wnętrzu Ziemi; zmiany o okresach krótszych indukują prądy w płytszych warstwach. Natężenie zewnętrznego, pierwotnego pola geomagnetycznego można przedstawić analitycznie jako szereg funkcyjny o wyrazach zależnych od współrzędnych geograficznych miejsca obserwacji oraz od promienia ziemskiego a i odległości miejsca obserwacji od środka Ziemi r w postaci czynników $(r/a)^n$. W wypadku wewnętrznej, wtórnej składowej pola ta ostatnia zależność jest inna i ma postać $(a/r)^{n+1}$. Dzięki tej różnicy można rozdzielić obserwowane na powierzchni Ziemi zmiany pola na części pochodzące od źródeł zewnętrznych i od prądów tellurycznych, a następnie otrzymać rozkład przewodności elektrycznej we wnętrzu Ziemi. Znaleziony w ten sposób globalny model rozkładu przewodności pozwala rozróżnić następujące warstwy: dobrze przewodzącą warstwę przypowierzchniową o kilkukilometrowej miąższości, następnie źle przewodzącą warstwę o miąższości ok. 250–300 km, poniżej której przewodność rośnie o dwa rzędy wielkości. Zmiany pola geomagnetycznnego o okresach nie przekraczających roku pozwalają na poznanie rozkładu przewodności do głębokości ok. 1000 km. O przewodności głębszych warstw można sądzić na podstawie prądów indukowanych przez zmiany wiekowe pola. Bardziej szczegółowy obraz rozkładu przewodności na mniejszych głębokościach, tzn. 10–150 km, daje metoda sondowań magnetotellurycznych zapoczątkowana w 1949 r. przez A. Tichonowa i L. Cagnarda. Polega ona na jednoczesnych obserwacjach zmian składowych poziomych pola magnetycznego i elektrycznego. Wyniki badań wskazują na duże regionalne odstępstwa od globalnego modelu.

metoda sondowań magnetotellurycznych

Rysunek 29 przedstawia rozkład przewodności elektrycznej z głębokością uzyskany na podstawie uśrednionych wyników sondowań magnetotellurycznych, wyników badań prądów indukowanych przez zmiany dobowe pola geomagnetycznego, burze magnetyczne i zmiany wiekowe. Zjawiska indukcji elektromagne-



Rys. 29. Rozkład przewodności elektrycznej we wnętrzu Ziemi (wg M. H. P. Bott *The Interior of the Earth*, London 1971)

tycznej we wnętrzu Ziemi przyciągają uwagę geofizyków ze względu na ścisły związek pomiędzy rozkładem przewodności elektrycznej a rozkładem temperatury we wnętrzu Ziemi. Więcej informacji na ten temat można znaleźć w artykule poświęconym fizyce skorupy i wnętrza Ziemi.

A. N. Chramow *Paleomagnietizm paleozoja*, Leningrad 1974; A. H. Cook *Physics of the Earth and Planets*, London 1973; B. Lisowski *O zmianach dziennych słonecznych deklinacji magnetycznej w Świdrze*, Acta Geophys. Pol. 3, 35, 1955; S. Matsushita *Solar quiet and lunar daily variation fields*, w *Physics of Geomagnetic Phenomena*, vol. 1, New York and London 1967; M. W. Mc Elhinny *Paleomagnetism and plate tectonics*, Cambridge 1973; A. T. Price *Electromagnetic introduction within the Earth*, w *Physics of Geomagnetic Phenomena*, vol. 1, New York and London 1967; T. Przypkowski *Zabytkowe kompasy magnetyczne na instrumentarium astronomicznym Marcina Bylicy z Olkusza z lat 1480–1487*, Acta Geophys. Pol. 4, 245 (1956); U. Schmucker, J. Jankowski *Geomagnetic induction studies and the electrical state of the upper mantle*, w *The Upper Mantle Developments in Geotectonics*, 4, Amsterdam 1972; R. Sroczyński *Rozwój eksperymentu, pojęć i teorii magnetycznych*, Warszawa 1969.

Fizyka morza

Dynamika morza

Czesław Druet

Około 70% powierzchni ziemskiego globu zajmują morza i oceany, olbrzymie masy wód, które są w ciągłym ruchu. Formy tego ruchu zależą od rodzaju czynnika zewnętrznego, wywołującego na powierzchniach granicznych zbiornika wodnego (powierzchnia swobodna wód i powierzchnia dna) oraz w jego wnętrzu, hydrostatyczne i hydrodynamiczne siły tarcia i ciśnienia. Woda jest ośrodkiem łatwo odkształcalnym i praktycznie rzecz biorąc — nieściśliwym. Jej struktura określona jest gęstością, wyrażającą stosunek masy m elementu cieczy do jego objętości q. W zbiornikach wodnych, których masa rozłożona jest równomiernie (struktura jednorodna), gęstość jest wielkością stałą $\varrho = m/q =$ const. Natomiast w dużych zbiornikach oceanicznych rozkład gęstości wody jest z reguły rozkładem niejednorodnym, zależnym od temperatury, zasolenia i ciśnienia. Na pewnych głębokościach, wskutek znacznych różnic gęstości warstw wodnych, tworzą się często tzw. powierzchnie skoku gęstości.

gęstość wody

Niezależnie od stanu zasolenia mórz i oceanów procentowy udział poszczególnych rodzajów soli w 1 kg wody jest wielkością prawie stałą i znając zawartość jednego ze składników (np. chlorku sodu) w 1 kg wody morskiej, możemy obliczyć przybliżony stan jej zasolenia. Stan ten waha się — w zależności od szerokości geograficznej, głębokości akwenu — od kilku do kilkudziesięciu promili.

zawartość soli

Temperatura wód morskich zależy również od szeregu czynników, spośród których najistotniejsze są promieniowanie słoneczne oraz mechanizmy wymiany i przenoszenia ciepła wewnątrz zbiornika. Rozkład temperatury wód morskich jest znacznie mniej stabilny od rozkładu zasolenia i zależy od szerokości geograficznej, głębokości akwenu oraz pory roku, miesiąca, doby. W zależności od tych czynników średnia temperatura wód morskich przybiera wartości od zera do kilkudziesięciu stopni Celsjusza.

temperatura

Wpływ ciśnienia atmosfery i górnych warstw wody na gęstość wody morskiej jest niewielki, ponieważ — jak powiedzieliśmy — woda jest ośrodkiem prawie nieściśliwym. Gdyby jednak wody morskie były idealnie nieściśliwe, wówczas poziom światowego oceanu byłby o ok. 30 m wyższy od poziomu obecnego.

Przyczyny ruchu mas wodnych

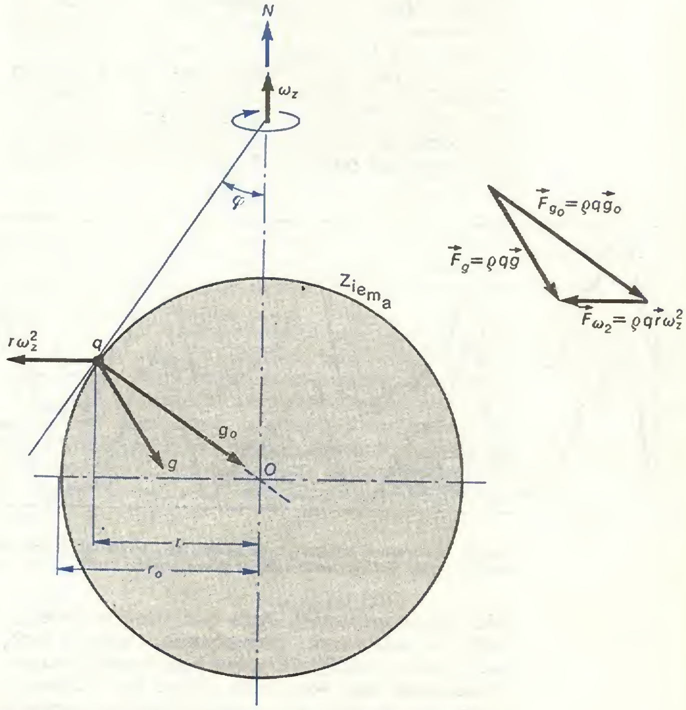
Stałymi czynnikami zewnętrznymi, od których zależy ruch mas wodnych w morzu są:

siła ciężkości

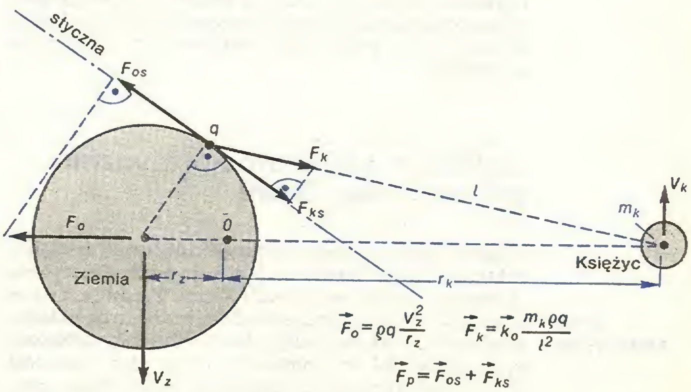
Siła ciężkości $\vec{F_g}$, będąca wypadkową dwóch składowych (rys. 1) — siły grawitacji ziemskiej $\vec{F_{g_0}}$, wywołanej zjawiskiem wzajemnego przyciągania się masy Ziemi i elementu wody q, oraz siły odśrodkowej $\vec{F_{\omega_z}}$, wywołanej obrotem Ziemi dookoła jej osi.

siła pływotwórcza

Siła pływotwórcza $\vec{F_p}$, będąca wypadkową rzutów dwóch sił na linię styczną do krzywizny swobodnej powierzchni wód, w punkcie położenia elementu q (rys. 2) — siły odśrodkowej ($\vec{F_o}$), wywołanej ruchem



Rys. 1. Rozkład sił i przyspieszeń działających na element q, zlokalizowany na powierzchni wirującej Ziemi; $\vec{g}$ jest rzeczywistym przyspieszeniem ziemskim



Rys. 2. Rozkład sił międzyplanetarnego oddziaływania w układzie Ziemia–Księżyc, działających na element q, położony na powierzchni Ziemi

okrężnym Ziemi dookoła środka („O") grawitacyjnej równowagi układu Ziemia–Księżyc (wszystkie punkty materialne na Ziemi i w jej wnętrzu krążą w tym ruchu po orbitach kołowych względem różnych środków obrotu O, ale promienie tych orbit r oraz liniowe prędkości ruchu ($\vec{V} = \omega_0 r$) punktów q są jednakowe, równe $r = r_z$ i $\vec{V} = \vec{V_z}$ (rys. 2)), oraz rzutu siły grawitacji księżycowej ($\vec{F_k}$) wywołanej zjawiskiem wzajemnego przyciągania się elementów wody i Księżyca. Analogiczne siły pływotwórcze wzbudzane są również w układzie grawitacyjnym Ziemia–Słońce. Obliczając siły $\vec{F_p}$ w układzie Ziemia–Księżyc i układzie Ziemia–Słońce zobaczymy, że siły pływotwórcze księżycowe są 2,2 razy większe od sił pływotwórczych słonecznych.

siła Coriolisa

Na elementy wody q poruszające się w akwenie z prędkością postępową $\vec{u}$ działa siła wywołana ruchem obrotowym globu ziemskiego dookoła własnej osi, zwana siłą Corliolisa $\vec{F_c} = 2\varrho q(\vec{u} \times \vec{\omega_z})$ (gdzie $\vec{u} \times \vec{\omega_z}$ jest iloczynem wektorowym, czyli wektorem o kierunku prostopadłym do kierunku $\vec{u}$ i $\vec{\omega_z}$, o długości równej $u\omega_z \sin(\sphericalangle \vec{u}, \vec{\omega_z})$. Siła ta odchyla kierunek wektora prędkości $\vec{u}$, na półkuli północnej zgodnie z ruchem wskazówek zegara (w prawo), a na półkuli południowej w kierunku przeciwnym (w lewo).

Pozostałe czynniki zewnętrzne, od których zależy ruch mas wodnych w morzu związane są z własnościami powierzchni granicznych zbiornika, tj. z warunkami panującymi na swobodnej powierzchni wód oraz na dnie i brzegach morskiego zbiornika.

ciśnienie atmosferyczne i aerodynamiczne tarcie

Atmosfera ziemska oddziałuje na swobodną powierzchnię akwenu siłą normalną do tej powierzchni, zależną od ciśnienia atmosferycznego i siłą styczną do niej zwaną siłą tarcia aerodynamicznego. Ciśnienie atmosferyczne (p_a), zmieniające się nieregularnie w czasie i przestrzeni, zależy głównie od warunków termicznych panujących w atmosferze na różnych wysokościach. Natomiast styczne siły tarcia aerodynamicznego ($\vec{\tau_a}$) wzbudzane są ruchem mas powietrza (wiatrem) w przywodnej warstwie atmosfery i wynikają z turbulentnej wymiany pędu między elementami powietrza i wody w przypowierzchniowej granicznej warstwie morza i atmosfery. Siły te są proporcjonalne do kwadratu prędkości wiatru ($\tau_a \approx 3 \cdot 10^{-6} U^2$, gdzie U jest prędkością wiatru).

kształt i dynamiczna aktywność akwenu

Drugą naturalną powierzchnię graniczną morskiego zbiornika tworzą dno i brzegi akwenu. Ich kształt i aktywność dynamiczna wywołują różnego rodzaju procesy brzegowe lub hydrodynamiczne osobliwości. Kształt i rodzaj brzegu morskiego oraz jego stromość wywierają znaczny wpływ na procesy hydrodynamiczne w strefie przybrzeżnej, transformując istniejące i wzbudzające nowe (np. fale odbite od stromych brzegów lub przybrzeżne spiętrzenia mas wodnych). Kształt dna morskiego odgrywa także dużą rolę w przydennych przepływach mas wodnych oraz związanej z nimi dyfuzji pędu, ciepła i zawartych w wodzie substancji biernych. Ale najpoważniejsze w skutkach są zjawiska wulkaniczne i trzęsienia dna morskiego, wywołujące groźne dla życia i mienia ludzkiego powodzie morskie.

Prawa rządzące ruchem mas wodnych

Ruch mas wodnych w morzach i oceanach, zależny od wyżej wymienionych czynników, przybiera w czasie i przestrzeni różne formy, spełniające zawsze trzy podstawowe prawa mechaniki ośrodków ciągłych: prawo zachowania masy,

$$\iiint_q \frac{\partial \varrho}{\partial t} dq + \iint_A \varrho u_n dA = 0;$$

prawa zachowania

prawo zachowania pędu,

$$\frac{d}{dt} \iiint_q \varrho \vec{u} dq = \iint_A \vec{p_n} dA + \iiint_q \varrho \vec{F} dq;$$

prawo zachowania energii,

$$\iiint_{q_2} \frac{\varrho u^2}{2} dq_2 - \iiint_{q_1} \frac{\varrho u^2}{2} dq_1 = $$

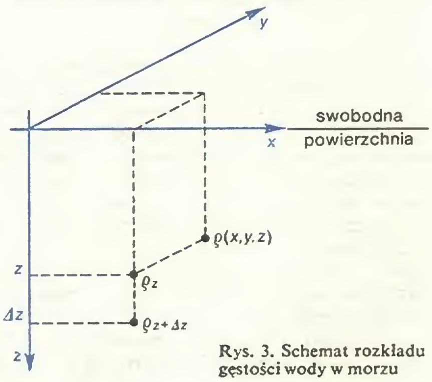
$$= \int_{t_1}^{t_2} \left[\iint_A \vec{p_n} \vec{u} dA + \iiint_q \varrho \vec{F} \vec{u} dq \right] dt,$$

gdzie ϱ jest gęstością wody zawartej w elementarnej objętości dq, A — powierzchnią objętości q, $\vec{u}$ — prędkością ruchu elementu dq, t — czasem, $\vec{p_n}$ — ciśnieniem (siłą powierzchniową normalną do elementu powierzchni dA), u_n — prędkością normalną do elementu powierzchni dA, $\vec{F}$ — siłą masową zewnętrzną, działającą na element objętości dq.

Prawo zachowania masy wyraża fakt, że zmiana masy w objętości q, uwarunkowana zmianą gęstości wody w czasie dt, musi równać się masie, która w tym samym czasie dopłynie do wnętrza objętości q lub wypłynie z niego przez powierzchnię A. Jeżeli zasada ta nie jest spełniona to oznacza, że wewnątrz obszaru q istnieją źródła produkujące lub pochłaniające wodę. Prawo zachowania pędu odwzorowuje podstawową zasadę Newtona, dotyczącą równowagi sumy sił zewnętrznych działających na ciało (siły powierzchniowe i siły masowe — prawa strona równania) i iloczynu masy ciała oraz jego przyspieszenia (lewa strona równania). Prawo zachowania energii wyraża zasadę, że ilość pracy wykonanej przez siły zewnętrzne w czasie $t_2 - t_1$ musi być równa różnicy energii kinetycznej masy wody w dwóch stanach chwilowych; w momencie t_2, w którym masa cieczy ma objętość q_2 i momencie t_1, odpowiadającym objętości q_1.

stan równowagi hydrostatycznej

Stanem spoczynkowym morskiego akwenu jest stan równowagi hydrostatycznej, w którym elementy wody q nie zmieniają w czasie i przestrzeni swojego położenia. Wówczas jedyną siłą zewnętrzną działającą na elementy wody jest siła ciężkości warunkująca ciśnienie hydrostatyczne. Stan równowagi hydrostatycznej możliwy jest oczywiście jedynie wówczas, gdy rozkład gęstości wody w morzu jest bądź rozkładem jednorodnym w przestrzeni: $\varrho(x, y, z) =$ $=$ const, bądź też rozkładem jednorodnym w płaszczyźnie poziomej: $\varrho(x, y) =$ const i stabilnym w płaszczyźnie pionowej: $\varrho_z < \varrho_{z+\Delta z}$ (rys. 3). W warunkach naturalnych stan taki nie może wystąpić

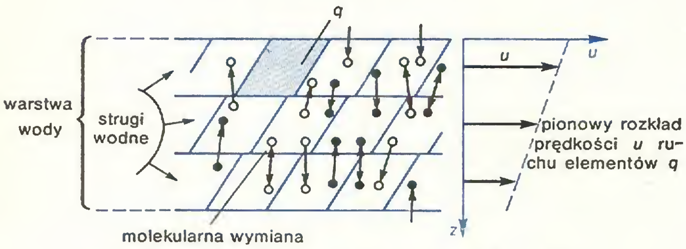


Rys. 3. Schemat rozkładu gęstości wody w morzu

w czystej postaci, ponieważ astronomiczne siły pływotwórcze oraz przestrzenna i czasowa zmienność pól atmosferycznego ciśnienia, wzbudzać będą zawsze ruch mas wodnych i jedynie w szczególnych okolicznościach (bardzo słabe pływy w małym akwenie, jednorodny i stacjonarny układ baryczny, brak wiatru) i tylko chwilowo mogą być spełnione w akwenie warunki kwazihydrostatycznej równowagi.

ruch laminarny

W razie naruszenia gęstościowej jednorodności czy stabilności ośrodka wodnego lub zadziałania innego czynnika, naruszającego stan równowagi hydrostatycznej akwenu, elementy wody rozpoczną ruch, który w fazie początkowej, bez względu na rodzaj czynnika wzbudzającego, będzie ruchem laminarnym. W ruchu tym poruszające się elementy wody *q* pozostaną zawsze w niezniennej względem siebie pozycji, a jedynym mechanizmem przekazywania energii ze źródła ruchu w głąb i wszerz zbiornika będzie molekularna wymiana pędu (wymiana pędu między molekułami) między sąsiadującymi elementami wody (rys. 4), stanowiąca istotę molekularnego tarcia wewnętrznego, zwanego lepkością wody.



Rys. 4. Schemat struktury laminarnego ruchu wody

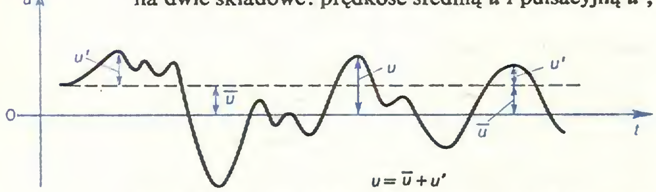
ruch turbulentny

Jeżeli prędkość elementów *q* przekroczy pewną wartość graniczną, ruch laminarny wody przechodzi w ruch turbulentny, który charakteryzuje się chaotyczną wymianą elementów wody między strugami. Elementy te poruszają się wówczas po nieregularnych krzywoliniowych drogach, tworząc w ten sposób różnorodne struktury wirowe, których wymiary (skale turbulencji) zmieniają się w oceanie od kilku centymetrów do kilku tysięcy kilometrów, w zależności od rodzaju czynnika wzbudzającego ruch. Tego typu wymiana pędu między sąsiadującymi warstwami mas wodnych stanowi istotę tzw. turbulentnego tarcia wewnętrznego.

Siły molekularnego i turbulentnego tarcia wewnętrznego i odpowiadające im naprężenia styczne do powierzchni elementów wody, łącznie z siłami ciśnienia hydrodynamicznego, działającego prostopadle do powierzchni tych elementów i wywołującego naprężenia normalne, warunkują procesy przekazywania energii mechanicznej mas wodnych wewnątrz morskiego akwenu.

Rodzaje zjawisk hydrodynamicznych w morzach i oceanach

Różnorodność czynników zewnętrznych wzbudzających i warunkujących ruch mas wodnych w morzu oraz różnoskalowy charakter ich oddziaływania powodują powstawanie złożonego mechanizmu przekazywania energii zewnętrznej w głąb morskiego zbiornika, od większych ustrojów wirowych do mniejszych. Mechanizm ten tworzy pole przepływu, którego struktura jest wynikiem przypadkowego (losowego) nakładania się i przenikania różnych ruchów składowych. Podstawową wielkością charakteryzującą taki przepływ jest prędkość chwilowa *u* elementów wody, która zmienia się w sposób przypadkowy w czasie i przestrzeni. Prędkość tę możemy zawsze rozłożyć na dwie składowe: prędkość średnią $\bar{u}$ i pulsacyjną u',



Rys. 5. Wycinek oscylogramu prędkości turbulentnego przepływu

charakteryzującą turbulencję (rys. 5). W zależności od wielkości przedziału czasu lub obszaru przestrzeni, w których wypadkowy ruch mas wodnych jest uśredniany, parametry $\bar{u}$ i u' mogą mieć różne znaczenie i wartości (ruch makroskalowy, mezoskalowy, mikroskalowy).

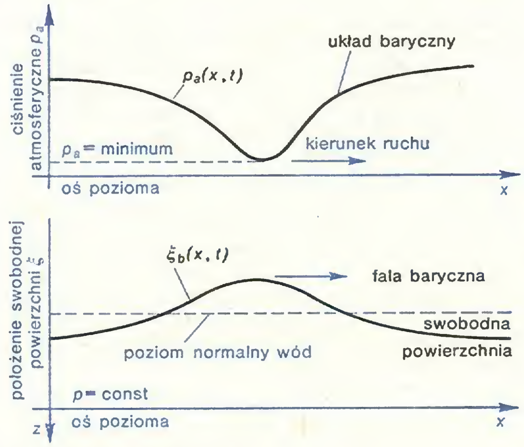
Formy mikroskalowe charakteryzują ruch elementów wody bezpośrednio i nie podlegają żadnemu podziałowi. Natomiast wielko- i średnioskalowe ruchy mas wodnych w morzu możemy podzielić na różne rodzaje. W zależności od przyjętego kryterium podziału może to być ruch oscylacyjny, postępowy, cyrkulacyjny itp., lub ruch swobodnej powierzchni morza, ruch powierzchni skoku gęstości, termohalinowa cyrkulacja, przepływy geostroficzne, dryfowe, gradientowe itp.

Scharakteryzujmy pokrótce najbardziej znane postacie zjawisk hydrodynamicznych w morzach i oceanach.

A) Ruch falowy

fale baryczne

Jeżeli poziomy rozkład ciśnienia atmosferycznego (układ baryczny) będzie niejednorodny, to zgodnie z zasadą naczyń połączonych, powierzchnia swobodna jednorodnego gęstościowo akwenu nie będzie miała położenia poziomego, lecz przybierze kształt równoważący ciśnienie atmosferyczne tak, że na dowolnej rzędnej poniżej powierzchni swobodnej ciśnienie będzie stałe (rys. 6). Jeżeli ponadto pole baryczne będzie się zmieniać w czasie i przemieszczać w określonym kierunku, to procesowi temu będzie towarzyszyć odpowiedni ruch odkształcenia swobodnej powierzchni morza zwany falą baryczną. Fale baryczne należą do kategorii fal nieokresowych, wymuszonych. Ich rozmiary i prędkość rozprzestrzeniania się zależą od charakterystyki układów barycznych wzbudzających te odkształcenia oraz od głębokości i rozmiarów morskiego akwenu.

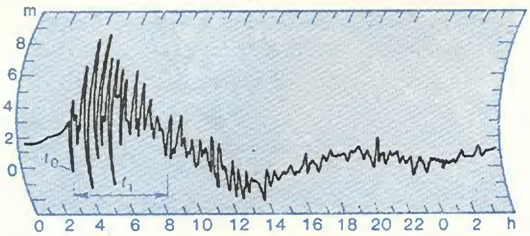


Rys. 6. Schemat fali barycznej

fale tsunami

Innym rodzajem występujących w morzu fal nieokresowych są fale samotne, zwane tsunami, należące do kategorii falowania swobodnego, a wzbudzane najczęściej wybuchem podwodnego wulkanu, trzęsieniami dna morskiego lub innymi zjawiskami sejsmicznymi. Tsunami w płytkowodnych rejonach przybrzeżnych osiągają nieraz wysokość rzędu kilkudziesięciu metrów i są zjawiskiem bardzo groźnym dla życia i mienia ludzkiego. Tak np. słynny wybuch wulkanu na wyspie Krakatau w 1883 r. był przyczyną katastrofalnej powodzi morskiej, wywołanej falą tsunami, która zniszczyła szereg miast na Jawie i Sumatrze oraz pochłonęła ok. 40 000 istnień ludzkich. Powodzie morskie zdarzają się dość często w rejonach le-

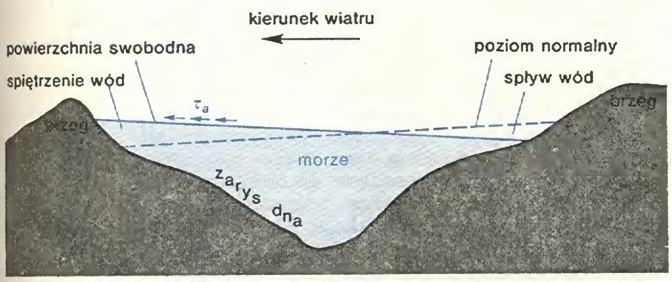
żących na trasach wędrówki fal tsunami. Z tych względów instaluje się tam specjalne systemy ostrzegawcze, rejestrujące zachowanie się swobodnej powierzchni wód w strefie przybrzeżnej. Na rys. 7 zamieszczony jest typowy oscylogram obrazujący przebieg fali tsunami. Podstawowym sygnałem mówiącym o zbliżeniu się tsunami jest chwilowe obniżenie się wód, zwane zwiastunem (moment t_0 na rys. 7). Później następuje faza przejścia tsunami właściwej t_1 i faza $t > t_1$ oscylacji wtórnych.



Rys. 7. Oscylogram tsunami (wg Takashiego)

spiętrzenia i spływy dryfowe

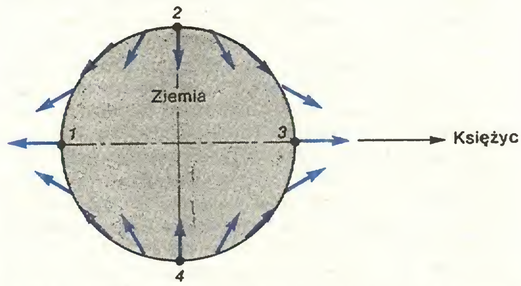
Najczęstszymi przyczynami nieokresowego podwyższania się i opadania swobodnej powierzchni wód w strefie przybrzeżnej morza są tzw. spiętrzenia i spływy dryfowe, zwane czasem wiatrowymi lub sztormowymi (przy silnych wiatrach). Wywołany siłami tarcia aerodynamicznego (wiatrem) ruch mas wodnych jest w strefie przybrzeżnej tłumiony konturem brzegu, stanowiącym dla przepływu wód naturalną zaporę, oraz nierównościami i szorstkością dna, których wpływ jest tym większy, im mniejsza jest głębokość akwenu. Dlatego średnia prędkość ruchu mas wodnych jest tym mniejsza, im bliżej brzegu znajduje się miejsce obserwacji przepływu. Zatem przy brzegach, na które wiatr wieje (nawietrznych) masy wodne spiętrzają się, a przy brzegach przeciwległych (podwietrznych) spływają. W pierwszym wypadku obserwujemy podnoszenie się poziomu wód, w drugim — obniżenie (rys. 8).



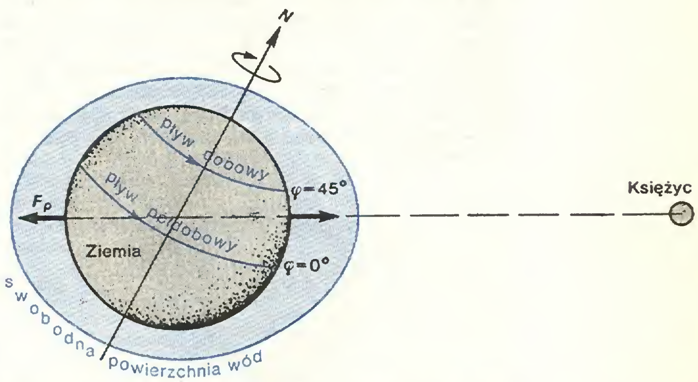
Rys. 8. Schemat spiętrzenia i spływu dryfowego

pływy

Główne okresowe oscylacje swobodnej powierzchni morza wzbudzane są astronomicznymi siłami pływotwórczymi. Oscylacje te, zwane pływami, należą do kategorii zjawisk makroskalowych w skali ziemskiego globu. Towarzyszą im zawsze w przybrzeżnej strefie morza cykliczne spływy wód i obniżanie się swobodnej powierzchni akwenu zwane odpływem oraz spiętrzanie wód — zwane przypływem. W celu zrozumienia mechanizmu tych procesów przyjrzyjmy się jeszcze raz rys. 2. Układ Ziemia–Księżyc pozostaje w równowadze dlatego, że średnia siła odśrodkowa ziemskiego globu $(\vec{F_o})_{śr}$ równa jest średniej sile grawitacyjnej $(\vec{F_k})_{śr}$. Ponieważ siły cząstkowe F_o, działające na dowolne elementy masy ziemskiego globu są sobie równe, więc $(\vec{F_o})_{śr} = F_o$. Natomiast siła $(\vec{F_k})_{śr}$ równa jest sile cząstkowej F_k jedynie w środku masy Ziemi. Z tych względów w każdym punkcie ziemskiego globu istnieć będzie różnica sił $\vec{F_o} - \vec{F_k} = \Delta\vec{F}$, przybierająca największe wartości w punktach *1* i *3*, a najmniejsze w punktach *2* i *4* (rys. 9), przy czym siły pływotwórcze na półkuli zorientowanej w stronę Księżyca (*2* → *3* → *4*) będą nieco większe od sił działających na półkuli przeciwnej (*4* → *1* → *2*). Układ sił uwidoczniony na rys. 9 wywoła odpływ wód z obszarów *2* i *4* i obniżenie swobodnej powierzchni morza w tych rejonach, przypływ zaś wód i podwyższenie powierzchni w obszarach *1* i *3*. Z kolei oś obrotu Ziemi nachylona jest pod pewnym kątem w stosunku do prostej łączącej środki mas Ziemi i Księżyca. Przy założeniu, że Ziemia jest całkowicie pokryta wodą, swobodna powierzchnia wód ukształtowałaby się jak na rys. 10.



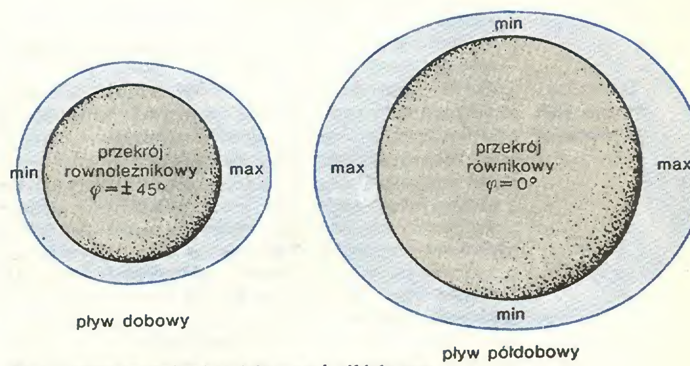
Rys. 9. Schematyczny rozkład sił pływotwórczych na powierzchni Ziemi



Rys. 10. Schemat odkształcenia swobodnej powierzchni wód pod wpływem sił pływotwórczych $\vec{F_p}$

pływy dobowe i półdobowe

Gdybyśmy teraz przecięli glob ziemski równoleżnikowo, to obwiednie przekroju (rys. 11) zawierałyby jeden wysoki i jeden niski stan wód w przekroju równoleżnika $\varphi = 45°$ szerokości geograficznej północnej i południowej oraz dwa wysokie i dwa niskie stany wód w przekroju równikowym $\varphi = 0$. I dlatego w czasie jednego obrotu Ziemi dookoła swojej osi obserwujemy równoleżnikowy ruch grzbietów

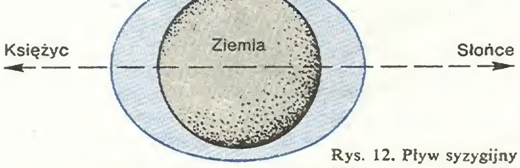


Rys. 11. Schematy pływów: dobowego i półdobowego

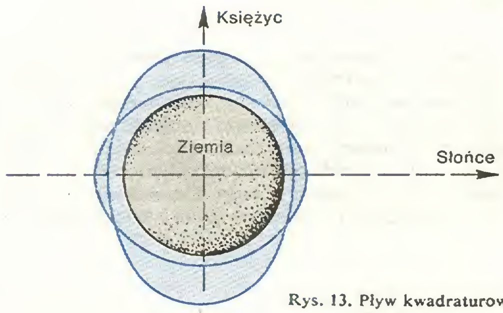
i dolin fali pływowej w prawo względem punktów stałych na Ziemi. W obszarach $\varphi = \pm 45°$ w czasie jednej doby pojawia się tylko jeden wysoki stan wód i jeden niski, będące tzw. pływami dobowymi księżycowymi o okresie 25,82 h. W obszarach równikowych natomiast pojawiają się w tym czasie dwa stany wysokie i dwa niskie, będące pływami księżycowymi półdobowymi o okresie 12,42 h. Te dwa rodzaje pływów przenikają się oczywiście wzajemnie i występują w każdej szerokości geograficznej, z tym że w rejonach $\varphi = 0$ i $\varphi = 45°$ przybierają wartości ekstremalne. Analogiczne pływy (dobowe i półdobowe) występują w układzie wzajemnego oddziaływania Ziemia–Słońce. Obserwujemy zatem na Ziemi pływy dobowe słoneczne o okresie 24,07 h i pływy półdobowe słoneczne o okresie 12 h.

inne rodzaje pływów

Pływy dobowe i półdobowe (słoneczne i księżycowe) stanowią główne składowe pływów astronomicznych, ale nie jedyne. Niezależnie bowiem od omówionych już mechanizmów — z faktu, że Ziemia krąży wokół Słońca i Księżyc wokół Ziemi po orbitach eliptycznych wynikają pływy długookresowe: księżycowy półmiesięczny o okresie 13,661 doby i miesięczny o okresie 27,555 doby, oraz słoneczny półroczny o okresie 182,621 doby i roczny. Ponadto płaszczyzny tych orbit nachylone są w stosunku do płaszczyzny równika, z czego wynikają pływy mieszane dobowe i półdobowe o okresach 23,93 h i 12,97 h. Inne pływy mieszane wynikają z nakładania się i przenikania różnych fal pływowych oraz zjawisk transformacji tych fal w rejonach kontynentalnych szelfów. Wzajemne położenie Księżyca i Słońca względem Ziemi wywiera również wpływ na zjawiska pływowe. Największe amplitudy fali pływowej wzbudzane są wówczas, gdy środki mas Ziemi, Słońca i Księżyca leżą na jednej prostej (pływ syzygijny, rys. 12), najmniejsze zaś — gdy proste łączące środki mas Ziemi, Słońca i Księżyca ustawione są prostopadle do siebie (pływ kwadraturowy, rys. 13).



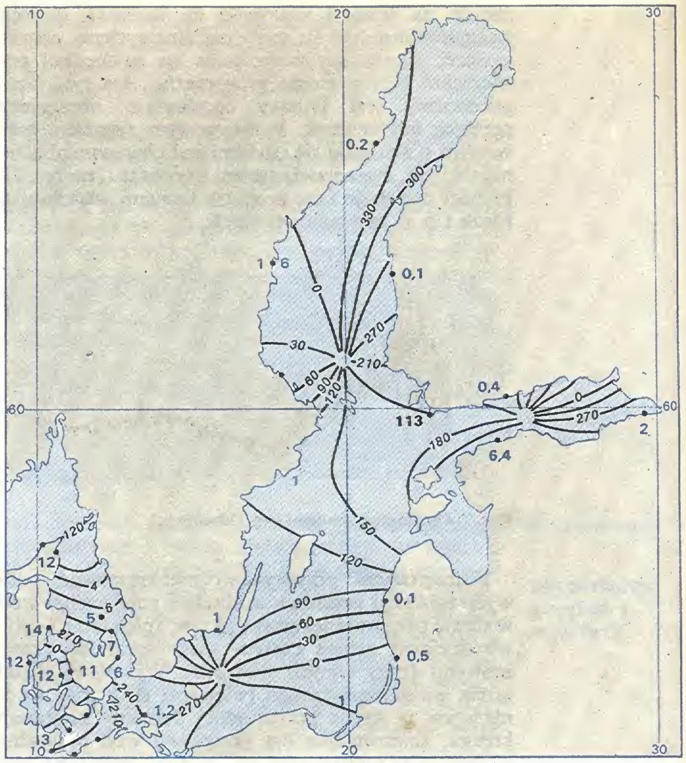
Rys. 12. Pływ syzygijny



Rys. 13. Pływ kwadraturowy

transformacja fali pływowej

Gdyby glob ziemski pokryty był całkowicie wodą, maksymalna wysokość fali pływowej księżycowej (różnica między najwyższym i najniższym stanem wód) nie przekraczałaby 55 cm, a fali pływowej słonecznej — 24 cm. Istnienie lądów i związanych z nimi płycizn szelfowych, zatok i cieśnin, powoduje zmniejszanie się prędkości rozprzestrzeniania się fali pływowej w miarę zbliżania się do lądu (wpływ tarcia przydennego i ograniczenie głębokości akwenu). Ponadto przy brzegach stromych część energii fali pływowej zostaje zamieniona na pracę sił wzbudzających falę „odbitą", rozprzestrzeniającą się w kierunku przeciwnym do ruchu fali nabiegającej. Wskutek tych zjawisk powstają procesy transformacji fal pływowych w obszarach przybrzeżnych, gdzie występuje silny wzrost amplitudy pływów, osiągającej średnio wartości od 1,30 do 3,00 m (pływy dobowe), a w szczególnych okolicznościach (zwężona zatoka) dochodzącej nawet do 18 m.



Rys. 14. Układy amfidromiczne w Morzu Bałtyckim, dla pływu półdobowego, księżycowego (wg Defanta). Położenia linii największego spadku podane są w stopniach, a amplitudy wahań przybrzeżnych w centymetrach

prądy pływowe

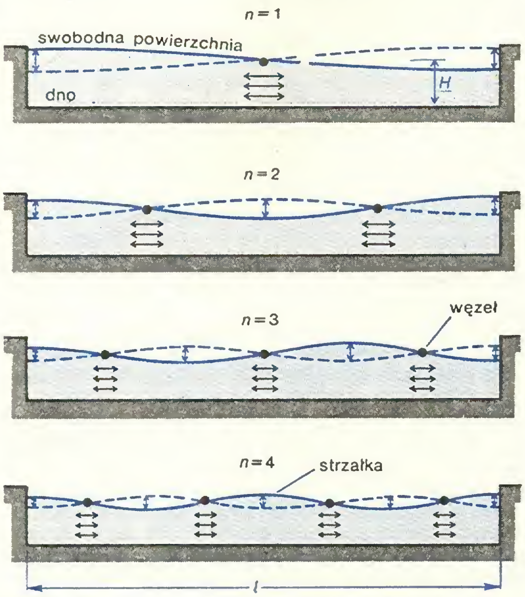
Długości fal pływowych są rzędu połowy długości obwodowej równika. Nic więc dziwnego, że w przybrzeżnych rejonach nawiedzanych przez pływy, fazy przepływu i odpływu obserwowane są w postaci postępowego ruchu wód zwanego prądem pływowym. Prądy te podlegają znacznemu oddziaływaniu sił Coriolisa, które odchylają tory ruchu elementów wody w kierunku prostopadłym do wektora prędkości prądu, w prawo na półkuli północnej i w lewo na półkuli południowej. Zjawisko to wywołuje w akwenach morskich spiętrzenie wód w jednych rejonach przybrzeżnych i równoczesne ich opadanie w rejonach przeciwległych. Linia największego spadku swobodnej powierzchni akwenu, łącząca przeciwległe punkty najwyższego i najniższego stanu wód, zatacza w ciągu całego okresu pływowego krąg, którego środkiem jest punkt amfidromiczny. Stan wód w tym punkcie przedstawia poziom średni. Linie największego spadku swobodnej powierzchni wraz z punktem amfidromicznym tworzą układ amfidromiczny (rys. 14), którego znajomość umożliwia przewidywanie momentów występowania ekstremalnych stanów pływowych w danym rejonie.

układ amfidromiczny

sejsze

Pływy astronomiczne to rodzaj fal grawitacyjnych postępowych. Innym rodzajem okresowych oscylacji swobodnej powierzchni morza są tzw. fale grawitacyjne stojące. Każdy akwen morski, w którym naruszona została równowaga hydrostatyczna mas wodnych, wykonuje okresowe ruchy swobodnej powierzchni wód, zwane sejszami. Stanowią one odpowiednik drgań własnych ciała materialnego, a charakteryzują się takimi szczegółami właściwymi falom stojącym (rys. 15), jak węzły, w których elementy wody znaj-

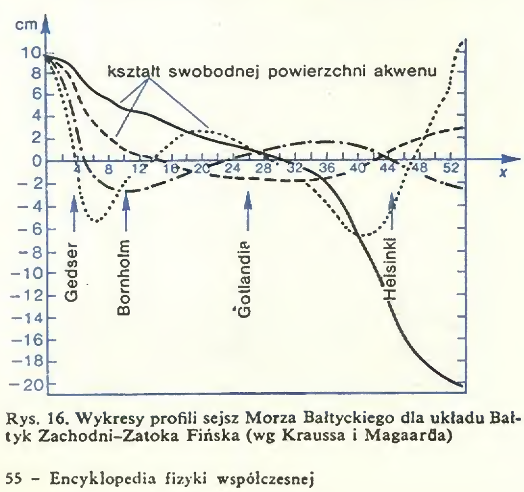
dują się w stanie nieruchomym lub wykonują słabe poziome ruchy oscylacyjne względem punktu centralnego (węzła), oraz tzw. strzałki, w których ruch elementów wody jest ruchem oscylacyjnym pionowym względem poziomu spokoju. Czas, w którym swobodna powierzchnia wód w strzałce powraca do swego pierwotnego położenia, nazywamy okresem drgań własnych. W zależności od rozmiarów i kształtu zbiornika okres ten może przybierać szereg wartości dla $n = 1, 2, 3 \ldots$ itd. punktów węzłowych. Mówimy wówczas o sejszy jedno-, dwu- lub n-węzłowej. Rysunek 15 obrazuje uproszczone schematy sejsz n-węzłowych w kanale prostopadłym o stałej głębokości.



Rys. 15. Schemat sejszy w prostokątnym, wydłużonym kanale

W akwenach szerokich mamy do czynienia z sejszami wzdłużnymi i sejszami poprzecznymi; ich charakterystyki mogą się znacznie różnić. W naturalnych akwenach morskich, których nie można w przybliżeniu traktować jak prostych kształtów kanonicznych (prostokąt, koło, elipsa) i które nie mają stałej głębokości, procesy sejszowe są bardzo złożone i opis ich jest bardzo trudny. Na rys. 16 zamieszczone są dla przykładu profile sejsz wzdłużnych Morza Bałtyckiego w układzie Bałtyk Zachodni–Zatoka Fińska.

W warunkach naturalnych (morze rzeczywiste) omówione oscylacje swobodnej powierzchni akwenu, wzbudzane siłami wiatrowego oddziaływania (spiętrzenia i spływy dryfowe), dynamiką dna (fale tsunami), ciśnieniem atmosferycznym (fale baryczne), astronomicznymi siłami pływotwórczymi (fale pływowe) oraz drganiami własnymi akwenu (sejsze), nakładają się wzajemnie tworząc wypadkowy stan wód w morzu. Charakter przestrzenno-czasowej zmienności tego stanu zależy od rodzaju dominującej oscylacji składowej. I tak np. jeżeli oscylacją dominującą będą pływy, zmiany stanu wód będą miały charakter oscylacji okresowych: jeżeli fale baryczne lub spiętrzenia i spływy dryfowe — zmiany te będą zmianami nieokresowymi, o charakterze losowym (przypadkowym).



Rys. 16. Wykresy profili sejsz Morza Bałtyckiego dla układu Bałtyk Zachodni–Zatoka Fińska (wg Kraussa i Magaarda)

falowanie wiatrowe

Na opisane wyżej drgania długookresowe i nieokresowe swobodnej powierzchni morskiego akwenu nakładają się różnorodne oscylacje krótkookresowe, wśród których falowanie wiatrowe jest zjawiskiem najbardziej powszechnym. Turbulentny ruch mas powietrza w dolnych warstwach atmosfery charakteryzuje się obecnością różnowymiarowych wirów, przemieszczających się nad swobodną powierzchnią morza z różnorodną prędkością. Wskutek przypadkowości przestrzennego układu tych wirów oraz ich zmienności w czasie ciśnienie atmosferyczne działające na swobodną powierzchnię morza nie jest wielkością stałą, lecz ulega zmiennym przypadkowym pulsacjom. Pulsacje te łącznie z siłami tarcia aerodynamicznego wzbudzają ruch swobodnej powierzchni akwenu w postaci przestrzenno-czasowych chaotycznych oscylacji, zwanych falowaniem wiatrowym (il. 198, tabl. 53).

Wzrost amplitudy niektórych składowych losowego pola fal wiatrowych następuje wskutek mechanicznego rezonansu powstającego wówczas, gdy częstość pulsacji ciśnień atmosferycznych, przemieszczających się w określonym kierunku z prędkością równą prędkości wiatru, jest równa częstości drgań swobodnej powierzchni akwenu, rozchodzących się z prędkością fazową równą prędkości wiatru. Gdy warunek ten nie jest spełniony, drgania ulegają bardziej lub mniej intensywnemu tłumieniu. Wskutek istnienia sił tarcia aerodynamicznego możliwy jest również wzrost wysokości i długości fali tych składowych falowania wiatrowego, których prędkość rozprzestrzeniania się jest mniejsza od prędkości wiatru.

Fizyczną strukturę losowego pola fal wiatrowych można odwzorować w postaci czasowo-przestrzennej superpozycji nieskończonej liczby regularnych fal elementarnych (składowych harmonicznych) o różnych amplitudach i częstościach (stałych w obrębie poszczególnych ciągów), rozprzestrzeniających się pod różnymi kątami względem kierunku działania wiatru. Fale te nakładają się wzajemnie z losowym przesunięciem fazowym, tworząc chaotycznie sfalowaną powierzchnię morza.

falowanie wewnętrzne

Ruch falowy swobodnej powierzchni morza jest w stosunku do akwenu ruchem powierzchni górnej. Natomiast w stosunku do atmosfery ziemskiej jest to ruch falowy powierzchni dolnej, atmosferycznego podłoża. W stosunku do obu warstw swobodna powierzchnia akwenu jest powierzchnią rozdziału dwóch ośrodków płynnych o różnej gęstości. Wewnątrz morskiego akwenu (szczególnie w akwenie oceanicznym) takie powierzchnie rozdziału występują wszędzie, gdzie istnieje pionowe zróżnicowanie (ciągłe lub skokowe) gęstościowej struktury wód. Ruch oscylacyjny takiej powierzchni to tzw. falowanie wewnętrzne.

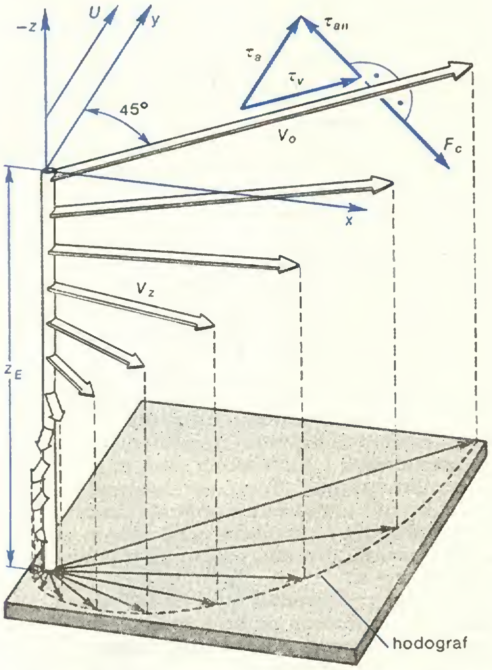
Przyczynami wzbudzającymi fale wewnętrzne mogą być ruchy swobodnej powierzchni akwenu, mechanizmy wywołujące przejście układu barotropowego w układ baroklinowy, siły pływotwórcze, drgania własne zbiornika, zjawiska sejsmiczne itp. Charakterystyki fal wewnętrznych są zatem bardzo różnorodne i zależą od rodzaju czynnika wzbudzającego. Ogólnie można powiedzieć, że fale wewnętrzne są formą ruchu analogiczną do fal powierzchniowych z tym uzupełnieniem, że intensywność tego ruchu w poważnej mierze zależy od gęstości górnej warstwy wód.

Zjawiskiem mającym duże znaczenie praktyczne dla podwodnej żeglugi jest zjawisko rezonansu fali wewnętrznej z czynnikiem wymuszającym. W warunkach takich wysokość fal wewnętrznych może przybierać bardzo duże wartości (kilkudziesięciu, a nawet kilkuset metrów) i może być powodem trudności w podwodnej nawigacji, a nawet awarii podwodnego pojazdu.

B) Prądy morskie

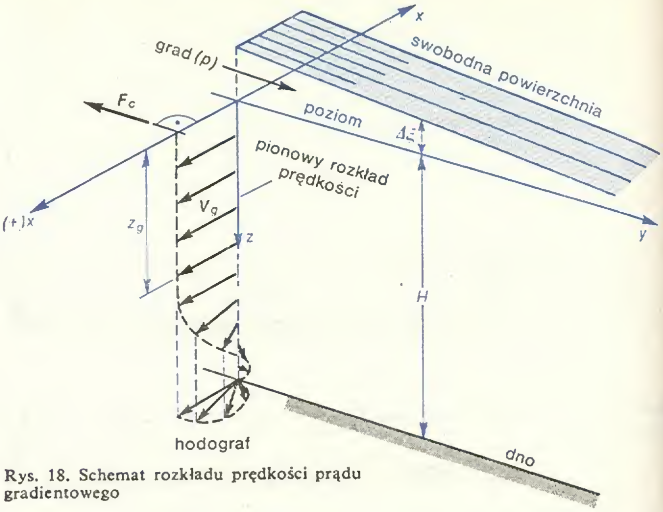
Drugim, po falowaniu, powszechnym zjawiskiem hydrodynamicznym w morzu jest postępowy ruch mas wodnych, zwany prądem morskim. Wśród różnorodnych rodzajów tego zjawiska na pierwszym miejscu wymienić należy prądy dryfowe. Żeby zrozumieć, jak powstają takie prądy, załóżmy, że osie współrzędnych prostokątnych x, y (rys. 17) leżą w płaszczyźnie swobodnej powierzchni akwenu i są tak ustawione, że wiatr wieje równolegle do osi y. Wektor prędkości ruchu wód w warstwie powierzchniowej ($\vec{V}_o$) wywołanego siłami aerodynamicznymi tarcia ($\vec{\tau}_a$) jest w początkowym okresie skierowany zgodnie z kierunkiem wiatru (kierunek sił $\vec{\tau}_a$). W miarę wzrostu prędkości przepływu siła Coriolisa ($\vec{F}_c$) bę-

prądy dryfowe



Rys. 17. Schemat rozkładu prędkości prądu dryfowego (spirala Ekmana)

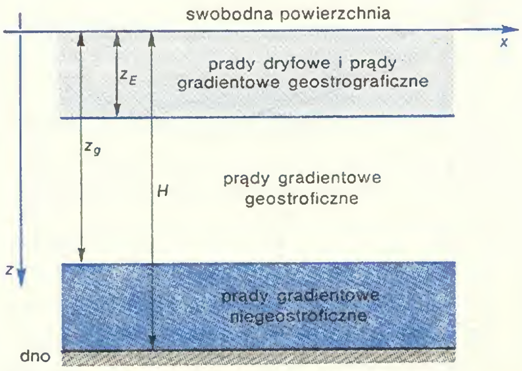
dzie coraz bardziej odchylać wektor $\vec{V}_o$ (w prawo na półkuli północnej, w lewo na półkuli południowej, patrząc zgodnie z kierunkiem ruchu) do położenia, w którym składowa siły tarcia aerodynamicznego ($\vec{\tau}_{an}$) prostopadła do wektora $\vec{V}_o$ zrównoważy siłę $\vec{F}_c$ (rys. 18). Stan równowagi wystąpi wówczas, gdy wektor $\vec{V}_o$ będzie odchylony o 45° od początkowego położenia. Ze wzrostem głębokości wpływ siły aerodynamicznego tarcia będzie stopniowo malał i jednocześnie będzie malał moduł prędkości ruchu mas wodnych, a kierunek prądu odchylany będzie przez siłę Coriolisa o kąt tym większy, im głębiej będzie się znajdować miejsce obserwacji, tj. im mniejszy będzie wpływ tarcia aerodynamicznego. Na pewnej głębokości z_E, zależnej od prędkości wiatru, kierunek prądu (wektora $\vec{V}_z$) będzie przeciwstawny kierunkowi prądu powierzchniowego (kierunkowi wektora $\vec{V}_o$), a moduł prędkości $|V_z|$ będzie tak mały, że z praktycznego punktu widzenia prąd dryfowy na głębokości z_E, zwanej grubością warstwy Ekmana, nie będzie istniał.



Rys. 18. Schemat rozkładu prędkości prądu gradientowego

prądy gradientowe

Drugim rodzajem postępowego ruchu mas wodnych w morzu są prądy gradientowe, wzbudzane różnicą ciśnień hydrostatycznych panujących w sąsiadujących ze sobą obszarach akwenu. Przypuśćmy, że wiatr, który wywołał spiętrzenie dryfowe, ucichł i spiętrzone masy wód znajdujących się pod przeważającym wpływem sił ciężkości. Nachylenie swobodnej powierzchni akwenu będzie w przybliżeniu takie, jak na rys. 18. Różnica ciśnień hydrostatycznych panujących w dwóch punktach znajdujących się na prostej poziomej, równoległej do osi y, zawsze będzie miała tę samą wartość, niezależnie od głębokości z. Wywołana tą różnicą ciśnień pozioma siła gradientu ciśnienia będzie powodować ruch mas wodnych. Siła ta będzie miała taką samą wartość w całym pionowym przekroju akwenu od dna do swobodnej powierzchni wód, ale prędkość mas wodnych na różnej głębokości będzie różna. Przy samym dnie ruch będzie tłumiony siłami tarcia przydennego i jego kierunek będzie zgodny z kierunkiem gradientu ciśnienia (kierunek osi y). W miarę oddalania się od dna tłumiący wpływ sił tarcia maleje, prędkość przepływu rośnie, a wraz z nią rośnie siła Cariolisa odchylająca wektor prędkości przepływu. Na pewnej głębokości z_g, zależnej od wartości gradientu ciśnienia, wektor siły Coriolisa skierowany będzie przeciwnie do wektora siły wywołanej różnicą ciśnień. Prędkość przepływu mas wodnych zorientowana będzie w tym wypadku równolegle do osi x. Powyżej głębokości z_g prędkość prądu będzie stała, prostopadła do kierunku działania



Rys. 19. Schemat rozkładu prądów w głębokim oceanie

gradientu ciśnienia. Zatem w warstwie akwenu od swobodnej powierzchni do głębokości z_g prądy gradientowe będą miały charakter przepływów geostroficznych.

prądy gęstościowe

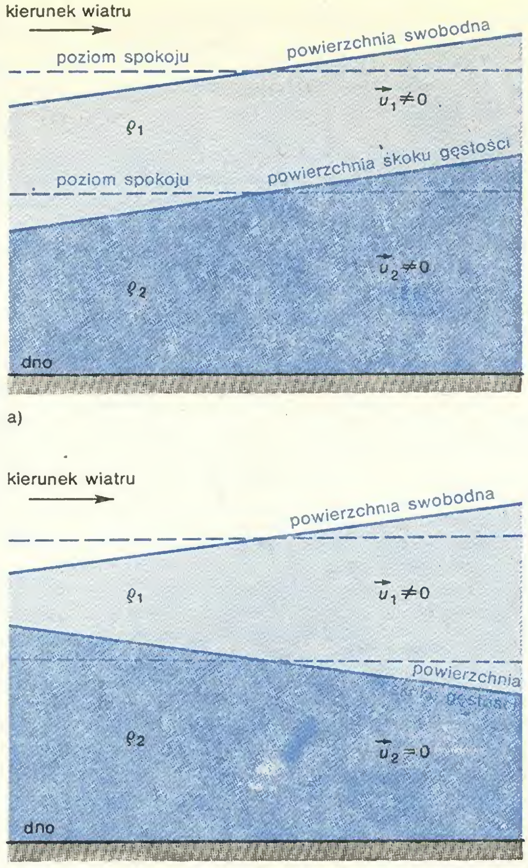
Jeżeli rozkład gęstości wód morskich jest rozkładem niejednorodnym w przestrzeni, to wskutek istnienia różnicy gęstości powstają poziome siły wywołane różnicą ciśnienia, analogiczne do wyżej opisanych, wzbudzające ruch mas wodnych w morzu, zwany prądem gęstościowym. Fizyczny opis tych prądów jest bardzo złożony, ponieważ wraz z przepływem wód zmieniają się gradienty gęstości (woda napływająca ma często inną temperaturę i zasolenie).

W warunkach naturalnych prądy dryfowe i prądy gradientowe przenikają się i nakładają wzajemnie, powodując wzrost lub osłabienie wypadkowego przepływu wód.

W górnych warstwach akwenu z reguły mamy do czynienia z prądami mieszanymi: dryfowymi i gradientowymi geostroficznymi (rys. 19). Poniżej warstwy Ekmana przepływ ma charakter prądów gradientowych geostroficznych, a jeszcze niżej — gradientowych niegeostroficznych.

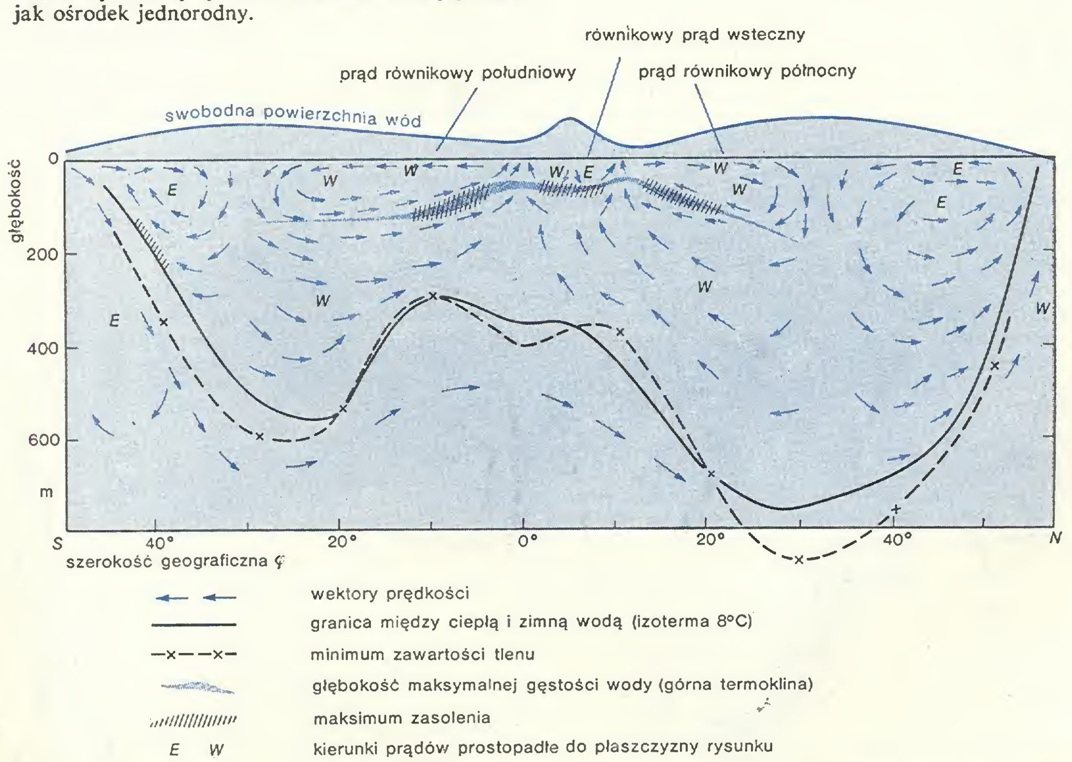
układy barotropowe i baroklinowe

Jeżeli w akwenie morskim występuje wewnętrzna powierzchnia skoku gęstości (rys. 20), co się często zdarza, to przez pewien stosunkowo krótki czas przepływy dryfowe są wzbudzane w obu warstwach. Układ taki nazywamy barotropowym. Po pewnym czasie różnice ciśnień wywołane spiętrzeniem dryfowym w górnej warstwie zostają wyrównane w dolnej warstwie i dryfowy przepływ mas wodnych odbywa się wyłącznie w warstwie górnej. Powierzchnia skoku gęstości zmienia przy tym swoje położenie na przeciwstawne. Układ taki nazywamy baroklinowym. Przy długotrwałym działaniu wiatru mamy do czynienia prawie wyłącznie z układami baroklinowymi. Zmiany prędkości pola wiatru w okresach krótszych od godziny tworzyć będą układy barotropowe. W okresach zmienności od godziny do doby mamy do czynienia z układami mieszanymi. Okresom zbliżonym do okresu tzw. doby wahadłowej towarzyszą zawsze układy baroklinowe, a w okresach większych od 1 doby a mniejszych od 7 dób akwen reaguje znów jak ośrodek jednorodny.

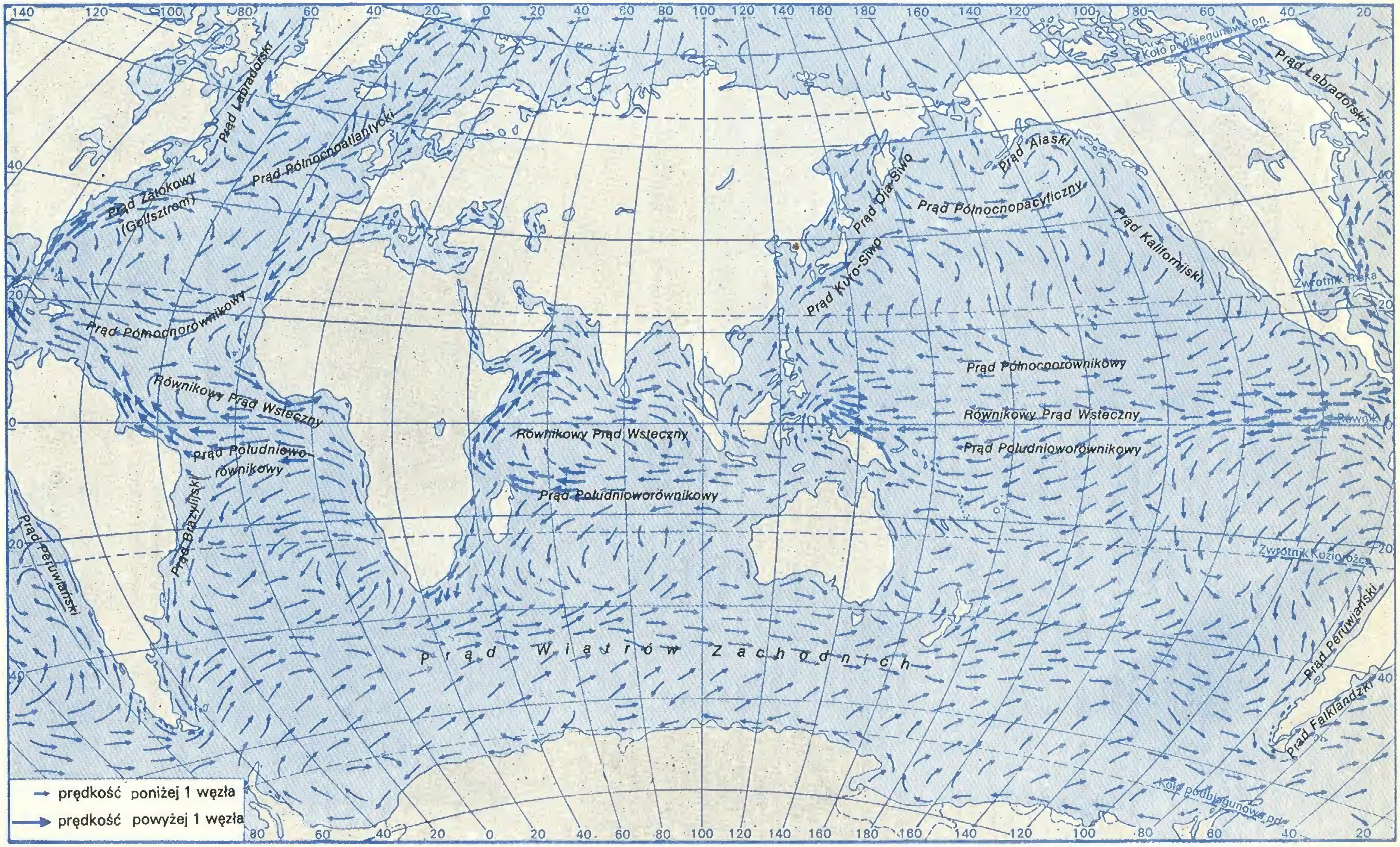


Rys. 20. Schemat układów: a) barotropowego i b) baroklinowego w uwarstwionym oceanie

układy cyrkulacyjne



Rys. 21. Schemat głównych układów cyrkulacyjnych wód w południkowym przekroju środkowej części Oceanu Atlantyckiego (wg Defanta)



Rys. 22. Rozkład prądów powierzchniowych w oceanie światowym w miesiącach luty–marzec (wg Sverdrupa). Strzałki oznaczają wektory prędkości

wielkoskalowa cyrkulacja

Postępowy ruch mas wodnych w wielkich zbiornikach oceanicznych jest niewspółmiernie bardziej złożony niż prądy w akwenach jednorodnych. Istnienie dwóch biegunów zimna i związany z tym niejednorodny rozkład gęstości wód jest czynnikiem wzbudzającym cyrkulację mas wodnych i zjawiska termohalinowej dyfuzji na wielką skalę, w pionowej płaszczyźnie południkowej (rys. 21). Z kolei zmienność siły Coriolisa ze zmianą szerokości geograficznej oraz oddziaływanie wielkoskalowych poziomych układów cyrkulacyjnych atmosfery w warstwie przywodnej wywołują obszerne zamknięte układy cyrkulacyjne oraz intensyfikują przepływy mas wodnych przy wschodnich brzegach kontynentów półkuli północnej. Na rys. 22 uwidoczniony jest przykładowy rozkład prądów powierzchniowych w oceanie światowym. Widoczna jest na nim wyraźna intensyfikacja przepływu w pobliżu wschodnich brzegów kontynentalnych półkuli północnej (Golfsztrom, Kuro-siwo, 1-2 m/s), strefy silnych prądów podzwrotnikowych oraz dwa równoległe układy cyrkulacyjne: jeden na półkuli północnej, gdzie ruch mas wodnych skierowany jest zgodnie z ruchem wskazówek zegara, i drugi na półkuli południowej o orientacji przeciwnej.

Metody badania morskich procesów dynamicznych

Badanie praw rządzących ruchem mas wodnych oraz różnorodnymi jego osobliwościami w morzach i oceanach realizowane jest współcześnie metodami empirycznego i matematycznego modelowania przy wykorzystaniu znajomości równań hydrodynamiki (modele hydrodynamiczne), oraz zasad teorii podobieństwa, statystyki matematycznej i teorii funkcji losowych (modele stochastyczne).

modelowanie hydrodynamiczne

Badania na modelach hydrodynamicznych (teoretyczne i empiryczne) mogą być prowadzone w dwóch układach analitycznych: w układzie analizy wędrownej (współrzędne Lagrange'a) opisującej w czasie i przestrzeni ruch konkretnego, ciągle tego samego elementu wody oraz w układzie analizy lokalnej (współrzędne Eulera) opisującej zmiany w czasie charakterystyk ruchu różnych elementów w wybranym stałym punkcie.

Modelowanie empiryczne polega na hydraulicznym odwzorowaniu w zmniejszonej skali, całego zbiornika morskiego, lub jego wycinka, oraz symulowaniu w takim modelowym akwenie (il. 199, tabl. 53) czynników wzbudzających ruch masy wodnej, zgodnie z zasadami teorii podobieństwa. Badania prowadzone tą metodą dotyczą przeważnie rozpoznawania czasowo-przestrzennych zmian takich charakterystyk jak fluktuacja położenia swobodnej powierzchni wód, stany i zmienność ciśnień hydrodynamicznych, prędkości, przyspieszeń, drogi ruchu elementów wody, linii prądu oraz różnorodnych charakterystyk pochodnych. Zarejestrowane za pośrednictwem specjalnej aparatury pomiarowej wartości bezwzględne tych charakterystyk poddaje się różnorodnej analizie, umożliwiającej rozpoznanie praw rządzących przebiegiem badanego procesu hydrodynamicznego.

Modelowanie matematyczne polega na numerycznym lub analitycznym odwzorowaniu procesu, w skali 1 : 1, w oparciu o hydrodynamiczne równania ruchu. W modelowaniu tym przyjmuje się pewne warunki początkowe i warunki brzegowe, określające a priori charakterystykę procesu w momencie rozpoczęcia badań (obserwacji procesu) oraz jego stanu na granicznych powierzchniach morskiego akwenu (powierzchnia swobodna wód, dno i brzegi morskiego zbiornika). Ponadto przyjmuje się szereg założeń upraszczających, jak np. że woda jest nieściśliwa i nielepka, a wektorowe pole prędkości ruchu elementów wody jest polem potencjalnym.

Rejestrując bezpośrednio w morzu (w stanie naturalnym) czasowe i przestrzenne zmiany różnorodnych parametrów ruchu masy wodnej otrzymujemy dane opisujące własności empirycznej funkcji odwzorowującej określoną charakterystykę ruchu (kinematyczną lub dynamiczną) zmienną w czasie i przestrzeni.

modelowanie stochastyczne

Te funkcje empiryczne mają z reguły charakter funkcji losowych, których własności mogą być opisane metodami opartymi na teorii prawdopodobieństwa i statystyce matematycznej. Charakterystyki morskich procesów hydrodynamicznych mają w znacznej większości rozkład normalny funkcji gęstości prawdopodobieństwa oraz taką własność, że operacje uśredniania w zespole realizacji funkcji losowej mogą być zastąpione operacją uśredniania w jednej, dostatecznej długości realizacji czasowej lub przestrzennej. Modelowanie stochastyczne tego rodzaju procesów może być realizowane za pośrednictwem czterech charakterystyk: wartości przeciętnej, charakteryzującej najbardziej prawdopodobną wartość danego parametru (np. średnią prędkość prądu, średnią wysokość falowania), wariancji charakteryzującej intensywność fluktuacji wartości chwilowych względem średniej (np. pulsacje prędkości ruchu elementów wody, chwilowe amplitudy falowania wiatrowego), funkcji korelacji charakteryzującej związek między chwilowymi wartościami danej charakterystyki lub wartościami dwóch dowolnych, różnych charakterystyk oraz funkcji widmowej gęstości mocy pulsacji procesu, charakteryzującej energię falowania lub moc fluktuacji dowolnej charakterystyki. Modelowanie stochastyczne może być realizowane również w oparciu o teoretyczną postać przyjętej a priori funkcji losowej.

modelowanie deterministyczno-stochastyczne

Często stosowaną metodą badania morskich procesów hydrodynamicznych jest modelowanie deterministyczno-stochastyczne. Przeważnie modele takie uzyskuje się przez wprowadzenie do hydrodynamicznych równań ruchu warunków brzegowych w postaci funkcji losowej lub jej statystycznych, lub stochastycznych charakterystyk.

Matematyczne modelowanie (hydrodynamiczne i stochastyczne) oraz system badań empirycznych prowadzonych na statkach, bojach oceanograficznych i różnorodnych stacjach autonomicznych (il. 197, tabl. 52), stanowią dziś główne źródło informacji o prawach rządzących złożonym, różnowymiarowym i wielorodzajowym ruchem mas wodnych w morzach i oceanach.

Praktyczne problemy dynamiki morza

Ruch mas wodnych w morzach i oceanach jest zjawiskiem ważnym dla technicznej i gospodarczej działalności człowieka. Niemal każde przedsięwzięcie techniczne w przybrzeżnej strefie morza uzależnione jest od intensywności i zasięgu tych procesów hydrodynamicznych. Budowle osłonowe nowoczesnych portów morskich, autonomiczne stanowiska przeładunkowe i morskie szyby wiertnicze w akwenie odsłoniętym, elektrownie pływowe, urządzenia dla zrzutu i poboru wód morskich i tym podobne inwestycje techniczne uwarunkowane są zawsze zakresem zmienności stanu wód i dynamicznym oddziaływaniem falowania morskiego. Przenikające do wnętrza portu fale wiatrowe są czynnikami generującymi wzmożone drgania swobodnej powierzchni portowego akwenu, które utrudniają lub wręcz uniemożliwiają przeładunkową działalność w porcie.

działanie niszczące

Działanie fal wiatrowych jest często przyczyną naruszania stateczności budowli hydrotechnicznej, stałego niszczenia naturalnych brzegów morskich i konstrukcji ochronnych (il. 200, tabl. 53) oraz zapiaszczenia torów i kanałów żeglugowych. Groźne powodzie morskie, wywołane falą tsunami i silnym spiętrzeniem sztormowym, są dla nawiedzanych przez nie krajów (np. Japonia lub Hawaje) istną plagą, pochłaniają znaczne środki materialne, a przede wszystkim narażają życie ludzkie.

Możliwości wykorzystania mechanicznej energii

wykorzystanie energii pływów i prądów morza do praktycznych potrzeb człowieka związane są dziś głównie z pływami oraz prądami morskimi. Wywołane nimi kilkunastometrowej wysokości spiętrzenia wód w przybrzeżnej strefie morza mogą służyć do produkcji energii elektrycznej (elektrownie pływowe). Astronomiczne pływy w fazie przepływu umożliwiają statkom morskim żeglugę w głąb lądu (np. do Londynu). Prądy morskie i termohalinowa cyrkulacja wód to w głównej mierze mechanizmy przenoszenia zawartych w wodzie substancji biernych (sole, zawiesiny, zanieczyszczenia). Każdemu rejonowi odpływu wód powierzchniowych towarzyszy skierowany ku górze napływ wód głębinowych, przenoszących substancje biogenne. W konsekwencji rejony te charakteryzuje wzmożona produkcja pożywienia dla ryb i wysoki stan zarybienia.

wpływ na pogodę i klimat Ponadto konieczność intensyfikacji lądowych źródeł surowców żywnościowych stawia dziś przed nauką i praktyką zadanie zwiększenia dokładności prognoz meteorologicznych i klimatycznych z dowolnym dystansem wyprzedzenia w czasie. Dziś wiemy już, że ziemskie lądy wywierają na zjawiska meteorologiczne i klimatyczne wpływ niewielki w porównaniu z wpływem mórz i oceanów. Z tych względów dokładne poznanie różnowymiarowych procesów wzajemnego oddziaływania morza i atmosfery oraz związanych z nimi zjawisk hydrodynamicznych, nabiera z roku na rok coraz większego znaczenia.

Turbulentne zjawiska dyfuzji i mieszania wód odgrywają dziś w dobie intensywnego zanieczyszczenia wód morskich, które towarzyszy współczesnej cywilizacji i pociąga za sobą ekologiczne zmiany środowiska, rolę naturalnego mechanizmu oczyszczania wód. Dzięki tym procesom do dziś jeszcze odprowadza się do morza różnego rodzaju ścieki i odpady bez katastrofalnie negatywnych dla siebie skutków

Optyka morza

Jerzy Dera

Optyka morza zajmuje się badaniem rozchodzenia się i oddziaływania światła w morzu, a szczególnie światła słonecznego, niosącego energię, która podtrzymuje życie wszystkich organizmów morskich oraz procesy termodynamiczne decydujące o klimacie ziemskim. Około 98% energii wchodzącego do morza promieniowania słonecznego (głównie o długościach fal $\lambda < 4\ \mu m$) pochłania woda morska, zamieniając je na ciepło. Energia ta, dostarczana nierównomiernie masom wodnym w przestrzeni oceanu, podtrzymuje potężne, międzykontynentalne cyrkulacje tych mas i powoduje wyparowywanie z powierzchni oceanu do atmosfery średnio ok. metrowej warstwy wody rocznie oraz wypromieniowywanie olbrzymich ilości ciepła. Pociąga to za sobą również cyrkulacje mas powietrza wraz z parą wodną, odbywające się czasem w sposób bardzo burzliwy, w postaci huraganów. Masy wodne i powietrze, niosące w ten sposób ciepło z mocno nasłonecznionych stref gorących ku biegunom, wyrównują i łagodzą klimat ziemski oraz podtrzymują odżywczy obieg wody w przyrodzie.

Nie mniej ważną rolę w kształtowaniu środowiska życia na Ziemi odgrywa pozostała drobna część energii światła słonecznego w morzu. Zaledwie ułamek procenta wchodzącej do morza energii zostaje zużywany na fotosyntezę materii organicznej w komórkach fitoplanktonu.

fotosynteza w morzu Ta produkowana z wody, dwutlenku węgla i zawartych w wodzie pierwiastków chemicznych materia organiczna (węglowodany, tłuszcze, białka) jest pierwszym ogniwem łańcucha pokarmowego wszystkich organizmów morskich (stąd fotosyntezę i towarzyszące jej procesy wzrostu komórek zwiemy biologiczną produkcją pierwotną). Produktem reakcji fotosyntezy jest także wolny tlen, który służy do oddychania roślinom i zwierzętom morskim oraz do utleniania nie zużywanych i zanieczyszczających wodę substancji organicznych. Znaczna część produkowanego w procesie fotosyntezy wolnego tlenu ulatnia się z morza do atmosfery, dzięki czemu 60–90% (wg różnych oszacowań) wolnego tlenu w atmosferze ziemskiej pochodzi z morza. Oprócz fotosyntezy materii organicznej, pod wpływem światła w morzu zachodzi wiele innych ważnych, mało jeszcze zbadanych reakcji fotochemicznych, prowadzących np. do rozpadu niektórych szkodliwych związków organicznych, do wydzielania jodu czy zabijania niektórych gatunków bakterii.

kolor czystego morza Najczystsze morza świata mają w świetle słonecznym wyraźny ciemnobłękitny kolor; jest to skutek silnego rozpraszania fioletu na cząsteczkach wody. Ze wzrostem głębokości w morzu energia światła dziennego znacznie maleje wskutek jego rozpraszania i absorpcji przez wodę, oraz przez zawarte w niej zawiesiny i rozpuszczone substancje organiczne, których ilość zależy od stopnia zanieczyszczenia danego akwenu. Zawarta w wodzie sól morska osłabia światło widzialne w minimalnym stopniu.

światło w głębi morza Pomimo to nawet w najczystszych i najbardziej nasłonecznionych morzach (np. w Morzu Sargassowym lub we wschodniej części Morza Śródziemnego) na głębokości 1000 m panuje ciemność. Nie jest to jednak ciemność absolutna, ponieważ wiele organizmów morskich posiada zdolność świecenia, zwanego bioluminescencją.

Wraz z głębokością zmienia się w morzu skład widmowy światła, stopień jego rozproszenia, stopień polaryzacji, fluktuacje natężenia, wywołane falowaniem powierzchni morza, i inne właściwości, do których przystosowują się rośliny i zwierzęta żyjące na różnych głębokościach. I tak np. barwniki w komórkach fitoplanktonu pochłaniające światło potrzebne do fotosyntezy są tak dobierane przez naturę, że maksymalnie wykorzystują całe widmo światła dostępne na danej głębokości.

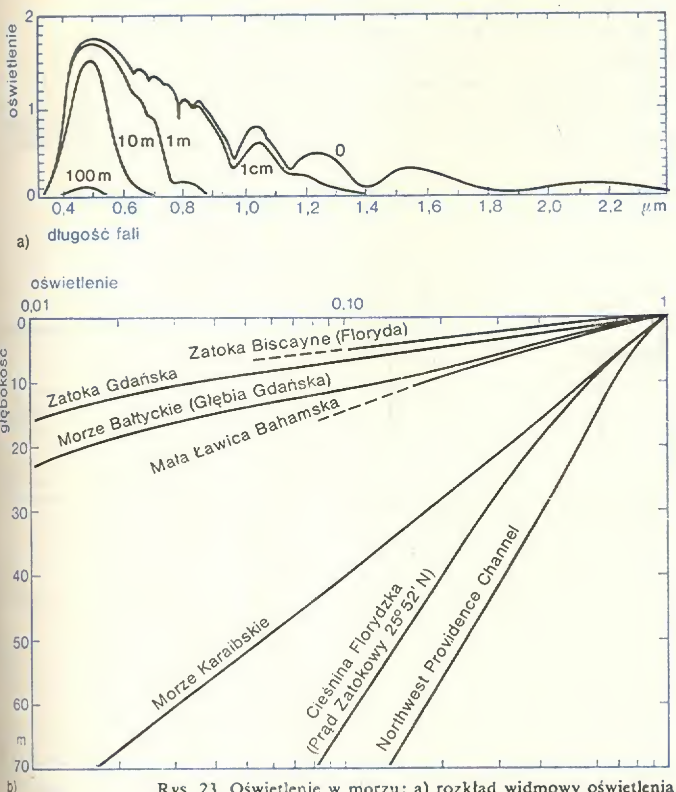
fotoadaptacja organizmów morskich Nie jest też przypadkowe zabarwienie ryb; służy im ono przede wszystkim do maskowania się w środowisku wodnym (jasne podbrzusze ryby jest zwykle słabo widoczne z dołu, na tle jasnej powierzchni morza, ciemny zaś grzbiet ryby jest źle widoczny z góry, na tle ciemnej toni wodnej; czerwony kolor ryb głębinowych jest zupełnie niewidoczny w pozbawionym czerwieni świetle naturalnym na dużych głębokościach, a to wskutek silnego pochłaniania czerwieni przez molekuły wody). Skomplikowane są także organy wzroku niektórych gatunków zwierząt morskich; dzięki temu są one zdolne rozróżniać wiele cech światła, jak barwa, kątowy rozkład natężeń, czy kierunek polaryzacji. Różne cechy podwodnego światła są wykorzystywane przez ryby i inne zwierzęta morskie do ich orientacji podczas wędrówek w przestrzeni wodnej, pozbawionej często innych znaków szczególnych; wiele organizmów wędruje wraz z dobowymi zmianami oświetlenia, przemieszczając się na dużą głębokość w dzień, a ku powierzchni — w nocy.

widoczność podwodna Zasięg widoczności przedmiotów w wodzie, zarówno gołym okiem, jak i za pomocą przyrządów optycznych lub kamer telewizyjnych, nie przekracza 100 m nawet w najczystszych morzach, a na ogół jest dużo mniejszy (w Bałtyku np. od kilku do kilkunastu metrów).

Badania rozchodzenia się światła w morzu

Ten rodzaj badań rozpoczęto pod koniec XIX w.; używano przy tym klisz fotograficznych naświetlanych naturalnym światłem panującym na różnych głębokościach. Już wówczas stwierdzono znaczne różnice w poszczególnych akwenach morskich. Stopniowe udoskonalanie fotograficznych metod badań doprowadziło w 1922 r. znanego oceanologa duńskiego, M. Knudsena, do zastosowania spektrografu. Za pomocą spektrografu stwierdzono, że nie tylko natężenie, ale i widmo światła dziennego zmienia się

wraz z głębokością w morzu i jest bardzo różne w różnych morzach (rys. 23). W tym samym czasie zastosowano pierwsze detektory fotoelektryczne. Spowodowało to przełom w badaniach, gdyż pozwoliło na zdalne dokonywanie pomiarów światła w głębi morza za pomocą przyrządu opuszczanego z burty statku, a połączonego kablem z miernikiem prądu fotoelektrycznego znajdującym się na pokładzie statku.



Rys. 23. Oświetlenie w morzu: a) rozkład widmowy oświetlenia dziennego (odgórnego) na powierzchni morza i na różnych jego głębokościach (w jednostkach względnych) wg N. G. Jerlova (1968 r.), b) przykłady osłabiania oświetlenia odgórnego z głębokością w różnych morzach dla fal długości ok. 530 nm (światło zielone)

Kolejny znaczny skok w postępie badań nastąpił dopiero w latach sześćdziesiątych naszego stulecia, kiedy to powszechnie wprowadzono do pomiarów światła w morzu czułe fotopowielacze i filtry interferencyjne, pozwalające precyzyjnie wydzielić z widma światła dowolne, wąskie pasmo badanych fal świetlnych. Postęp ten wiąże się oczywiście z ogólnym rozwojem fizyki i techniki badań naukowych oraz z nowymi, doskonalszymi metodami matematycznej analizy procesów fizycznych za pomocą komputerów.

Współczesne urządzenia do badań optycznych w morzu działają prawie wyłącznie na zasadzie wykorzystania zjawiska fotoelektrycznego w takich przetwornikach jak fotopowielacze, fotokomórki próżniowe, ogniwa selenowe i fotooporniki półprzewodnikowe. Przetwornik umieszcza się w wodoszczelnej zanurzanej części przyrządu, zwanej sondą optyczną. Sonda jest ponadto wyposażona w precyzyjny układ optyczny, w skład którego wchodzi wodoszczelne okienko (wejście dla światła), układ filtrów optycznych lub monochromator oraz odpowiedni układ soczewek i przesłon (zależnie od potrzeb i przeznaczenia przyrządu). Na ogół niezbędnym elementem sondy optycznej jest też precyzyjny wzmacniacz liniowy, który wzmacnia prąd fotoelektryczny z przetwornika przekazywany za pośrednictwem kabla do odbiornika na pokładzie statku. Niektóre przyrządy optyczne, służące np. do pomiarów rozpraszania światła lub osłabiania wiązki światła, wymagają także

morskie sondy optyczne

wyposażenia sondy we własne źródło światła. Jest to najczęściej żarówka punktowa z kolimatorem optycznym lub laser umieszczony na konstrukcji nośnej sondy w oddzielnym, wodoszczelnym pojemniku. Bardziej złożone urządzenia mają ponadto zdalny elektromechaniczny zmieniacz filtrów lub układ sterowania monochromatora (wybierania żądanej długości fali), urządzenie do zmiany zakresu czułości przyrządu, urządzenie do zamiany prądu fotoelektrycznego w impulsy elektryczne, które mogą być przesyłane kablem o mniejszej liczbie przewodów, a nawet poprzez wodę w postaci impulsów dźwiękowych. Gdy w sondzie znajduje się fotopowielacz, musi tam również być zasilacz wysokiego napięcia, regulowany napięciem niskim (z akumulatora) poprzez kabel z pokładu statku. Przy wnikliwych pomiarach rozpraszania światła konieczny jest także elektromechaniczny układ do zmiany i zdalnego odczytu kąta położenia odbiornika w stosunku do rozpraszanej wiązki światła. W badaniach dopływu strumieni światła z różnych kierunków w przestrzeni wodnej sonda musi być ponadto wyposażona w kierunkowy odbiornik światła oraz skomplikowany i zdalnie regulowany układ śrub napędowych i sterów, aby wisząc na linie na dowolnej głębokości w morzu, ustawiała się w żądanym i kontrolowanym kierunku.

Zależnie od budowy i przeznaczenia rozróżnia się sondy zwane przezroczomierzami, miernikami podwodnego oświetlenia, miernikami rozpraszania światła, miernikami radiacji itp. (rys. 24 i il. 201, 202 z tabl. 53).

pokładowa aparatura rejestrująca

Drugą część aparatury pomiarowej stanowią zwykle urządzenia zasilające, sterujące i rejestrujące, instalowane na pokładzie statku i połączone z sondą wodoszczelnym kablem elektrycznym. Ze względu na statystyczne fluktuacje mierzonych wielkości i wynikającą stąd potrzebę wykonywania dużych serii pomiarów, sygnały z sond optycznych coraz rzadziej odczytuje się mikroamperomierzem. Używa się raczej miliwoltomierzy z automatycznym zapisem analogowym na taśmie papierowej, a w najnowszych urządzeniach — perforatorów lub rejestratorów magnetycznych rejestrujących sygnały w postaci dostosowanej do komputera. Najnowocześniejsze statki badawcze, np. polski statek badawczy „Profesor Siedlecki", wyposażone są w komputer, który może bezpośrednio odbierać sygnały z sond pomiarowych i automatycznie obliczać żądane wielkości, takie jak współczynniki absorpcji i rozpraszania światła, współczynniki osłabiania oświetlenia z głębokością, a także korelacje tych wielkości z innymi wielkościami opisującymi stan badanego środowiska wodnego. Tak dalece skomplikowana automatyzacja jest jednak opłacalna tylko przy masowych pomiarach przestrzennych.

Wszystkie morskie urządzenia pomiarowe, zarówno pracujące pod wodą, jak i nad wodą, muszą być szczególnie odporne na działanie niszczących czynników żywiołu morskiego (korozję, przyspieszenia wstrząsy i wibracje, ciśnienie hydrostatyczne, sól morska itp.).

Pomimo tych znacznych trudności technicznych, praktycznie wszystkie właściwości optyczne wody morskiej i mas wodnych w morzu badane są zdalnie *in situ* (w naturalnych warunkach w danym miejscu toni wodnej), a nie w laboratorium na próbkach wody. Jest to konieczność podyktowana szybkimi zmianami zachodzącymi w próbce wody morskiej po jej zaczerpnięciu, spowodowanymi działaniem bakterii i planktonu, obumieraniem komórek planktonu, koagulacją i opadaniem zawiesin, zmianami chemicznymi pod wpływem zmian oświetlenia, ciśnienia, temperatury itp. Niektóre trudniejsze pomiary można wykonać w laboratorium na pokładzie statku natychmiast po zaczerpnięciu próbki wody z morza.

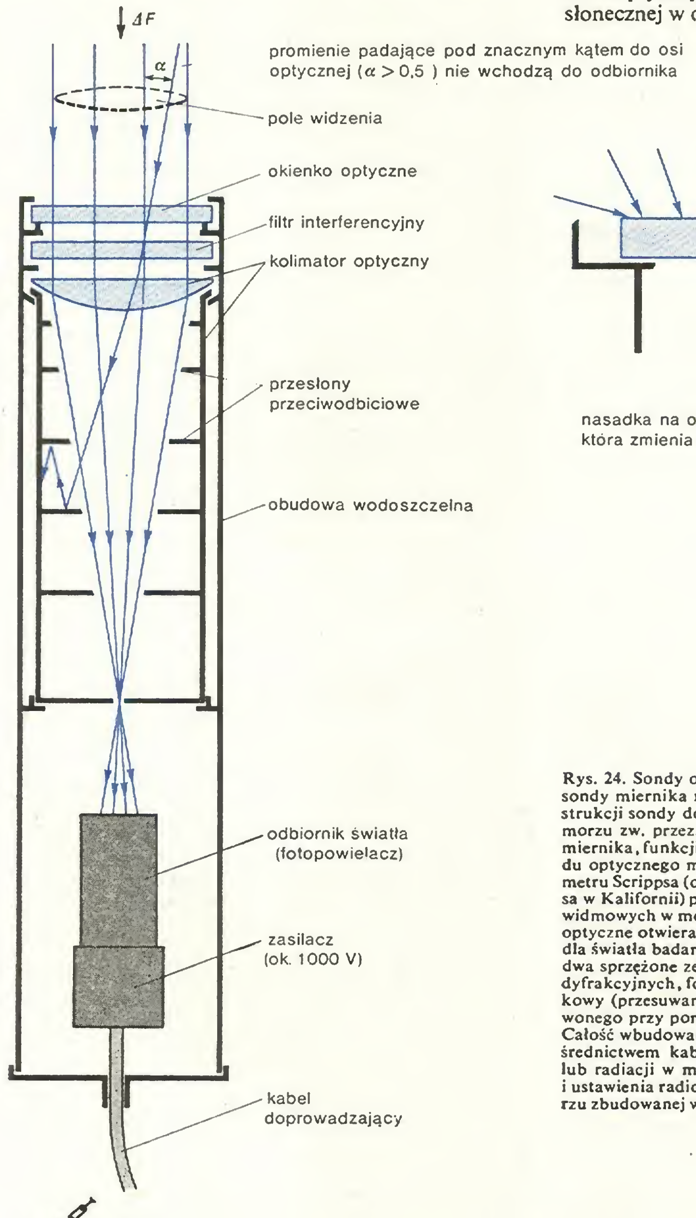
osiągnięcia naukowe

Postęp w rozwoju metod i techniki optycznych badań morza umożliwił rozpoznanie większości skomplikowanych procesów optycznych w morzu. Zbada-

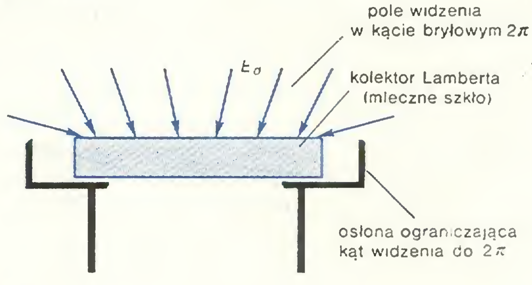
no absorpcję i rozpraszanie energii światła słonecznego w przestrzeni wodnej, powstawanie specyficznych pól podwodnego światła dziennego, o różnych właściwościach widmowych w różnych wodach i w różnych warunkach zewnętrznych (zależnych np. od położenia Słońca na niebie, od zachmurzenia, falowania powierzchni morza), zbadano selektywne oddziaływanie różnych składników wody morskiej ze światłem, polaryzację podwodnego światła, przenoszenie informacji w postaci optycznego obrazu przedmiotów zanurzonych w wodzie (widoczność podwodna, zastosowanie kamer telewizyjnych i fotograficznych), obraz i kolor morza widoczny z przestrzeni nadmorskiej (interpretacja zdjęć lotniczych i satelitarnych morza) i wiele innych.

obecne badania

Obecnie trwają intensywne badania i poszukiwania ilościowych związków między właściwościami składników wody morskiej i rozmieszczeniem tych składników w oceanach a optycznymi właściwościami morza oraz uzależnieniem tych właściwości od innych procesów przyrodniczych zachodzących w środowisku morskim. Opracowuje się modele matematyczne tych zjawisk, tak by na podstawie mniej licznych pomiarów można było przewidywać przebieg procesów optycznych i ich wpływ na przemiany energii słonecznej w olbrzymich przestrzeniach oceanów.

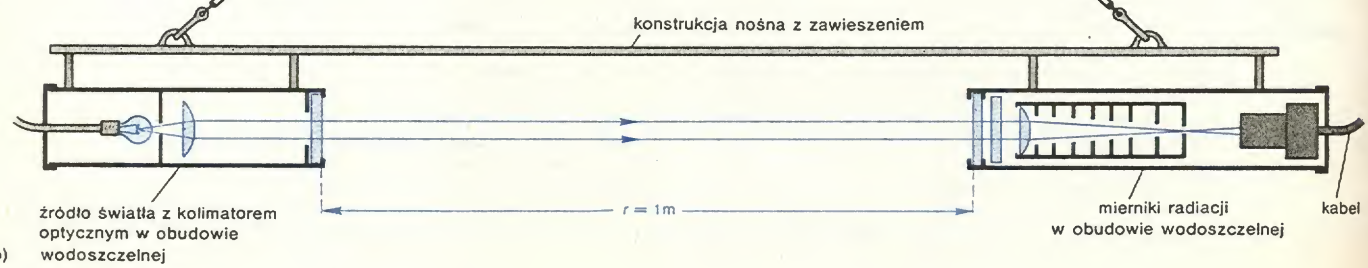


a)

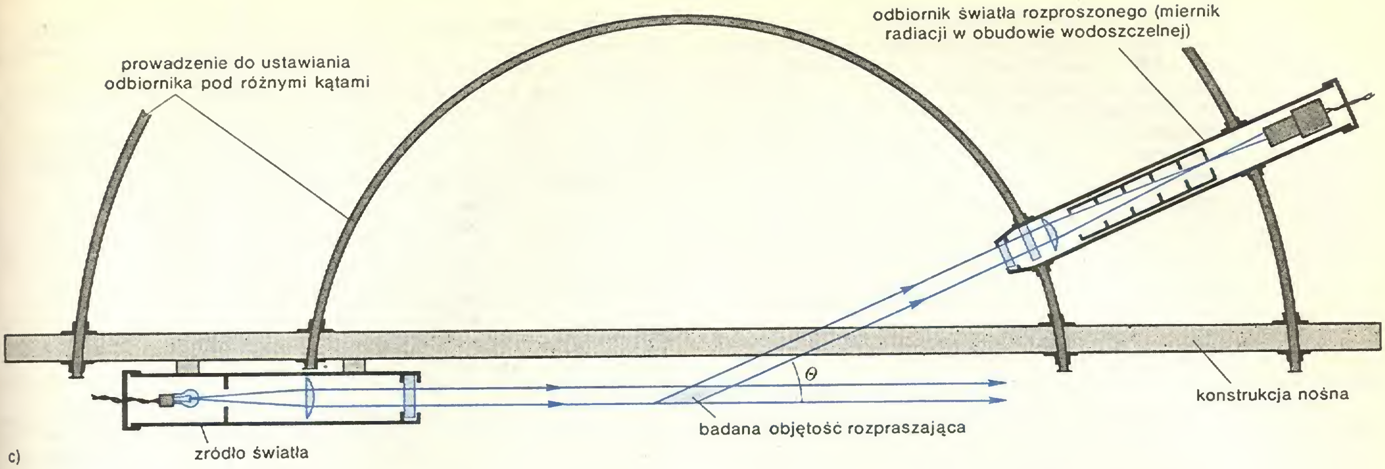


nasadka na okienko miernika radiacji która zmienia go w miernik oświetlenia

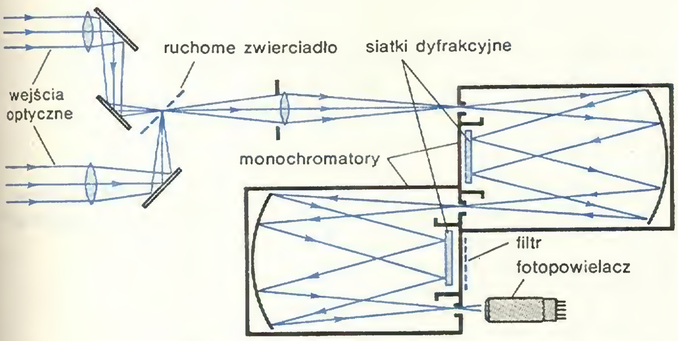
Rys. 24. Sondy optyczne do badań morza: a) Schemat przekroju sondy miernika radiacji i oświetlenia w morzu. b) Schemat konstrukcji sondy do pomiarów współczynnika osłabiania światła w morzu zw. przezroczomierzem morskim. c) Schemat konstrukcji miernika, funkcji rozpraszania światła w morzu. d) Schemat układu optycznego miernika radiacji i oświetlenia tzw. spektroradiometru Scrippsa (od nazwy słynnego Instytutu Oceanografii Scrippsa w Kalifornii) przeznaczonego do bardzo dokładnych pomiarów widmowych w morzu. Na schemacie rozróżnić można dwa wejścia optyczne otwierane alternatywnie przez ruchome zwierciadło (np. dla światła badanego i wzorcowego lub dla radiacji i oświetlenia), dwa sprzężone ze sobą monochromatory zbudowane na siatkach dyfrakcyjnych, fotopowielacz jako odbiornik światła oraz dodatkowy (przesuwany) filtr optyczny odcinający ślady światła czerwonego przy pomiarach w obszarze krótkofalowej części widma. Całość wbudowana jest w wodoszczelną obudowę i zdalnie za pośrednictwem kabla rejestruje automatycznie widma oświetlenia lub radiacji w morzu (J. E. Tyler, 1965 r.). e) Szkic konstrukcji i ustawienia radiopławy do zdalnych pomiarów optycznych w morzu zbudowanej w Zakładzie Oceanologii PAN w Sopocie w 1971 r.



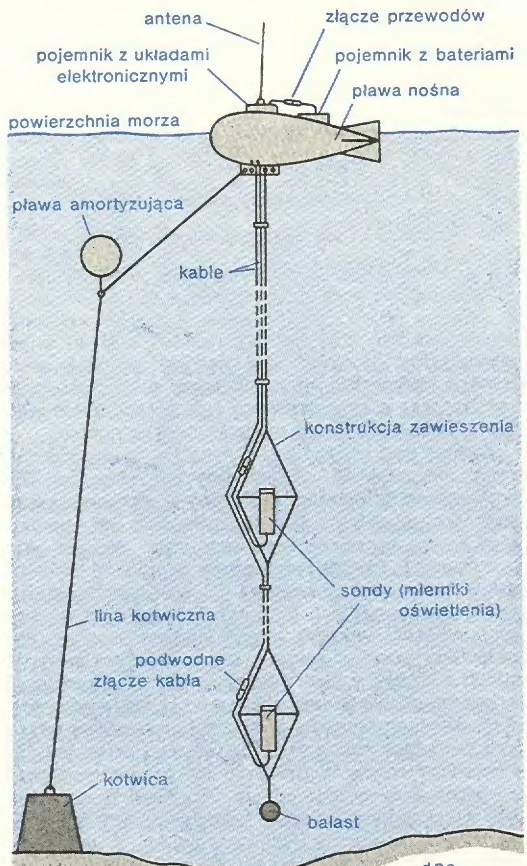
b)



c)



d)



e)

Właściwości optyczne morza rzeczywiste i pozorne

właściwości rzeczywiste i pozorne

Rzeczywiste właściwości optyczne morza zależą wyłącznie od natury samej wody morskiej i zawartych w niej składników, nie są zależne od zewnętrznych warunków oświetlenia morza. Do rzeczywistych właściwości optycznych zalicza się współczynnik załamania światła n, funkcję rozpraszania światła β opisującą kątowy rozkład rozpraszania, całkowity współczynnik rozpraszania światła b, współczynnik absorpcji a i współczynnik osłabiania (wiązki) światła c. Pozorne właściwości optyczne morza zależą zarówno od właściwości rzeczywistych, czyli od natury wody morskiej, jak i od warunków zewnętrznego oświetlenia morza. Pozornymi właściwościami optycznymi morza są np.: współczynnik dyfuzyjnego osłabiania oświetlenia odgórnego z głębokością K_d, współczynnik odbicia światła w toni wodnej (na skutek rozpraszania) R_d i kilkanaście innych.

radiacja

Podstawową wielkością fotometryczną stosowaną w optyce morza jest funkcja radiacyjna, zwana w skrócie radiacją L, która opisuje wartość strumienia energii światła F przychodzącego z dowolnego, lecz określonego kierunku w jednostkowym kącie bryłowym ω wokół tego kierunku i przypadającego na jednostkę powierzchni A_n prostopadłej do tego kierunku. Jeśli ten dowolny kierunek oznaczymy wektorem jednostkowym $\vec{\xi}$, to w myśl definicji i oznaczeń na rys. 25 radiacja

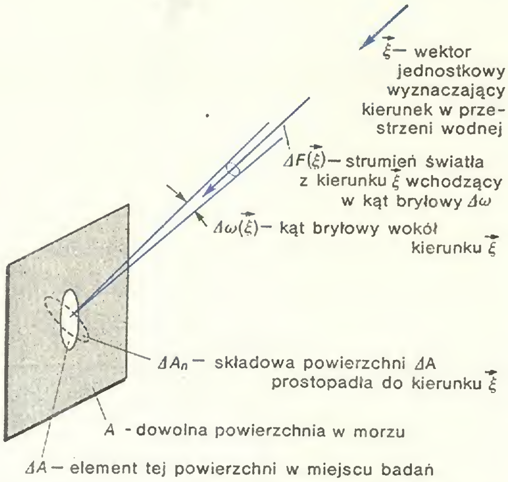
$$L(\vec{\xi}) = \frac{\Delta F(\vec{\xi})}{\Delta A_n \cdot \Delta\omega(\vec{\xi})}, \qquad (1)$$

co zapisuje się ściśle stosując różniczki:

$$L(\vec{\xi}) = \frac{d^2F(\vec{\xi})}{dA_n \cdot d\omega(\vec{\xi})}. \qquad (1a)$$

Promieniowanie słoneczne wchodzące do morza przenika w głąb toni wodnej, ulegając stopniowemu pochłanianiu przez składniki wody morskiej oraz rozpraszaniu na molekułach wody i na licznie występujących w morzu zawiesinach. W przestrzeni wodnej powstaje w rezultacie pole światła o ciekawych i złożonych właściwościach, które najdokładniej można by opisać zdefiniowaną wyżej funkcją radiacyjną, gdyż zmierzona lub obliczona dla wszystkich kierunków $\vec{\xi}$, funkcja ta informuje po prostu o wartościach liczbowych strumieni światła przychodzących z tych wszystkich kierunków w badanym miejscu. Jeślibyśmy znali wartości radiacji ze wszystkich kierunków, w każdym miejscu toni wodnej i w dodatku dla wszystkich długości fal światła, mielibyśmy pełną infor-

mację o rozkładzie energii światła i jej przenoszeniu w morzu. Na ogół takiej informacji nie można uzyskać, gdyż radiacja przybiera w morzu bardzo różne wartości, zależne od kierunku i głębokości, a rozkłady przestrzenne radiacji są bardzo różne w różnych akwenach i zmieniają się w zależności od pory roku, pory dnia i warunków meteorologicznych. Często bada się względne zmiany wartości radiacji z różnych kierunków $\vec{\xi}$ (pod różnymi kątami θ do pionu) tylko w płaszczyźnie padania promieni słonecznych, które pozwa-



Rys. 25. Szkic wyjaśniający definicję radiacji $L(\vec{\xi})$ zgodnie ze wzorem

lają wyznaczyć tzw. kątowy rozkład radiacji w tej płaszczyźnie. Rozkład taki zawiera również wiele ścisłych informacji o panującym w morzu polu światła oraz o absorpcyjnych i rozpraszających właściwościach wód, które otaczają badane miejsce.

kątowy rozkład radiacji w morzu

Kątowy rozkład radiacji wykreślony we współrzędnych biegunowych dla niedużych głębokości jest w słoneczne dni bardzo wydłużony w kierunku, z którego dochodzą bezpośrednie, załamane na powierzchni morza promienie słoneczne (rys. 26); jest przy tym znacznie bardziej wydłużony w wodach czystych niż w wodach zmętnionych — to wynik rozpraszania światła we wszystkich kierunkach w każdym miejscu toni wodnej, wskutek czego badane miejsce jest oświetlane również rozproszonym światłem ze wszystkich kierunków, lecz z każdego kierunku w stopniu zależnym od właściwości rozpraszających wody morskiej. W miarę wzrostu głębokości procentowy wkład światła rozproszonego i dyfundującego ze wszystkich kierunków rośnie w stosunku do osłabianych wraz z głębokością bezpośrednich promieni słonecznych i z tej przyczyny rozkład kątowy radiacji staje się na wykresie coraz bardziej zaokrąglony, a jego maksimum zbliża się do kierunku pionowego (zenitalnego). Na pewnej głębokości (np. 100, 200 lub 800 m — zależnie od przezroczystości wód) rozkład kątowy radiacji staje się symetryczny względem pionu, niezależnie od położenia słońca na niebie. Pole światła poniżej tej głębokości nazywa się polem asymptotycznym lub granicznym. Cechują je swoiste prawa, np. dalsze osłabianie radiacji lub oświetlenia z głębokością ma przebieg idealnie wykładniczy, jeśli koncentracja zawiesin i składników wody morskiej w otaczającej przestrzeni wodnej jest stała. Światło rozproszone jest też w znacznym stopniu spolaryzowane.

pole światła asymptotyczne

Pomiary kątowych rozkładów radiacji w morzu wymagają właśnie tych wspomnianych wyżej, najbardziej skomplikowanych urządzeń, tj. detektorów kierunkowych zaopatrzonych w śruby napędowe i stery (il. 202, tabl. 53) celem zdalnego regulowania ustawienia mierników radiacji w przestrzeni wodnej. Jest to bardzo złożone i rzadko spotykane urządzenie, lecz zmierzone za jego pomocą rozkłady radiacji pozwalają wyznaczyć wszystkie inne właściwości optyczne badanego obszaru morza.

detektor kierunkowy

W ostatnich latach do pomiarów kątowych rozkładów radiacji próbuje się zastosować metodę fotograficzną z wykorzystaniem specjalnego obiektywu, zwanego rybim okiem, o kącie widzenia 2π (bryłowym). Wykorzystanie tej metody wniesie zasadniczy postęp do pomiarów, gdyż pozwoli na jednej fotografii odwzorować rozkład kątowy radiacji z całej półsfery. Ciągle jednak istnieje problem dokładności takiego odwzorowania z uwzględnieniem widmowych właściwości podwodnego światła. Można by oczywiście zamiast kliszy fotograficznej umieścić pod takim obiektywem odpowiedni zestaw detektorów fotoelektrycznych, ale to by znacznie skomplikowało konstrukcję i działanie całego urządzenia.

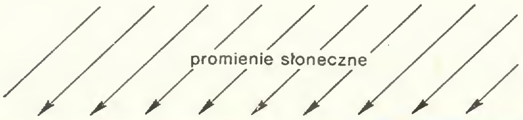
obiektyw „rybie oko"

W teoretycznym opisie właściwości optycznych morza przyjmuje się, najczęściej z dobrym przybliżeniem, model morza poziomo uwarstwionego, tzn. wychodzi się z założenia, że wielkości optyczne w płaszczyźnie poziomej na dowolnej głębokości są stałe, a zmieniają się jedynie z głębokością. W układzie współrzędnych prostokątnych, z płaszczyzną xy ustawioną poziomo i osią z skierowaną pionowo w dół, na dowolnej głębokości składowe radiacji przy tym założeniu są $L_x(\vec{\xi}) = L_y(\vec{\xi}) = \text{const}$, czyli radiacja $L(x, y, z, \vec{\xi}) = L(z, \vec{\xi})$ jest tylko funkcją głębokości i orientacji wektora $\vec{\xi}$ w przestrzeni.

model poziomo uwarstwionego morza

Często zakłada się także, że rozkład przestrzenny i kątowy radiacji (pole radiacji) w morzu nie zmienia się w pewnym okresie czasu (jest stacjonarny np. w ciągu 1 godziny, podczas której niewielkie są zmiany naturalnego oświetlenia powierzchni morza). Wówczas zmiany przestrzenne (osłabienie) radiacji $L(z, \vec{\xi})$ z kierunku $\vec{\xi}$ na odcinku drogi r opisuje się tzw.

stacjonarne pole radiacji



Rys. 26. Szkic rozkładów radiacji w morzu w płaszczyźnie padania promieni słonecznych — w wodach czystych (z lewej) i w wodach zmętnionych (z prawej). Przerywana linia wyznacza we współrzędnych biegunowych wartość radiacji (długość strzałki obrazującej porównawczo wartość radiacji $L(\vec{\xi})$ z dowolnego kierunku $\vec{\xi}$ (lub pod dowolnym kątem θ) w danym punkcie przestrzeni wodnej na głębokości z

równanie przenoszenia energii promienistej

równaniem przenoszenia energii promienistej, uproszczonym do postaci:

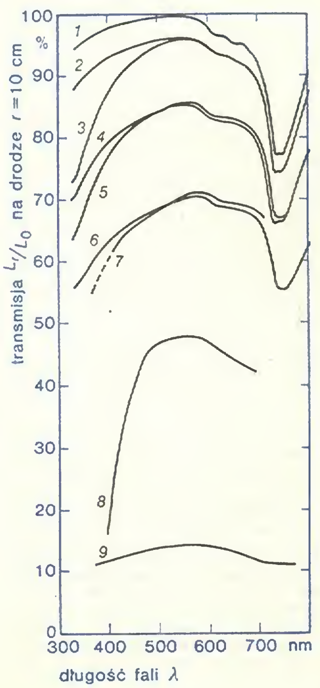
$$\frac{dL(z,\vec{\xi})}{dr} = -cL(z,\vec{\xi})+L_*(z,\vec{\xi})+L_n(z,\vec{\xi}). \quad (2)$$

W równaniu tym $dL(z,\vec{\xi})$ oznacza nieskończenie małe osłabienie radiacji na nieskończenie małym odcinku drogi dr (różniczka dL po dr), c oznacza rzeczywisty współczynnik osłabienia światła, L_* jest tzw. funkcją drogową, która opisuje ilość energii światła, jaka się dodaje do radiacji osłabionej na jednostkowym odcinku drogi r w wyniku rozpraszania w tym kierunku wszystkich innych strumieni światła dopływających z otaczającej przestrzeni wodnej, wreszcie L_n jest tzw. funkcją źródłową, która opisuje zwiększenie radiacji na jednostkowym odcinku drogi r spowodowane istnieniem źródeł światła na tej drodze, np. świecących mikroorganizmów, których wpływ może być znaczny na dużych głębokościach (poniżej 300 m) lub w nocy, tj. wówczas, kiedy udział bioluminescencji w oświetleniu toni wodnej jest znaczny w porównaniu z oświetleniem przez źródła zewnętrzne.

Jeżeli wiązkę światła biegnącą w kierunku $\vec{\xi}$ wzdłuż drogi r osłonimy od świateł dochodzących z innych kierunków ($L_* = 0$) i przyjmiemy, że nie ma w badanej wodzie źródeł światła ($L_n = 0$), to równanie (2) przybierze prostą postać, którą można zapisać w skrócie $dL/dr = -cL$ i która określa najprostszy sposób wyznaczania współczynnika osłabiania światła c. Scałkowanie powyższego równania różniczkowego wzdłuż drogi r daje rozwiązanie $L_r/L_0 = e^{-cr}$, gdzie L_0 jest wartością radiacji wchodzącej do warstwy wody, a L_r — wartością tej radiacji po przejściu odległości r w tej wodzie. Stosunek L_r/L_0, który można precyzyjnie mierzyć (w sposób pokazany na rys. 24b), nazywa się transmisją radiacji (na drodze r) lub transmisją równoległej wiązki światła, albo krótko — transmisją światła, a wartość tej transmisji dla warstwy wody $r = 1$ m nazywa się przezroczystością wody.

transmisja wiązki światła w wodzie

Przykładowe widma transmisji L_r/L_0 w różnych wodach naturalnych ilustruje rys. 27. Krzywa *1* na tym rysunku przedstawia typowy przebieg widma transmisji wiązki światła w stosunkowo czystej, lecz bardzo zasolonej (3,5% soli) wodzie oceanicznej.



Rys. 27. Typowe widma transmisji światła w różnych wodach naturalnych. *1* względnie czysta woda oceaniczna (Prąd Zatokowy w Cieśninie Florydzkiej), *2–7* wody szelfowe i przybrzeżne (*2–6* z okolic Florydy, *7* z Zatoki Gdańskiej), *8* wody ujściowe Wisły, *9* wody przybrzeżne silnie zmętnione w czasie sztormu (zatoka Biscayne)

Widmo to różni się od widma transmisji czystej wody destylowanej nieznacznie (w granicach błędu pomiaru), co świadczy o tym, że osłabianie przez sól morską światła w obszarze fal widzialnych i bliskiego nadfioletu jest niezauważalne. Z rysunku widać, że maksimum transmisji przypada w przedziale długości fal 470–550 nm. Spadek transmisji (spadek przezroczystości wody) w obszarze fioletu spowodowany jest w bardzo czystych wodach przede wszystkim rozpraszaniem światła na molekułach wody (rozpraszanie molekularne lub Rayleigha), którego natężenie jest odwrotnie proporcjonalne do czwartej potęgi długości fali ($I \sim 1/\lambda^4$). To rozpraszanie jest np. powodem pojawiania się ciemnobłękitnego koloru morza (także nieba) wtedy, gdy jego wody są bardzo czyste i panuje w nich dzienne światło rozproszone na molekułach.

pochłanianie światła czerwonego i podczerwonego

Natomiast spadek transmisji światła w obszarze czerwieni (długość fal powyżej 600 nm) i podczerwieni jest w czystych wodach spowodowany przede wszystkim intensywnym pochłanianiem fal o tej energii przez molekuły wody, w których wzbudzają się oscylacje atomów i rotacje całych molekuł — w efekcie zwiększa się ruch molekuł, czyli wzrasta temperatura. Jest to pochłanianie rezonansowe, tzn. energia pochłanianych kwantów świetlnych musi odpowiadać energii określonego sposobu drgań molekuły. Fale świetlne różnych długości mogą być przez wodę pochłaniane, bo istnieją różne sposoby drgań molekuł wody. Woda absorbuje najsilniej fale świetlne długości ok. 730 nm i fale świetlne z zakresu podczerwieni, co jest uwidocznione na wykresie. Z tej przyczyny światło czerwone, a tym bardziej podczerwone, pochłonięte zostaje w znacznym stopniu już w powierzchniowej kilkunastocentymetrowej warstwie wód w morzu, jak to widać z rys. 23. Tym się tłumaczy nagrzewanie powierzchniowej warstwy wód w słoneczne dni.

rozpraszanie przez zawiesiny

Wracając do przedstawionych na rys. 27 wykresów widm transmisji światła w różnych wodach naturalnych, zwróćmy uwagę na ich zróżnicowanie. Optyczne i inne, zarówno fizyczne, jak chemiczne badania tych wód wykazały, że przesuwanie się kolejnych wykresów w dół, czyli spadek transmisji w całym obszarze widma widzialnego, spowodowany jest absorpcją i przede wszystkim rozpraszaniem światła przez znajdujące się w wodzie zawiesiny. Wykresy 2–7 przedstawiają widma transmisji światła w wodach o coraz większej koncentracji zawiesin. Wobec dużej różnorodności zawiesin w wodach i ich dużych rozmiarów (na ogół znacznie większych od długości fali światła) rozpraszanie przez te zawiesiny wszystkich długości fal światła jest prawie jednakowo silne, co w efekcie obniża transmisję w całym przedziale widma światła — niezależnie od osłabiania z przyczyn wymienionych powyżej, gdy była mowa o czystej wodzie.

Poszczególne widma na wykresie różnią się także średnim nachyleniem po stronie fal krótkich. Jest to wyrazem zróżnicowania absorpcji światła w pasmie fioletu i ultrafioletu, spowodowanej zawartością w wodzie różnych organicznych substancji żółtych (humusy, melanoidy i in.). Im więcej żółtych substancji zawiera woda, tym mniej przepuszcza światła niebieskiego i fioletowego, co się szczególnie ostro zaznacza np. w wodach ujść rzecznych — w tym wypadku Wisły, a także w wodach Bałtyku (w porównaniu np. z Atlantykiem).

funkcja rozpraszania

Współczynnik osłabienia światła $s(z)$ na głębokości z w morzu jest sumą współczynnika absorpcji $a(z)$, czyli pochłaniania, i współczynnika rozpraszania $b(z)$, co można zapisać: $c(z) = a(z)+b(z)$. Wyznaczanie współczynników a i b w morzu jest nieco bardziej skomplikowane niż wyznaczanie c, przy czym b oznacza tzw. całkowity objętościowy współczynnik rozpraszania, który jest sumą kierunkowych współczynników rozpraszania, czyli sumą wartości funkcji rozpraszania molekularnego $\beta(\vec{\xi}_0, \vec{\xi})$, opisującej pro-

centowy rozkład kątowy rozpraszania strumienia światła biegnącego w kierunku $\vec{\xi}_0$ i rozpraszanego we wszystkich poszczególnych kierunkach $\vec{\xi}$ dookoła rozpraszającego elementu objętości wody. Ze względu na znaczenie rozpraszania w transmisji energii światła w morzu funkcja rozpraszania $\beta(\vec{\xi}_0, \vec{\xi})$ jest jedną z podstawowych wielkości charakteryzujących ośrodek wodny. Jej sens fizyczny i typowe wykresy ilustruje rys. 28.

Funkcja rozpraszania jest na wykresie symetryczna (jednakowe rozpraszanie w tył i w przód), a to dlatego, że w myśl teorii Rayleigha bardzo małe zbiory molekuł (dużo mniejsze od długości fali światła) zachowują się jak pojedyncze dipole elektryczne pobudzane do drgań elektrycznych przez fale światła padającego i emitujące we wszystkich kierunkach fale światła rozproszonego (jak antena dipolowa). Takie mikrozbiory molekuł powstają w wodzie w wyniku termicznych fluktuacji jej gęstości. Funkcja rozpraszania na zawiesinach morskich jest na wykresie bardzo wydłużona do przodu, gdyż zawiesiny te są na ogół większe od długości fali światła, w różne zatem obszary ich objętości dochodzi w tym samym czasie różne pole elektryczne padającej fali światła (różne natężenie i faza drgań), które pobudza te różne obszary do drgań elektrycznych w różnych fazach. Tak więc duża cząstka zachowuje się jak zbiór dipoli drgających w różnych fazach. Światło emitowane przez te dipole interferuje (z różnym efektem w różnych kierunkach), dając w rezultacie taki właśnie wydłużony do przodu wypadkowy rozkład natężeń, opisany funkcją β. Wokół kierunku padającego światła wszystkie funkcje rozpraszania są symetryczne, tak że ich wykresy przestrzenne są bryłami obrotowymi o takich samych przekrojach jak na rysunku.

Na rys. 28c przedstawiono typowe wykresy funkcji rozpraszania w oceanie we współrzędnych prostokątnych. Widać z nich, że w wodzie morskiej występuje bardzo silne rozpraszanie w małym kącie do przodu (bardzo trudne do zmierzenia) oraz minimum rozpraszania w bok pod kątami 90–100° — z przyczyn wyjaśnionych powyżej. Widać też, że procentowo mniej rozpraszane do przodu, lecz w całości bardziej rozpraszane (wyżej położona krzywa) jest światło niebieskie, co wskazuje na to, iż w całkowitym rozpraszaniu światła przez czyste wody Morza Sargassowego dużą rolę odgrywa rozpraszanie Rayleigha.

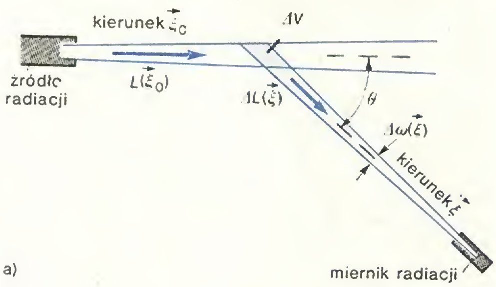
Dokładny wykres tej funkcji we współrzędnych biegunowych byłby tu w wypadku światła niebieskiego mniej wydłużony do przodu niż w wypadku światła czerwonego. Niestety — nie jest możliwe zmierzenie wartości funkcji rozpraszania, dla bardzo małych kątów (< 1°), a dla kątów małych (1, 2, 3°), gdzie rozpraszanie jest najsilniejsze, potrzebne jest zastosowanie specjalnej techniki ze światłem laserowym (il. 201, tabl. 53).

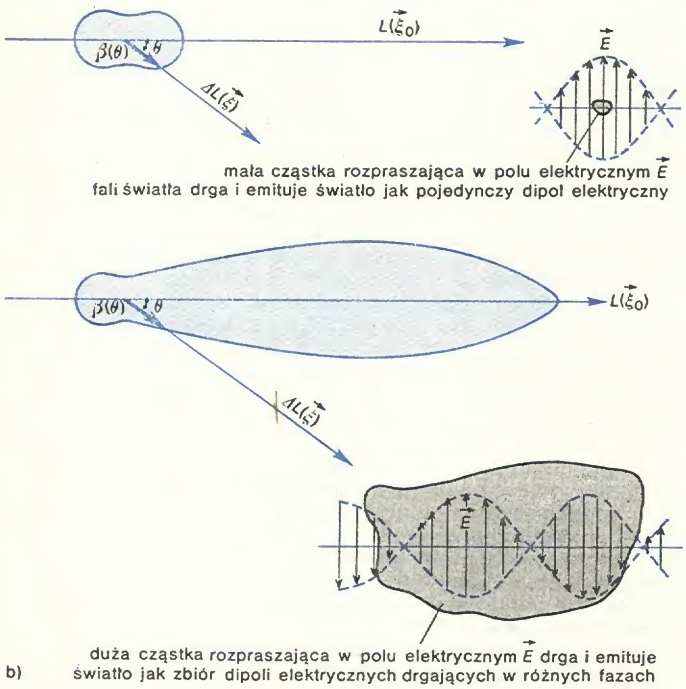
współczynnik rozpraszania

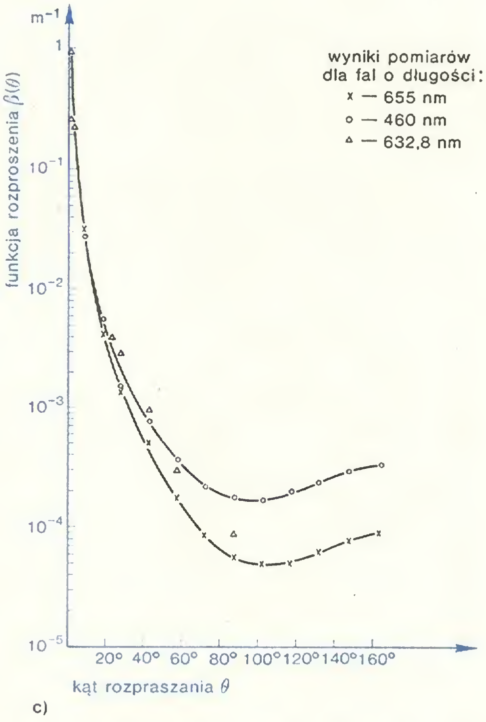
Całkowity współczynnik rozpraszania na głębokości z w morzu $b(z)$ jest sumą, a ściślej — całką funkcji $\beta(z, \vec{\xi}_0, \vec{\xi})$ po wszystkich kierunkach $\vec{\xi}$ sfery Ω otaczającej rozpraszający element objętości wody:

$$b(z) = \int_{\Omega} \beta(z, \vec{\xi}_0, \vec{\xi})\, d\omega(\vec{\xi}),$$

gdzie $d\omega(\vec{\xi})$ jest elementem kąta bryłowego wokół kierunku $\vec{\xi}$.







Rys. 28. Funkcja rozpraszania światła w wodzie morskiej. a) Sposób wyznaczania funkcji rozpraszania $\beta(\vec{\xi}_0, \vec{\xi}) \equiv \beta(\theta) = \Delta L(\vec{\xi})/L(\vec{\xi}_0)\Delta\omega\vec{\xi}_0$, gdzie $L(\vec{\xi}_0)$ radiacja ze źródła w kierunku $\vec{\xi}_0$, $\Delta L(\vec{\xi})$ część tej radiacji rozproszona w kąt bryłowy $\Delta\omega(\vec{\xi})$ w kierunku $\vec{\xi}$, θ kąt pomiędzy kierunkami $\vec{\xi}_0$ i $\vec{\xi}$. ΔV oznacza na rysunku badaną objętość rozpraszającą. b) Szkice wykresów (we współrzędnych biegunowych) funkcji rozpraszania na molekułach wody (u góry) i na zawiesinach morskich (u dołu) oraz rysunki (z prawej) wyjaśniające mechanizm rozpraszania. c) Funkcje rozpraszania zmierzone w Morzu Sargassowym przez G. Kullenberga (1968 r.) i wykreślone we współrzędnych prostokątnych dla 3 różnych długości fal światła

oświetlenie odgórne

Wielkością powszechnie mierzoną i stosowaną przy analizie optycznej morza jest tzw. oświetlenie odgórne $E_d(z)$ panujące na danej głębokości z. Charakteryzuje ono moc odgórnego strumienia promieniowania przenikającego w głąb morza. Oświetlenie odgórne na dowolnej głębokości z w morzu jest sumą wszystkich składowych radiacji z górnej półsfery, prostopadłych do elementu płaszczyzny poziomej z otaczającego rozpatrywany punkt w morzu, co można krótko zapisać:

$$E_d(z) = \int_{\Omega+} L_n(z, \vec{\xi})\, d\omega(\vec{\xi}), \tag{3}$$

gdzie Ω_+ oznacza półsferę górną wektorów jednostkowych $\vec{\xi}$, $L_n(z, \vec{\xi})$ — składową radiacji $L(z, \vec{\xi})$ prostopadłą do płaszczyzny poziomej na głębokości z.

Oświetlenie $E_d(z)$ można bezpośrednio mierzyć sondą wyposażoną w odbiornik światła z wejściem optycznym w postaci płytki z odpowiedniego mlecznego szkła (lub tworzywa), wystawionej tak, by padało na nią światło z całej górnej półsfery (rys. 24a). Taka płytka zwana jest kolektorem Lamberta lub kolektorem kosinusowym ($L_n = L\cos\theta$), ponieważ światło, które „widzi" odbiornik przez taki kolektor, jest automatycznie scałkowane z półsfery w myśl równania (3).

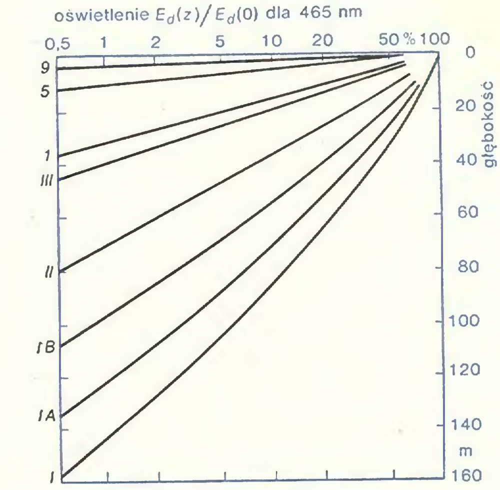
spadek oświetlenia z głębokością w morzu

Przebieg spadku oświetlenia odgórnego w miarę wzrostu głębokości zależy od pochłaniania i wielokrotnego rozpraszania światła w ośrodku. Z tej przyczyny procentowy spadek oświetlenia wraz ze wzrostem głębokości charakteryzuje dany akwen morski do tego stopnia, że ta okoliczność posłużyła — według koncepcji duńskiego oceanologa N. G. Jerlova — do optycznej klasyfikacji mórz i oceanów. Wyróżniono mianowicie 2 szeregi typów wód: szereg *I*, *IA*, *IB*, *II* i *III*, który nazwano typem wody oceanicznej, oraz szereg *1*, *2*, *3* ... *9*, nazwany typem wody przybrzeżnej, i tym sposobem scharakteryzowano poszczególne regiony oceanu światowego. Procentowy spadek oświetlenia odgórnego wybranej długości fali oraz widma transmisji oświetlenia w poszczególnych typach wód ilustruje rys. 29. Na rysunku jest także mapka przedstawiająca rozmieszczenie tych wód w oceanie światowym.

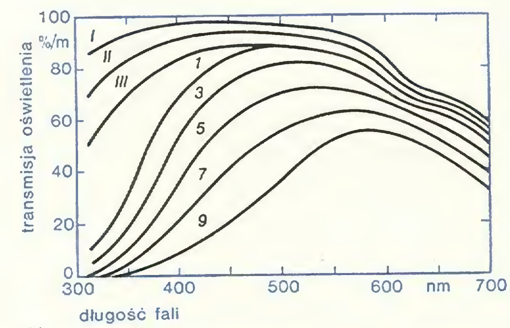
Nachylenie linii na wykresie (rys. 29a) wskazuje wielkość spadku oświetlenia na jednostkę wysokości kolumny wody. Dokładnie spadek ten opisuje się za pomocą tzw. współczynnika dyfuzyjnego osłabiania oświetlenia odgórnego:

$$K_d(z, \lambda) = -\frac{1}{E_d(z, \lambda)} \frac{dE_d(z, \lambda)}{dz},$$

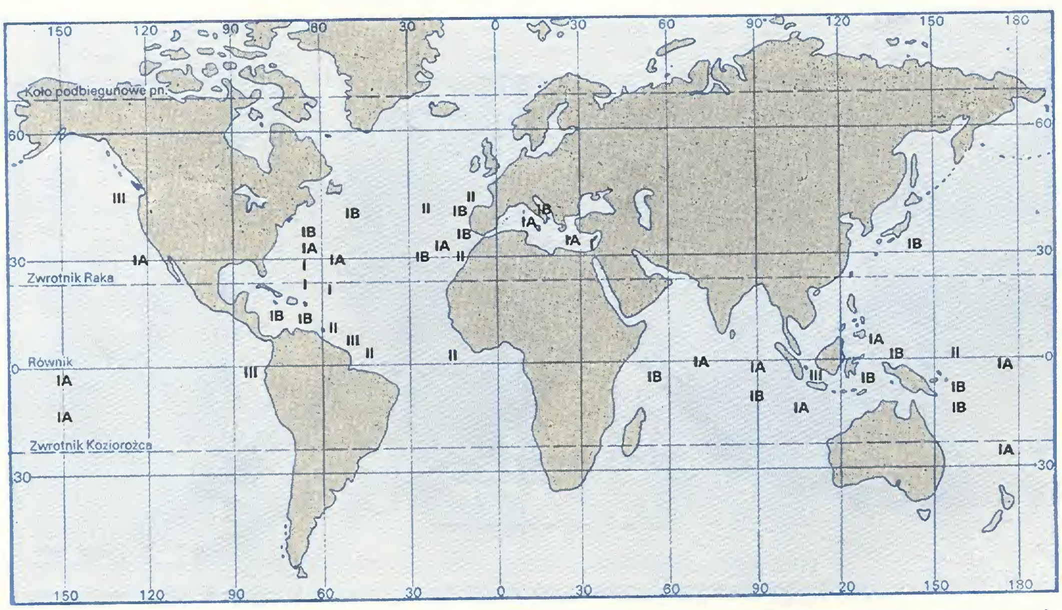
o którym wspomniano wyżej jako o jednej z ważnych pozornych właściwości optycznych morza. Po raz



a)



b)



Rys. 29. Optyczna klasyfikacja wód morskich wg N. G. Jerlova (1968 r.): a) Względny spadek oświetlenia z głębokością (światło niebieskie — 465 nm) w wodach oceanicznych typu *I*, *IA*, *IB*, *II* i *III* oraz wodach przybrzeżnych różnych typów *1*, *2*,..., *9*. b) Widma transmisji oświetlenia (tj. zmiany na drodze 1 m) w wodach różnych typów. c) Rozmieszczenie różnych typów wód w oceanie światowym

pierwszy wypisano tu wyraźnie, że oświetlenie $E_d(z, \lambda)$ i współczynnik jego osłabiania $K_d(z, \lambda)$ są funkcjami długości fali λ, lecz zależność od λ dotyczy także radiacji $L(z, \vec{\xi}, \lambda)$, współczynników $c(z, \lambda)$, $a(z, \lambda)$, $b(z, \lambda)$, $\beta(z, \vec{\xi}_0, \vec{\xi}, \lambda)$ oraz innych funkcji optycznych, a uproszczony zapis tych funkcji, bez zaznaczania zależności od λ, odnosi się umownie do światła monochromatycznego o dowolnie wybranej długości fali.

zróżnicowanie wód morskich

Poszczególne typy wód morskich różnią się wieloma właściwościami optycznymi (wyraźne różnice wykazują m.in. widma transmisji oświetlenia zilustrowane na rys. 29b). Pociąga to za sobą wielkie zróżnicowanie dobowych i rocznych bilansów energii światła słonecznego o różnej długości fal, dopływającej i zużywanej na procesy fizyczne i chemiczne na różnych głębokościach w morzu. Z tej również przyczyny zróżnicowane są warunki życia i samooczyszczania wód różnego typu: np. w wodach oceanicznych typu *I* do głębokości kilkudziesięciu metrów dociera jeszcze silny strumień ultrafioletu, który może uśmiercać niektóre mikroorganizmy, podczas gdy w wodach przybrzeżnych typu *1* (także w Bałtyku), a tym bardziej typu *2*, *3*, *4* itd., poniżej 10 m panuje już tylko słabe światło zielonożółte, a promieniowanie ultrafioletowe zostaje rozproszone i pochłonięte praktycznie w kilkucentymetrowej warstwie wody przypowierzchniowej.

Główną przyczyną zróżnicowania wód morskich pod względem optycznym są, jak już wspomniano, rozpuszczone w wodzie substancje organiczne, a głównie ich niektóre grupy (np. substancje humusowe, melanoidy i in.), zwane ogólnie substancjami żółtymi, oraz zawiesiny morskie. Ogólnie wiadomo, że substancje żółte są wynikiem metabolizmu organizmów morskich (także wnoszone są do morza przez rzeki) i że charakteryzuje je silna absorpcja ultrafioletu i fioletu, wskutek czego nadają wodzie żółtawe zabarwienie. Zawiesiny morskie mają wiele źródeł: są produktami kruszenia skał, pyłami z atmosfery, mikroorganizmami żywymi, szczątkami obumarłych roślin i zwierząt itp. Ich oddziaływanie ze światłem polega przede wszystkim na znacznym rozpraszaniu fal świetlnych wszystkich długości.

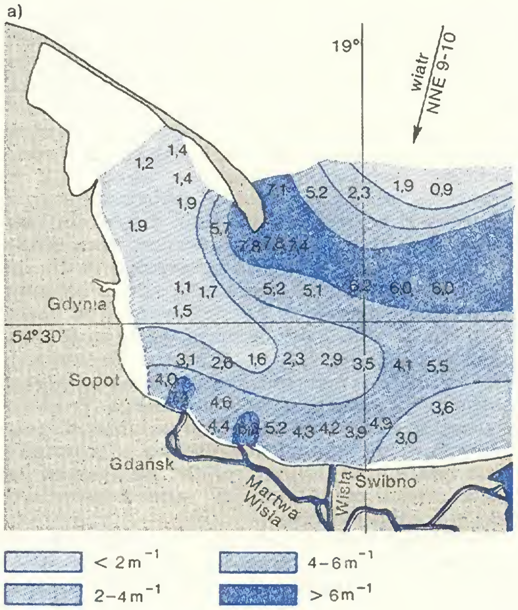
Mechanizm oddziaływania obu tych grup składników na pole światła oraz na transmisję energii promienistej w wodzie morskiej jest obecnie przedmiotem intensywnych badań.

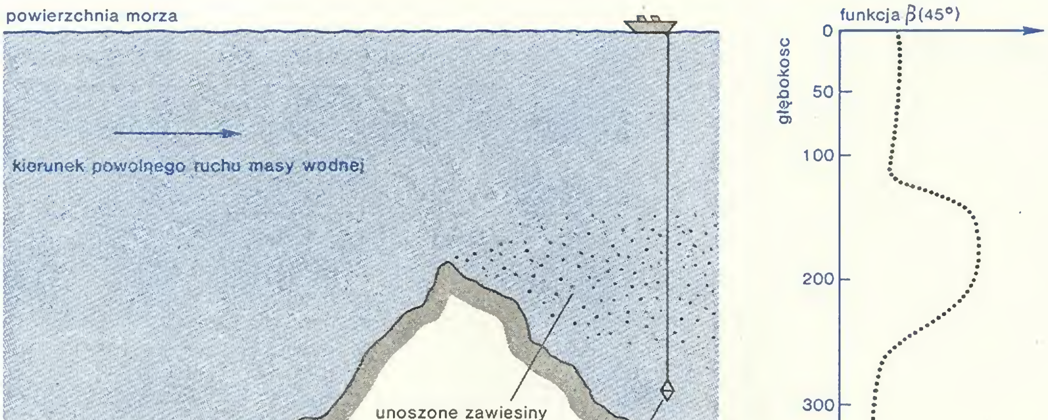
Zastosowanie optyki morza w innych dziedzinach oceanologii

Na zakończenie warto wspomnieć, że optyka morza jest szeroko stosowana w innych badaniach i pracach w morzu. Optyczne właściwości morza są często wyjątkowo czułymi wskaźnikami wielu zachodzących w oceanie procesów geofizycznych, geochemicznych, biologicznych i innych, których dokładne poznanie jest niezbędne dla racjonalnego gospodarowania w środowisku morskim i zapobiegania jego przeobrażeniom w niepożądanym kierunku. Bardzo czułe są np. optyczne wskaźniki rozmieszczenia i przenoszenia się mas wodnych różnego pochodzenia w morzu. Widać to szczególnie wyraźnie w rejonach ujść rzecznych, gdzie wody rzeczne, zawierające dużo zawiesin i substancji żółtych, rozpływają się i mieszają w skomplikowany często sposób z wodami morskimi. Mierząc rozkład przestrzenny np. współczynnika osłabiania światła *c* (szybko i na dużym obszarze wód), można

badanie ujść rzecznych

a)





b)

Rys. 30. Przykłady optycznych wskaźników rozmieszczenia i ruchu mas wodnych w morzu: a) Przykładowy rozkład wartości współczynnika *c* osłabiania światła zielonego ($\lambda = 0,5\,\mu m$) w powierzchniowej warstwie wód Zatoki Gdańskiej w czasie wiosennych wysokich stanów spływających wód Wisły. Duże wartości współczynnika *c* pokazują rozmieszczenie wody wiślanej, nie wymieszanej jeszcze z wodą morską. b) Szkic pomiaru pionowego rozkładu funkcji rozpraszania (pod jednym stałym kątem, np. $\theta = 45°$), który informuje o kierunku ruchu masy wodnej na podstawie rozpraszania światła przez zawiesiny unoszone z opływanej podwodnej góry

łatwo rozpoznać drogi rozchodzenia się wód ujściowych rzeki, a wraz z nimi zanieczyszczeń, ścieków itp. Przykład taki, ukazujący skomplikowane rozmieszczenie wód ujściowych Wisły w Zatoce Gdańskiej przy wysokim stanie tej rzeki w okresie wiosennym, przedstawiono na rys. 30a. Obszary wód, na których wartości współczynnika c są duże (6–7 m^{-1}), były zajęte przez bardzo zanieczyszczone wody wiślane nie wymieszane jeszcze z wodą morską. Widać, że w sprzyjających warunkach dotarły one aż do wysokości Półwyspu Helskiego (w warstwie powierzchniowej). W wyniku późniejszej zmiany wiatru i cyrkulacji wód głębinowych domieszka mniej zanieczyszczonej wody morskiej była w pobliżu ujścia rzeki dużo większa, a w Zatoce Puckiej w tym samym czasie — woda była prawie tak czysta jak na pełnym morzu. Podobne pomiary wykonane na różnych głębokościach wskazują na to, że woda wiślana rozpływa się w cienkiej warstwie powierzchniowej Zatoki Gdańskiej (1–2 m), a poniżej podpływają wody z otwartego morza Bałtyckiego.

badanie ruchu mas wodnych

Na wielkich przestrzeniach oceanu oprócz wyraźnych prądów morskich występuje bardzo powolne przemieszczanie się olbrzymich mas wodnych; ich wykrycie jest także często możliwe za pomocą metod optycznych. Kiedy np. taka powoli przemieszczająca się masa wodna opływa podwodne góry, to po drugiej stronie gór w kierunku przepływu notuje się wyraźny wzrost rozpraszania światła, spowodowany zawiesinami unoszonymi z osadów dennych na podwodnej górze, jak to obrazuje rys. 30b.

badanie strefy eufotycznej

Od optycznych właściwości wód danego regionu morza w znacznym stopniu zależy szybkość, wydajność energetyczna i pionowy rozkład produkcji materii organicznej, która ma miejsce w procesie fotosyntezy w komórkach fitoplanktonu morskiego. Grubość wystarczająco naświetlonej powierzchniowej warstwy wód, w której się ta produkcja odbywa z przewagą wydzielania tlenu nad jego zużyciem przez komórki, nazywa się strefą eufotyczną w morzu. Strefa ta dochodzi w czystych morzach do ok. 150 m głębokości, podczas gdy np. w Bałtyku zaledwie do ok. 20 m, co jest wynikiem różnic we właściwościach widmowych tych wód. W dodatku różne widma panującego w różnych wodach światła przyczyniają się do różnicowania nie tylko gatunków żyjącego w nich fitoplanktonu, lecz także składu barwników w komórkach tego samego gatunku. Komórki te mają bowiem zdolność fotoadaptacji, tzn. zmieniają skład swoich barwników wyłapujących dostępną dla fotosyntezy energię światła pod wpływem zmian w zewnętrznych warunkach oświetleniowych (np. podczas zmian zanieczyszczeń wody, a także w czasie powolnego tonięcia i znajdowania się w coraz głębszych wodach, gdzie panuje inne pole światła). Niekiedy jednak duże zmiany oświetlenia, spowodowane np. ogniskowaniem promieni słonecznych przez fale morskie, mogą znacznie hamować proces fotosyntezy.

Tak więc warunki optyczne mają zasadniczy wpływ na wzrost fitoplanktonu, a stąd — na podtrzymywanie wszelkich form życia w morzu, gdyż jak już wspomniano, wyprodukowana przez fitoplankton w procesie fotosyntezy materia organiczna i tlen są niezbędne do życia wszystkich organizmów morskich. Badanie pola światła jest zatem również bezpośrednio przydatne do określania warunków życia w danym morzu i sygnalizowania ich zmian.

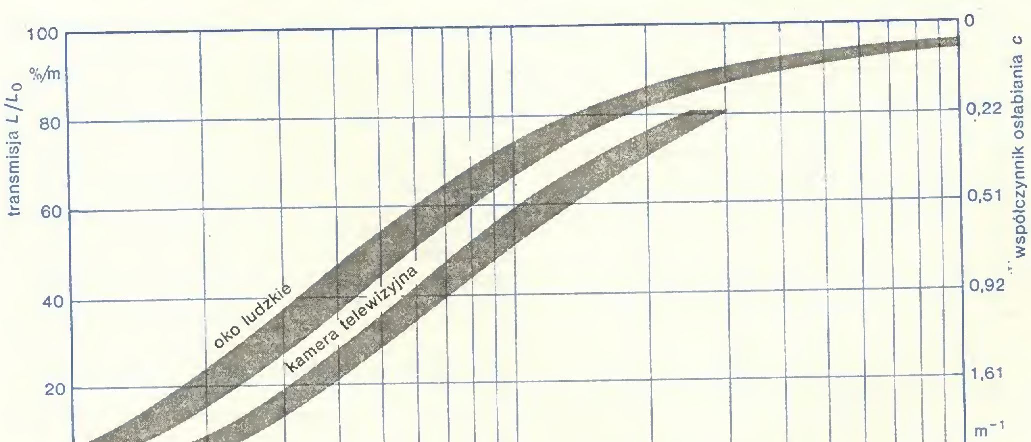
zdalne metody optyczne

W ostatnich latach coraz częściej wykorzystuje się do różnych badań morza zdjęcia powierzchni morza wykonane z samolotów i sztucznych satelitów oraz pomiary radiacji powierzchni morza. Interpretacja tych zdjęć, a zwłaszcza badanie na podstawie radiacji powierzchni morza różnych właściwości fizycznych, chemicznych i biologicznych dużych przestrzeni morskich, wymaga wielu szczególnie dokładnych informacji z optyki morza, w tym bowiem celu potrzebna jest dokładna znajomość zależności rozkładów radiacji światła wychodzącego z morza od wielu złożonych czynników środowiskowych. Te zdalne metody optyczne wnoszą jednak kolosalny postęp w rozpoznanie pól temperatur powierzchni morza, pól lodowych, rozmieszczenia planktonu (pośrednio ryb), prądów morskich i in. na wielkich obszarach oceanu. Z tej przyczyny prowadzone w tym kierunku badania optyczne rozwijają się obecnie bardzo dynamicznie.

zastosowanie optyki w inżynierii morskiej

W wielu pracach podwodnych związanych z nowymi kierunkami inżynierii morskiej (budownictwo morskie, górnictwo morskie, nowe techniki połowów itp.) niezbędne jest wykorzystywanie telewizyjnych i fotograficznych kamer podwodnych oraz świateł i szkieł optycznych umożliwiających obserwację różnych obiektów. Ze względu na mały zasięg widzenia (rys. 31) i zmiany barwy przedmiotów w wodzie konieczne jest rozpoznanie danych optycznych warunków środowiskowych i na tej podstawie dobieranie parametrów stosowanych urządzeń optycznych, takich jak widmo emisji sztucznych źródeł oświetlenia, filtry optyczne i polaryzacyjne w odbiornikach, geometria ustawienia układu itp. Odpowiedni dobór tych parametrów pozwala znacznie zwiększyć zasięg widzenia i kilkakrotnie poprawić kontrast i jakość obrazu.

Podstawowe znaczenie procesów oddziaływania światła słonecznego na środowisko morskie oraz praktyczne możliwości wykorzystania optyki morza inspirują współcześnie wiele badań doświadczalnych



Rys. 31. Zasięg widzenia w wodzie dla oka ludzkiego i dla kamery telewizyjnej (wg R. O. Briggs i G. L. Hatchett, 1965 r.)

i teoretycznych i przyczyniają się do dynamicznego rozwoju tej gałęzi fizyki. Badania w tym kierunku prowadzą liczne współpracujące ze sobą instytuty naukowe w świecie, szczególnie w ZSRR, Danii, USA, we Francji, w Japonii i NRD, a także w Polsce — w Zakładzie Oceanologii Polskiej Akademii Nauk w Sopocie.

J. Dera *Charakterystyka oświetlenia strefy eufotycznej w morzu*, Oceanologia nr 1, 1971; J. Dera, J. Kalinowski *Przenoszenie energii promienistej w morzu*, Post. Fiz. 17, 537 (1966); J. Dera, J. Olszewski *Widzialność podwodna*, Post. Fiz. 20, 473 (1969); *Fizyka morza*, Studia i Materiały Oceanologiczne Komitetu Badań Morza PAN nr 6 i nr 7, 1973; N. G. Jerlov *Marine Optics*, Amsterdam 1976; J. Kalinowski, J. Dera *Metody badań zjawisk optycznych w morzu*, Post. Fiz. 19, 219 (1968); *Optical Aspects of Oceanography* (N. G. Jerlov, E. Steemann Nielsen, ed.), London 1974; W. W. Szulejkin *Fizyka moria*, Moskwa 1968.

Akustyka morza

Antoni Śliwiński

Akustyka morza, inaczej: akustyka podwodna, hydroakustyka, obejmuje całokształt zjawisk i procesów dotyczących rozchodzenia się fal sprężystych w dużych zbiornikach wodnych (w hydrosferze). Zjawiska akustyczne w morzu odpowiadają bardzo szerokiemu widmu fal, o częstościach od rzędu ułamków herca (infradźwięki) aż do kilkuset kiloherców (ultradźwięki). Górną granicę tego widma wyznacza natura środowiska morskiego, które silnie tłumi fale o dużych częstościach, a więc fale krótkie.

zakres przenoszonych fal

Środowisko morskie bardzo dobrze nadaje się do generacji fal akustycznych i przesyłania informacji za ich pomocą. Pasmo częstości użyteczne ze względu na komunikację podwodną i hydrolokację rozciąga się w granicach 10–100 kHz. Fale takie można wytwarzać w morzu za pomocą przetworników elektroakustycznych, wykorzystujących zjawiska piezoelektryczne lub magnetostrykcyjne. Często też stosuje się impulsowe źródła akustyczne, którymi są podwodne eksplozje materiałów wybuchowych, eksplozje sprężonych gazów lub też wyładowania elektryczne dużej mocy.

szumy własne

Niezależnie od pól akustycznych wytwarzanych celowo istnieją w morzu pola akustyczne stanowiące tło, tzw. szumy własne, pochodzące z różnego rodzaju źródeł, zarówno naturalnych, jak i sztucznych, związanych z działalnością człowieka na morzu.

Zagadnienia akustyki morza są przedmiotem zainteresowania od kiludziesięciu lat i poświęca się im wiele prac naukowych, bada się fizyczne zjawisko rozchodzenia się fal dźwiękowych w morzu i szuka się możliwości wykorzystania go.

badania w Bałtyku

Badaniem akustycznych zjawisk w hydrosferze zajmuje się obecnie wiele ośrodków na świecie, także w Polsce. Badania, które prowadzi się w Bałtyku, obejmują głównie następujące zagadnienia:

charakter i zasięg rozchodzenia się dźwięków w morzu,

rozkład prędkości rozchodzenia się dźwięku wywołanego refrakcją i występowanie z tym związanych tzw. kanałów akustycznych,

rozpraszanie fal akustycznych w morzu i pogłos,

tłumienie fal akustycznych w morzu,

dyspersja fal akustycznych i jej rola w przenoszeniu sygnałów ciągłych i impulsowych,

szumy własne morza — ich natężenie i skład widmowy,

wykorzystanie fal akustycznych.

Istnieją pewne prawidłowości w przebiegu zjawisk akustycznych we wszystkich morzach i oceanach, ale jednocześnie w każdym zbiorniku wodnym zjawiska te mają specyficzny charakter, inny w morzach głębokich, inny w płytkich; przebieg ich zależy od szerokości geograficznej, warunków klimatycznych, pór roku, zmian dobowych oraz wielu dodatkowych czynników naturalnych i modyfikowanych przez człowieka. Do różnego rodzaju badań akustycznych, zarówno podstawowych, jak i stosowanych, skłaniają nie tylko rozległe przestrzenie hydrosfery i duża liczba czynników wpływających na przebieg zjawisk, ale również sama natura fal sprężystych, które jak dotąd — stanowią najlepsze narzędzie do przenoszenia sygnałów podwodnych na duże odległości, tysięcy metrów a nawet tysięcy kilometrów (zależy to od częstości sygnału i warunków fizycznych w danym akwenie), a więc wielokrotnie większe niż zasięgi fal elektromagnetycznych, które w morzu są silnie tłumione.

Charakter rozprzestrzeniania się dźwięku

Warunki, w jakich się rozchodzi dźwięk w morzu, różnią się znacznie od idealnych, z jakimi mamy do czynienia w płynnych środowiskach jednorodnych (→ Przedmiot i zakres akustyki), przy rozpatrywaniu których wychodzi się z założenia, że w całej objętości własności sprężyste są takie same. Niejednorodne własności sprężyste toni morskiej wypływają ze zmian temperatury, zasolenia, ciśnienia hydrostatycznego, ruchu mas wodnych (Rozdz. Dynamika morza). Ma też znaczenie zmieniająca się zawartość ciał obcych w wodzie, np. pęcherzy powietrza, których źródłem powstawania w toni morskiej są: falująca powierzchnia morza, różne procesy fizykochemiczne, a także organizmy żywe. Wszystkie te czynniki decydująco wpływają na prędkość, tłumienie i kierunek rozchodzenia się dźwięku w morzu.

Ponieważ zasięg sygnałów akustycznych rozchodzących się w morzu jest daleki, duży wpływ na charakter rozchodzenia się dźwięku mają ograniczenia zbiorników wodnych, w szczególności brzegi, dno i powierzchnia morza.

potencjał akustyczny

Często pole akustyczne w morzu podobnie jak w innych ośrodkach opisuje się funkcją zwaną potencjałem prędkości drgań cząstki akustycznej. Potencjał akustyczny Φ definiuje się jako skalarną funkcję określającą pole, której pochodna przestrzenna (gradient) daje bezpośrednio mierzalną wielkość wytwarzającą to pole. W polu akustycznym wielkością tą może być przesunięcie $\vec{\xi}$ (→ Przedmiot i zakres akustyki, rozdział Teoria ośrodków ciągłych i fale sprężyste), albo prędkość cząstki $\dot{\vec{\xi}} = \partial\vec{\xi}/\partial t$, albo też ciśnienie akustyczne p. Jeżeli Φ oznacza potencjał prędkości $\dot{\vec{\xi}}$, wtedy $\dot{\vec{\xi}} = \operatorname{grad} \Phi$, co z kolei oznacza, że $p = \varrho c \operatorname{grad} \Phi$.

Potencjał $\Phi(x, y, z, t)$ spełnia równanie falowe

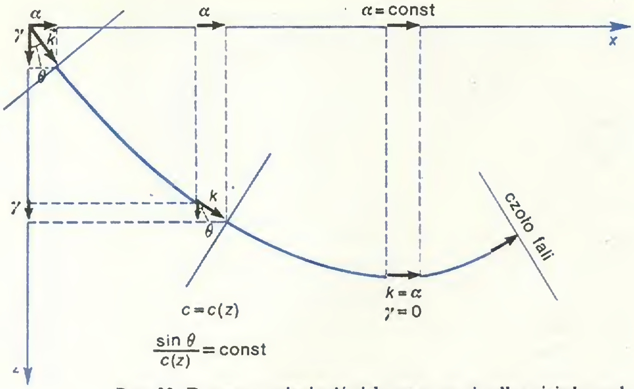
$$\frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial y^2} + \frac{\partial^2\Phi}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2}, \qquad (4)$$

przy czym c jest prędkością rozchodzenia się dźwięku, a symbole $\partial^2/\partial x^2, \partial^2/\partial y^2, \partial^2/\partial z^2, \partial^2/\partial t^2$ oznaczają pochodne cząstkowe po odpowiednich zmiennych. Równanie to jest słuszne tylko w odniesieniu do fali nie tłumionej, tzn. że można je stosować tylko w przybliżeniu, gdyż w rzeczywistości, jak już wyżej wspomniano, fala akustyczna w toni morskiej ulega tłumieniu. Do przybliżonego opisu pola akustycznego w morzu używa się modelu poziomo uwarstwionych mas wodnych, w którym prędkość dźwięku c jest funkcją tylko głębokości morza $c = c(z)$. Wynika on stąd, że, w toni morskiej zarówno temperatura wody, gęstość i zasolenie, jak i ciśnienie hydrostatyczne zależą regularnie od głębokości (wpływ pola ciężkości Ziemi).

Przy założeniu uwarstwionego modelu potencjał akustyczny jest tylko funkcją współrzędnych x i z oraz czasu t. Wtedy rozwiązaniem równania falowego jest funkcja opisująca falę płaską, a ogólna jej postać jest następująca:

$$\Phi(x, z, t) = \varphi(z)e^{i(\omega t \pm \varkappa x)}, \qquad (5)$$

gdzie $\varphi(z)$ oznacza przestrzenną część potencjału (funkcja ta określa rozkład potencjału wzdłuż osi z), ω — częstość kołową fali, znak $-$ lub $+$ odpowiada dodatniemu lub ujemnemu kierunkowi rozchodzenia się fali wzdłuż osi x z prędkością fazową $c' = \omega/\alpha$ a α jest liczbą falową dla kierunku x (rys. 32).



Rys. 32. Droga promienia dźwiękowego w ośrodku niejednorodnym

Po podstawieniu do równania (4) rozwiązania (5) (biorąc tylko znak $-$, oznaczający dodatni kierunek rozchodzenia się fali) otrzymuje się równanie określające funkcję $\varphi(z)$:

$$\frac{\partial^2\varphi}{\partial z^2}+\gamma^2\varphi = 0, \tag{6}$$

gdzie $\gamma^2 = \frac{\omega^2}{c^2}-\alpha^2 = \alpha^2\left(\frac{c'^2}{c^2}-1\right)$.

Rozwiązanie równania (6) ma ogólną postać następującą:

$$\varphi = Ae^{-i\gamma z}+Be^{i\gamma z},$$

gdzie A i B są stałymi całkowania. Pierwszy składnik opisuje zaburzenie rozchodzące się w dół (w dodatnim kierunku osi z), drugi — w górę. Jeśli się ograniczymy do zaburzenia w dół, to rozwiązanie (2) można będzie napisać w postaci:

$$\Phi(x, z, t) = Ae^{-i\gamma z}e^{i(\omega t-\alpha x)}.$$

Dla fali rozchodzącej się nie w kierunku osi x, lecz w płaszczyźnie x, z rozwiązanie będzie miało postać:

$$\Phi(x, z, t) = Ae^{i(\omega t-kr_z)},$$

gdzie $r_z = \sqrt{x^2+y^2}$, a $k^2 = \alpha^2+\gamma^2 = \omega^2/c^2$,

k jest liczbą falową fali wypadkowej, której kierunek rozchodzenia się wyznacza wektor falowy $\vec{k}$, przy czym $|\vec{k}| = k$, α i γ są liczbami falowymi fal składowych biegnących odpowiednio wzdłuż osi x oraz z. Wektory $\vec{\alpha}$ i $\vec{\gamma}$ są składowymi wektora $\vec{k}$ dla tych kierunków (sytuację ilustruje rys. 32). Zachodzi przy tym zależność: $\alpha = k\sin\theta$, $\gamma = k\cos\theta$, gdzie θ jest kątem między osią z a wektorem $\vec{k}$. W środowisku poziomo uwarstwionym θ odpowiada kątowi załamania na granicy dowolnych dwu warstw.

W środowisku zupełnie jednorodnym, w którym $c = \text{const}$, również wektor $\vec{k}$ jest stały i jego wartość $k = \omega/c = 2\pi f/c = 2\pi/\lambda$, gdzie λ — długość fali, f — częstość. Natomiast w morzu, dla którego przyjęliśmy poziome uwarstwienie i $c = c(z)$, wielkości k, λ i γ są funkcjami głębokości z, podczas gdy α pozostaje stałe. Wynika to z wyżej rozpatrzonych zależności i z prawa Snella, określającego współczynnik załamania „promienia" dźwiękowego:

$$n = \frac{\sin\theta}{\sin\theta_0} = \frac{c(z)}{c_0},$$

skąd wynika, że

$$\frac{\sin\theta}{c(z)} = \text{const.}$$

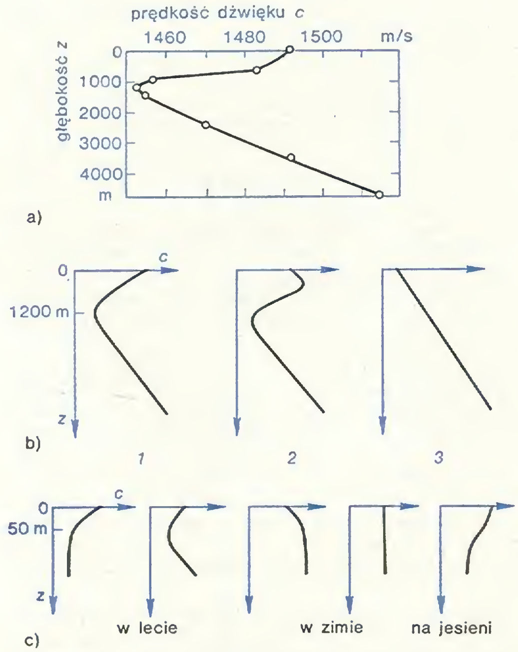
Jeśli fala biegnie w kierunku poziomym ($\theta = \pi/2$) $\alpha = k$ i $\gamma = 0$, wtedy (por. rys. 32)

$$c' = c(z). \tag{7}$$

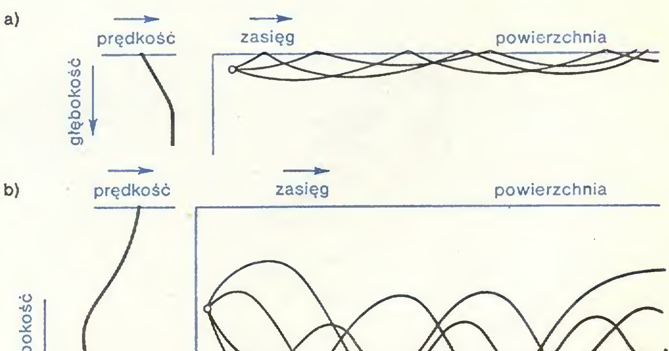
refrakcja fal dźwiękowych

Jeżeli fala biegnie na określonej głębokości z, a wektor $\vec{k}$ tej fali przybiera kierunek poziomy, to promień fali w dalszym swym biegu zmienia zwrot: przestaje biec w dół, zawraca ku górze, lub w przeciwnym przypadku — przestaje biec ku górze i zawraca w dół. Obydwa te przypadki zachodzą w morzu, a zjawisko nazywa się refrakcją fal dźwiękowych i prowadzi do powstawania tzw. kanałów dźwiękowych, zwanych z angielskiego SOFAR (Sound Fixing and Ranging).

kanały dźwiękowe



Rys. 33. Pionowe rozkłady prędkości dźwięku w morzu: a) typowy rozkład wg W. W. Szulkina; b) rozmaite rozkłady w morzu głębokim; *1* rozkład typowy w oceanie, temperatura wody przy powierzchni wyższa niż głębiej, *2* sezon zimowy w strefie umiarkowanej, *3* morze arktyczne, c) rozkłady przy powierzchni w morzach płytkich w zależności od pory roku



Rys. 34. Schemat biegu promieni dźwiękowych w morzu głębokim: a) w kanale przypowierzchniowym, b) w kanale głębinowym

Zjawisko to odpowiada znanemu w optyce zjawisku całkowitego wewnętrznego odbicia na nieciągłej granicy dwóch ośrodków przy przejściu światła z ośrodka optycznie gęstszego do rzadszego, co zachodzi wtedy, gdy przy danej wartości współczynnika załamania kąt padania w ośrodku optycznie gęstszym przekracza wartość kąta granicznego. Tutaj w wodzie morskiej sytuacja jest jednak ogólniejsza, a przy tym zjawisko może wystąpić na różnych głębokościach i przy różnych kątach padania. Wskutek ciągłej zmiany współczynnika załamania wraz z głębokością (ciągła zmiana prędkości $c(z)$ związana ze zmianą gęstości wynikającej z gradientu temperatury, rys. 33 i 34 kierunek rozchodzenia się promienia dźwiękowego zmienia się od punktu do punktu (rys. 32).

Prędkość rozchodzenia się dźwięku

Zróżnicowanie prędkości rozchodzenia się dźwięku w toni morskiej na ogół komplikuje i znacznie wydłuża drogę dźwięku od źródła do odbiornika w porównaniu z linią prostą.

Prędkość dźwięku jest bezpośrednio powiązana ze ściśliwością i gęstością środowiska, co przy założeniu, że rozchodzenie się fal akustycznych ma charakter procesu adiabatycznego, wyraża wzór:

$$c = \sqrt{\frac{1}{\beta_{ad}\varrho}} = \sqrt{\frac{\varkappa}{\beta_{iz}\varrho}}, \tag{8}$$

gdzie β_{ad} — współczynnik ściśliwości w procesie adiabatycznym, β_{iz} — współczynnik ściśliwości w procesie izotermicznym, $\varkappa = c_p/c_V$ ($\varkappa$ wody morskiej wynosi 1,00–1,02), c_p — ciepło właściwe przy stałym ciśnieniu, c_V — ciepło właściwe przy stałej objętości, ϱ — gęstość.

Ściśliwość wody morskiej jest mniejsza od ściśliwości wody czystej i maleje zarówno ze wzrostem ciśnienia (głębokości), jak i temperatury oraz zasolenia. Na przykład przy temperaturze $T = 0°C$ i zasoleniu $S = 34{,}85‰$ w warstwie powierzchniowej ściśliwość wody morskiej ma wartość $4{,}658 \cdot 10^{-5}$ cm²/kG, a na głębokości 10 000 m — $3{,}993 \cdot 10 \cdot 10^{-5}$ cm²/kG. Gęstość wody morskiej jest w porównaniu z wodą czystą trochę większa i maleje ze wzrostem temperatury, natomiast rośnie ze wzrostem zasolenia i ciśnienia.

Zależność prędkości dźwięku od zasolenia, temperatury i ciśnienia w wodzie morskiej ujmują różne wzory empiryczne, z których najbardziej znany w oceanografii jest wzór Wilsona:

wzór Wilsona

$$c(S, T, P) = c(35, 0, 0) + c_S + c_T + c_P + c_{STP},$$

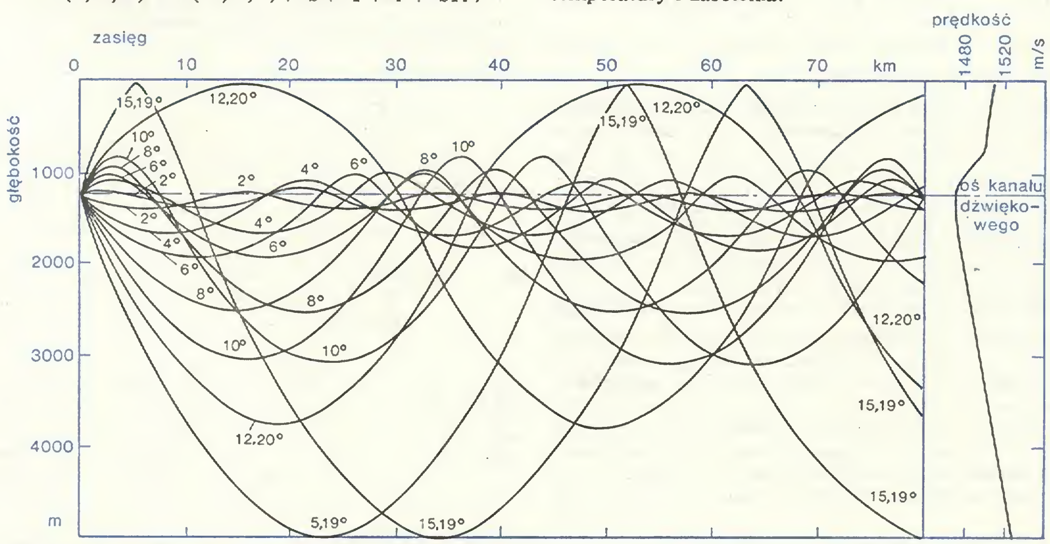
gdzie $c(35, 0, 0) = 1449{,}22$ m/s jest wartością prędkości przy zasoleniu 35‰, temperaturze 0°C i ciśnieniu atmosferycznym na poziomie morza, dalsze zaś składniki po prawej stronie wzoru stanowią poprawki, które trzeba uwzględnić przy zmianie każdej wielkości z osobna; ostatni wyraz określa poprawkę wynikającą z ich wzajemnego wpływu. Poszczególne poprawki w postaci wielomianów o różnych współczynnikach liczbowych podane są w literaturze fachowej i oblicza się je dla danych warunków.

Wzór Wilsona pozwala obliczyć prędkość dźwięku w morzu w zakresie temperatur od −4°C do 30°C, ciśnień — od 1 kG/cm² do 1000 kG/cm² i zasoleniu od 0 do 37‰, z dokładnością do +0,3 m/s. Dokładność ta jest tego samego rzędu co dokładność pomiarów prędkości w wodzie.

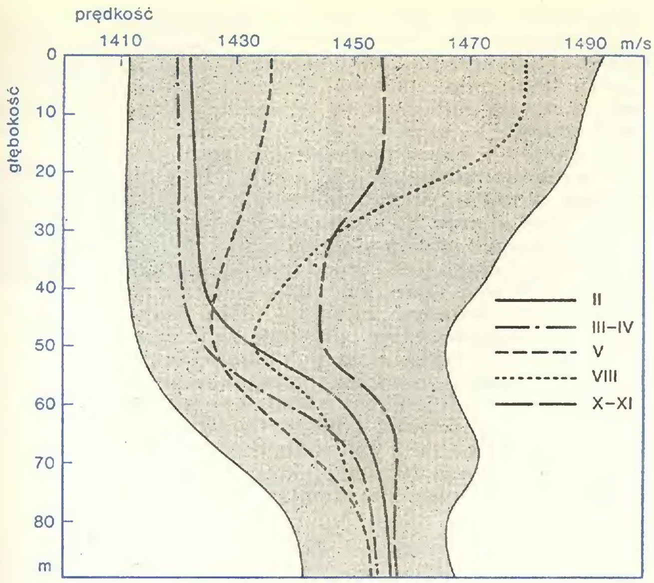
W miarę spadku temperatury wraz z głębokością zmniejsza się prędkość dźwięku — głównie wskutek wzrostu gęstości. W warstwie przypowierzchniowej obniżenie to wynosi ok. 3 m/s na 1°C. Wzrostowi zasolenia wraz z głębokością towarzyszy wzrost prędkości mniej więcej o 1,3 m/s na 1‰ — wynika to z faktu, że wzrost gęstości przy wzroście zasolenia jest powolniejszy niż spadek ściśliwości (zob. wzór 8). Wzrost ciśnienia hydrostatycznego wraz z głębokością daje wzrost prędkości dźwięku mniej więcej o 1,8 m/s na każde 100 m słupa wody.

Rozkłady prędkości $c(z)$ są więc bardzo zróżnicowane w przestrzeni i zmieniają się także w czasie (zmiany dobowe i sezonowe). Kilka najbardziej charakterystycznych rozkładów przedstawia rys. 33. Niektóre z nich wykazują wyraźne minimum prędkości dźwięku na określonej głębokości. W obrębie warstw wyznaczonych przez takie minima funkcji $c(z)$ wytwarzają się kanały dźwiękowe. Są to obszary, w których wskutek refrakcji fale dźwiękowe biegną jak w falowodzie, oscylując pomiędzy dwoma granicznymi powierzchniami, na których zachodzi warunek (7) całkowitego wewnętrznego odbicia. Przykład schematycznego biegu promieni dźwiękowych w charakterystycznych wypadkach — w kanale dźwiękowym przypowierzchniowym (na głębokości 40–70 m) oraz w kanale głębinowym (1000–1500 m) — pokazuje rys. 34. Na rys. 35 przedstawione są wyniki pomiarów wykonanych w oceanie na dużej głębokości.

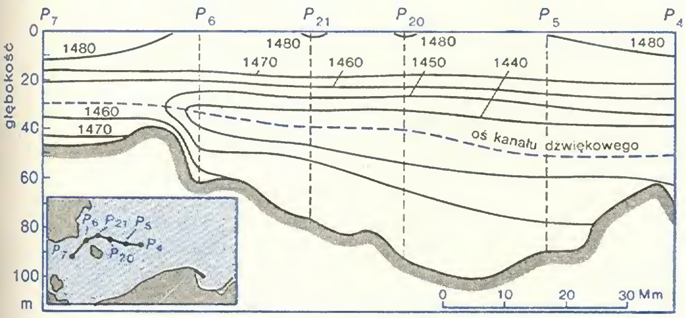
Szczególnego rodzaju falowód dla dźwięku może stanowić akwen morza płytkiego, gdzie rozchodzące się fale dźwiękowe ulegają kolejno odbiciom od dna i od powierzchni morza. Ilustrację wyników badań kanałów dźwiękowych w Bałtyku przedstawiają rys. 36 i 37. Na rys. 37 jest uwidoczniona oś kanału dźwiękowego $c(z)$ = minimum, oszacowana z pomiarów temperatury i zasolenia.



Rys. 35. Zmierzone przebiegi sygnałów dźwiękowych w kanale głębinowym SOFAR wg M. Ewinga i in.



Rys. 36. Zmiany sezonowe pionowego rozkładu średnich prędkości dźwięku w Głębi Bornholmskiej, wyznaczone na podstawie dziesięcioletniej serii pomiarów temperatury i zasolenia wody. Krzywe umieszczono na szarym tle, którego granice stanowią pionowe rozkłady minimalnych i maksymalnych prędkości dźwięku. Cyfry rzymskie oznaczają miesiące (wg P. Tymańskiego)



Rys. 37. Rozkład poziomy średnich prędkości dźwięku w morzu w (m/s) wg danych z sierpnia oraz profil dna morskiego wzdłuż przekroju pionowego Głębia Gdańska-Rynna Słupska (P_7-P_4), wyznaczony na podstawie wyników serii pomiarów temperatury i zasolenia wody, trwającej dziesięć lat (wg P. Tymańskiego)

Rozpraszanie dźwięku

Zjawisko rozpraszania dźwięku występuje zarówno przy jego odbiciu od pofalowanej powierzchni morza oraz nierównego dna, jak i przy odbiciu od niejednorodności wewnętrznych, takich jak pęcherzyki gazów, fluktuacje termiczne ośrodka, elementy biologiczne i in. Rozpraszaniu sygnałów akustycznych w różnych kierunkach towarzyszą fluktuacje amplitudy i fazy tych sygnałów, utrudniające w praktyce interpretację przy ich rejestracji. Ze względów teoretycznych i praktycznych zjawisku temu poświęca się w fizyce morza wiele uwagi. Zagadnienia związane z poszukiwaniem odpowiedniego modelu rozpraszania dźwięku przez nierówne powierzchnie można podzielić na dwie grupy. W pierwszej grupie wychodzi się z założenia, że charakter powierzchni rozpraszających jest zdeterminowany (ma np. kształt o przekroju sinusoidalnym lub piłowatym), natomiast w drugiej grupie przyjmuje się statystyczny rozkład nierówności elementów rozpraszających i również pole akustyczne opisuje się statystycznie. Wiele prac eksperymentalnych przeprowadza się na modelach w basenach przeznaczonych specjalnie do badań akustycznych.

Wskutek rozpraszania dźwięku w morzu część energii akustycznej rozchodzi się we wszystkich kierunkach — także pod kątem 180° (wstecz) — i powraca do źródła dźwięku. Odbiornik umieszczony w takim polu akustycznym rejestruje nie tylko falę, która do niego dochodzi bezpośrednio od źródła, ale również szereg fal rozproszonych przychodzących z różnych stron. Jeśli w pewnej chwili źródło dźwięku zostanie nagle wyłączone, odbiornik po zaniknięciu fali bezpośredniej będzie jeszcze przez pewien czas rejestrować fale rozproszone. Jest to tzw. zjawisko pogłosu lub rewerberacji w morzu. Zniekształca ono przesyłane w morzu sygnały dźwiękowe, powoduje „rozciąganie" ich w czasie i maskowanie sygnału dźwięku pożądanego przez powstające sygnały pasożytnicze (rozproszone).

zjawisko pogłosu

Tłumienie dźwięku

Wskutek absorpcji przez ośrodek dźwięk w morzu jest tłumiony, czyli zmniejsza się jego amplituda w miarę rozchodzenia się sygnału. Wpływają na to głównie dwa czynniki. Jeden określa się jako tzw. tłumienie klasyczne, w którym decydującą rolę odgrywa lepkość środowiska, mniejszą — przewodnictwo cieplne; drugi wiąże się z tzw. relaksacją strukturalną (→ Przedmiot i zakres akustyki, rozdział: Akustyka molekularna), a tu główną rolę odgrywa oddziaływanie fali z drobinami rozpuszczonego w wodzie morskiej siarczanu magnezu $MgSO_4$. W praktyce do określenia współczynnika tłumienia amplitudy fali dźwiękowej w wodzie morskiej używa się zwykle empirycznego wzoru, ustalonego przez W. W. Schulkina i Marsha, który uwzględnia zarówno istniejące teorie absorpcji dźwięku w cieczach, jak i szereg wyników doświadczalnych uzyskanych w morzu przez wielu autorów. W szczególności wzór ten uwzględnia zależność od głębokości (poprzez ciśnienie hydrostatyczne) oraz od temperatury:

wzór Schulkina–Marsha

$$\alpha = 8{,}686\cdot 10^{-6}\nu^2\left(\frac{2{,}34S\cdot\nu_T}{\nu_T^2+\nu^2}+\frac{3{,}38}{\nu_T}\right)\times$$
$$\times(1-6{,}54\cdot 10^{-4}p), \quad (9)$$

gdzie $\nu_T = 21{,}9\cdot 10^{6-1520/(T+273)}$, S — zasolenie w ‰, T — temperatura bezwzględna, ν — częstość fali w kHz, p — ciśnienie hydrostatyczne w atm.

W przestrzeni otwartej (w praktyce — w morzu głębokim) natężenie fali dźwiękowej wytworzonej przez źródło bezkierunkowe, mierzone w pewnej odległości od tego źródła, wyraża się wzorem

$$I = I_0\frac{e^{-2\alpha R}}{R^2},$$

gdzie I — natężenie w odległości R od źródła, I_0 — natężenie odniesienia w odległości $R_0 = 1$ m.

Natężenie dźwięku maleje odwrotnie proporcjonalnie do kwadratu odległości od źródła oraz eksponencjalnie wskutek absorpcji w toni wodnej. Współczynnik absorpcji zależy od częstości fali dźwiękowej, od zasolenia, temperatury i ciśnienia hydrostatycznego. Poziom natężenia dźwięku (w dB) określony jest jako:

$$L = 10\ \lg\frac{I}{I_{00}}.$$

Uwzględniając, że natężenie dźwięku jest proporcjonalne do kwadratu amplitudy, otrzymujemy

$$L = 20\ \lg\frac{A}{A_{00}},$$

gdzie A i A_{00} — amplitudy fali akustycznej odpowia-

dające natężeniom I i I_{00} odpowiednio (I_{00} — poziom odniesienia, np. 1 W/m²). Zatem

$$L = L_0 - (20\lg e)\alpha R - 20\lg R,$$

wyznaczanie współczynnika tłumienia

gdzie e — podstawa logarytmów naturalnych. Na podstawie tego wzoru wyznacza się współczynnik tłumienia dźwięku mierząc poziom natężenia fali w dwóch odległościach R_0 i R od źródła. Współczynnik tłumienia α wyrażany jest w dB/m lub dB/km — w zależności od tego, czy R mierzymy w metrach czy kilometrach. Ostatni składnik $20\lg R$ uważać można za regularną poprawkę, wynikającą z kulistego rozchodzenia się fali, a liniowa zależność poziomu natężenia L od R pozwala z nachylenia prostej znaleźć

$$\bar{\alpha} = (20\lg e)\alpha = 8{,}686\alpha,$$

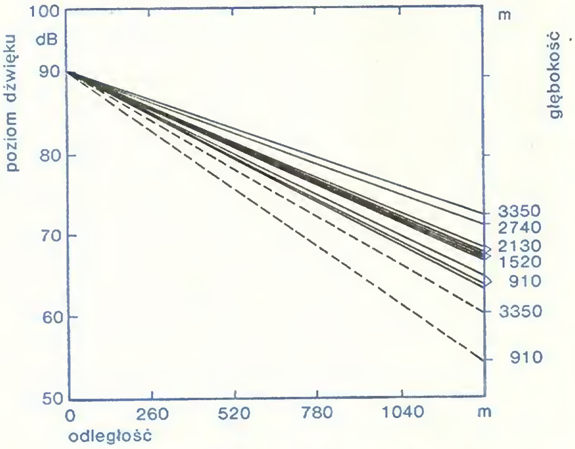
gdzie $\bar{\alpha}$ — współczynnik tłumienia w dB/km.

Pomiary takie były w oceanie wielokrotnie przeprowadzane i wykazały liniową zależność od głębokości (od ciśnienia hydrostatycznego):

$$\bar{\alpha} = \bar{\alpha}_0 - ap,$$

gdzie $\bar{\alpha}_0$ — współczynnik tłumienia w dB/km dla ciśnienia odniesienia (na poziomie morza), a — współczynnik ciśnieniowy tłumienia w dB/(km·at).

Na rys. 38 są porównane wyniki pomiarów wykonane w morzu z wynikami uzyskanymi z obliczeń na podstawie wzoru (6).



Rys. 38. Zależność natężenia fali ultradźwiękowej o częstości 75 kHz od odległości od źródła. Linie ciągłe — zależności wyznaczone w Oceanie Spokojnym przy różnych głębokościach — linie przerywane — zależności przewidywane przez wzór Schulkina–Marsha (wg Bezdeka)

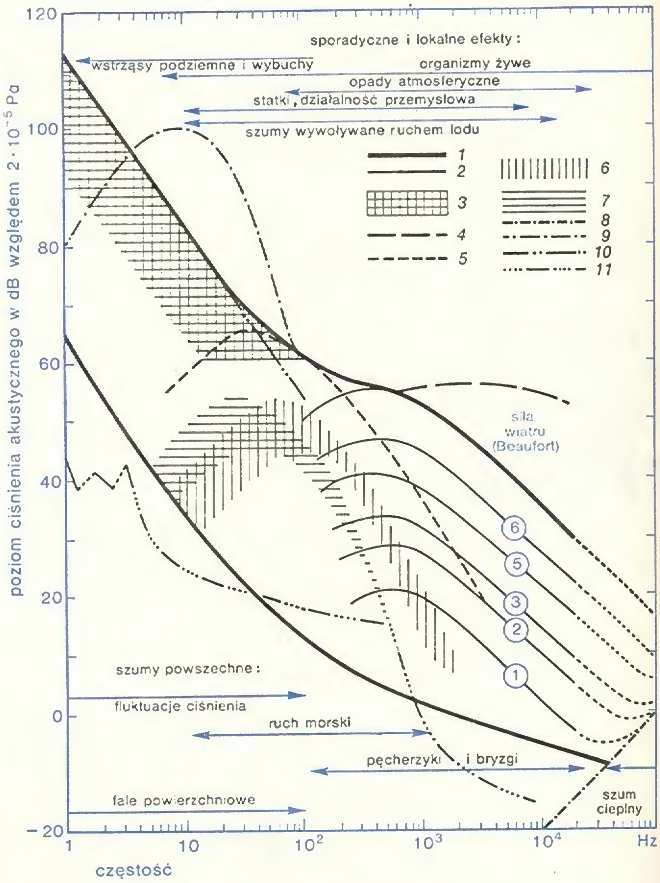
dyspersja prędkości

Zjawisko dyspersji prędkości dźwięku, czyli zależność prędkości od częstości fali, pozostaje w związku głównie z występowaniem wspomnianej wyżej relaksacyjnej części absorpcji dźwięku. Ta dyspersja wykrywana jest w morzu w zakresie częstości 100–150 kHz. Dyspersyjny wpływ, szczególnie w morzach płytkich, mają także procesy tłumienia i rozpraszania dźwięku przez dno i sfalowaną powierzchnię morza. Dyspersyjne własności środowiska morskiego wyraźnie mogą wpływać na przesyłane sygnały dźwiękowe, gdy te mają szerokie widmo częstości, co ma przede wszystkim miejsce przy krótkich w czasie przebiegach impulsowych.

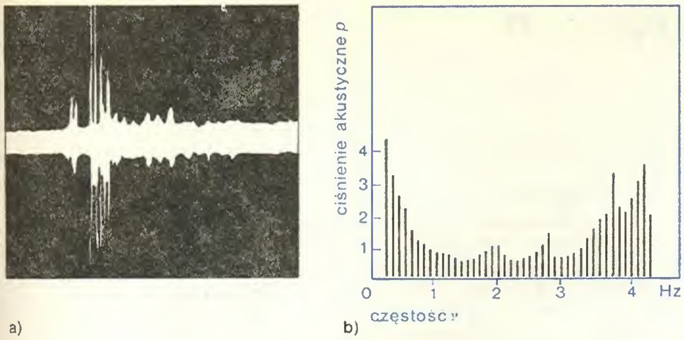
Szumy własne morza

Szumy własne morza, które ze względu na pochodzenie dzielimy, jak już wspomniano, na naturalne i sztuczne, stanowią nie mniej bogaty świat dźwięków niż otaczający nas bardziej bezpośrednio poznawalny świat dźwięków w atmosferze. Jako przykłady źródeł naturalnych szumu morza należy wymienić: falowanie powierzchni morza, tarcie wiatru o powierzchnię, opady atmosferyczne, fluktuacje termiczne, burzliwe przepływy mas wodnych oraz aktywność różnych gatunków fauny morskiej (krewetki, ryby, delfiny i wiele innych). Szumy pochodzące od technicznej działalności człowieka — to głównie hałasy związane z ruchem statków i innymi operacjami na morzu (np. poszukiwaniami geologicznymi), a w strefie przybrzeżnej — również z działalnością na lądzie, gdyż hałasy łatwo się przez ląd przenoszą do wody.

Pole akustyczne szumów własnych panujące w toni morskiej stanowi ciągle — mimo intensywnych badań — temat mało znany, gdyż jest ono zależne od bardzo wielu zmiennych czynników. Jako pewną ogólną charakterystykę tego pola często się w literaturze przytacza widmo szumów własnych morza (rys. 39), w którym się rozróżnia szereg składowych pochodzących z rozmaitych źródeł. Rysunek 40 przedstawia jako przykład dźwięków biologicznych w morzu zapis sygnału (i jego widmo) wydawanego przez rybę. Ilustracja 26 (tabl. 8) ukazuje hydrofon rejestrujący szumy podwodne w toni morskiej.



Rys. 39. Składowe akustycznego pola szumów własnych w morzu. Wykresy przedstawiają uśrednione wyniki wielu badań określające poziomy ciśnienia akustycznego i obszary widmowe (niebieskie strzałki poziome) przyporządkowane prawdopodobnym źródłom szumów w zakresie od 1 Hz do 100 Hz (wg Z. Kluska) *1* minimalny i maksymalny poziom szumu własnego; *2* szumy zależne od prędkości wiatru, niebieskie okółkowane cyfry — siła wiatru w skali Beauforta, których źródłem mogą być pęcherzyki gazów i bryzgi; linie przerywane określają ekstrapolację danych pomiarowych do poziomu szumów cieplnych; *3* tzw. pseudodźwięki — pulsacje ciśnienia, których prędkość fazowa rozchodzenia się jest mniejsza od prędkości dźwięku w danym środowisku w zakresie małych częstości; *4* szumy rejestrowane przy obfitych opadach atmosferycznych; *5* szumy powstające przy ożywionym ruchu statków; *6* hałasy statków na wodach płytkich; *7* hałasy statków na wodach głębokich; *8* szumy cieplne; *9* wybuchy i wstrząsy sejsmiczne; *10* krzywa ekstrapolowana składowej zależnej od wiatru; *11* szumy rejestrowane w akwenie całkowicie pokrytym lodami



Rys. 40. Dźwięk wydawany przez rybę (*Opicephalus argus*) w chwili schwytania zdobyczy zarejestrowany w basenie rzeki Amur: a) zapis dźwięku na oscylografie i b) jego widmo (wg W. R. Protasowa i E. W. Romanienko)

Zastosowanie

Całokształt omówionych wyżej zjawisk akustycznych w morzu warunkuje możliwości wykorzystania fal akustycznych. Są one szerokie i obejmują zastosowania: komunikacja podwodna, zdalne pomiary wielkości fizycznych charakteryzujących hydrosferę (temperatura, ciśnienie hydrostatyczne, zasolenie, prędkość prądów mas wodnych itp.), echosondaż (pomiary głębokości, analiza dna i poddennych warstw geologicznych, wykrywanie zasobów rybnych, sterowanie i obserwacja procesów połowów ryb), sonar (przeszukiwanie toni wodnej w określonym sektorze azymutalnym i wykrywanie obiektów podwodnych, np. przeszkód nawigacyjnych, ławic ryb itp.), nawigacja (np. log dopplerowski do określania szybkości względnej ruchomych obiektów).

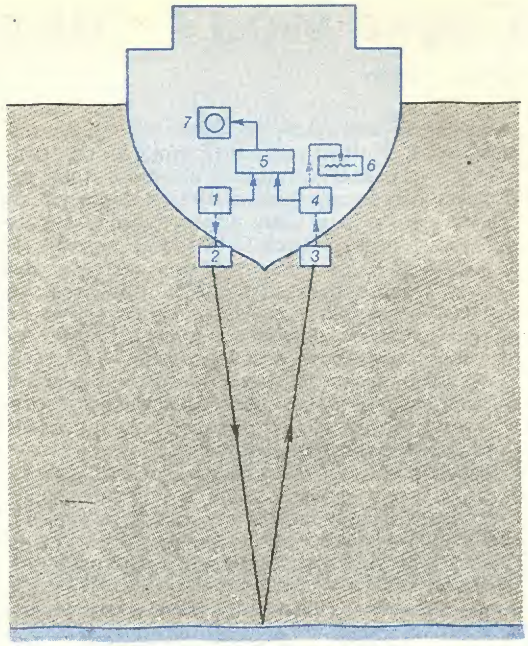
hydrolokacja w nawigacji

Urządzenia ultradźwiękowe służące do komunikacji podwodnej i hydrolokacji, przy korzystaniu z których ważny jest zarówno zasięg, jak i kolimacja wiązki fal, pracują zwykle w zakresie częstości optymalnych, rzędu kilkudziesięciu kHz. Kiedy chodzi o uzyskanie dużego zasięgu, stosuje się sygnały o częstościach mniejszych (z uwagi na mniejsze tłumienie). W niektórych zagadnieniach specjalnych, jak badanie przepływów czy wykrywanie obiektów małych na podstawie zjawiska Dopplera, stosuje się także ultradźwięki o większych częstościach — kilkuset kHz, a nawet MHz.

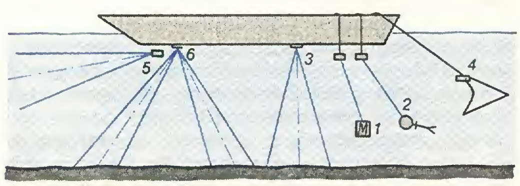
Po raz pierwszy zastosowano ultradźwięki do celów podwodnych podczas I wojny światowej, kiedy to P. Langevin, fizyk francuski, zbudował urządzenie echolokacyjne — echosondę. Echosonda wysyła sygnał (krótki impuls), a następnie odbiera go po odbiciu od przeszkody; długość odstępu czasu między wysłaniem i odebraniem sygnału i zmiana jego natężenia daje informacje o przeszkodzie, zwłaszcza o odległości, w jakiej się znajduje (rys. 41).

SONAR

Układy echolokacyjne oparte na wykorzystaniu fal ultradźwiękowych w morzu określa się obecnie nazwą SONAR (*Sonic Navigation and Ranging*). Współczesne statki wyposażone są w szereg urządzeń ultradźwiękowych do badania i penetrowania hydrosfery; najważniejsze z nich przedstawia poglądowo rys. 42. W zależności od typu statku, jego wielkości i przeznaczenia liczba tych urządzeń jest różna. Na statkach dużych instaluje się często po kilka echosond, o różnych zasięgach i różnych szerokościach wiązki. Urządzenia hydrolokacyjne na nowoczesnych statkach, szczególnie okrętach wojennych, statkach rybackich i pomiarowo-badawczych (np. m/s „Profesor Siedlecki"), współdziałają z komputerami, które w bardzo szybki i dokładny sposób przetwarzają rejestrowane sygnały akustyczne na pożądane informacje.



Rys. 41. Schemat i zasada działania echosondy ultradźwiękowej na statku: *1* generator impulsów elektrycznych, *2* przetwornik nadawczy impulsów ultradźwiękowych, *3* przetwornik odbiorczy impulsów ultradźwiękowych, *4* wzmacniacz, *5* układ synchronizujący, *6* rejestrator piszący, *7* rejestrator oscyloskopowy



Rys. 42. Urządzenia hydroakustyczne na statku morskim: *1* urządzenie do zdalnego pomiaru wielkości fizycznych — dane rejestrowane przez czujnik pomiarowy *M* przekazywane są na statek badawczy (lub na ląd) przy pomocy łącza ultradźwiękowego; *2* hydrotelefon — urządzenie do dwustronnej komunikacji telefonicznej z nurkiem swobodnym lub innym obiektem zanurzonym; *3* echosonda ultradźwiękowa do pomiaru głębokości morza, badań struktury dna oraz wykrywania ławic ryb; *4* echosonda sieciowa — do określania położenia włoka (zwłaszcza pelagicznego), jego rozwarcia, naprowadzenia na ławicę i obserwacji ryb wpadających do włoka; *5* sonar do przeszukiwania toni wodnej w określonym sektorze azymutalnym (zwykle $\pm 120°$) celem wykrycia i lokalizacji obiektów podwodnych, np. gór lodowych, wraków, ławic ryb itp.; *6* log dopplerowski — do pomiaru prędkości rzeczywistej statku. Jest to jedyny przyrząd umożliwiający pomiar prędkości statku względem dna morza bez błędów powstających wskutek prądów, wpływu wiatru itp. czynników (wg Z. Jagodzińskiego)

R. D. Bobber *Hydroakusticzeskije izmierienija*, Moskwa 1974; L. M. Brekhowskikh (ed.) *Akustika okieana*, Moskwa 1974; J. Dera i in. *Wybrane zagadnienia fizyki morza. Część V. Rozpraszanie dźwięku na sfalowanej powierzchni morza i dnie morskim*, Post. Fiz. 25, 175 (1974); J. Dera i in. *Wybrane zagadnienia fizyki morza. Część IV. Fale dźwiękowe w morzu*, Post. Fiz. 23, 667 (1972); *Fizyka morza, cz. II — Akustyka morza*, Sopot 1974; Z. Kowalik i in. *Podstawy hydroakustyki*, Gdynia 1965; A. B. Staszkiewicz *Akustika moria*, Leningrad 1966; R. W. B. Stephens *Underwater Acoustics*, London 1970; Studia i materiały oceanologiczne 7, 109 (1973); W. W. Szulkin *Fizyka moria*, Moskwa 1968.

Fizyka wód śródlądowych

Bohdan Utrysko

wody śródlądowe

Przez pojęcie wód śródlądowych rozumie się w technice cieki i zbiorniki naturalne i sztuczne, tzn. strumienie, rzeki, kanały, jeziora i zbiorniki zaporowe. W fizyce rozszerzamy pojęcie wód śródlądowych na całą tę część obiegu wody w przyrodzie, która się odbywa na lądzie. Woda śródlądowa pojawia się w postaci opadu lub kondensacji pary wodnej na ciałach stałych, a opuszcza ląd parując lub odpływając do morza. Woda opadowa częściowo paruje, spływa po powierzchni ziemi lub wsiąka w grunt, by zasilić potem znów wody powierzchniowe albo wyparować. Wyodrębnić więc można dwa środowiska występowania wód lądowych: w korytach i zbiornikach otwartych (wody powierzchniowe) oraz w porach gruntów i szczelinach skał (wody podziemne).

Przepływ wody powierzchniowej podlega prawom badanym przez dynamikę cieczy (hydrodynamikę, hydraulikę), zmiany cieplne obejmuje termika wód, efekty oddziaływania płynącej wody na jej koryto — dynamika koryt. Całość zagadnień przepływów podziemnych obejmuje nauka zwana najczęściej teorią filtracji. W każdej z tych nauk istnieje szereg problemów, które naukowcy starają się rozwiązać w aktualnie prowadzonych badaniach. Aby je zrozumieć, konieczne jest wprowadzenie w najogólniejsze pojęcia i prawa rządzące w wymienionych dziedzinach.

Mechanika przepływów wód powierzchniowych

Wody powierzchniowe pozostają w kontakcie z atmosferą całą swą powierzchnią, zwaną zwierciadłem albo swobodną powierzchnią wody. Na powierzchni tej panuje wszędzie jednakowe ciśnienie.

Woda znajdująca się w stanie spoczynku w nieruchomym zbiorniku poddana jest działaniu sił ciężkości i ciśnienia, równoważących się wzajemnie. Z warunku równowagi sił wynika wzór na ciśnienie hydrostatyczne:

$$p = p_0+\varrho g h,$$

w którym p_0 jest ciśnieniem na powierzchni, h — zagłębieniem punktu, w którym ciśnienie wynosi p, ϱ — gęstością wody.

W wodzie płynącej związki są bardziej złożone. W dowolnym punkcie obszaru wypełnionego cieczą i w dowolnej chwili t występuje pewna prędkość v, ciśnienie p, a woda posiada gęstość ϱ. Dla znalezienia wartości tych trzech niewiadomych potrzebne są trzy równania. Są to: równanie stanu, wyrażające zależność gęstości wody od jej ciśnienia i temperatury, równanie ciągłości, wyrażające warunek zachowania masy, oraz równanie ruchu, wyrażające zależność wszystkich sił działających na cząstkę cieczy.

charakterystyka strumienia cieczy

W wielu zagadnieniach praktycznych, szczególnie dotyczących przepływów wód powierzchniowych, w rozważaniach traktuje się cały strumień cieczy łącznie i bada się parametry opisujące globalnie sytuację w danym przekroju strumienia. Do podstawowych wielkości charakteryzujących strumień należą:

— objętościowe natężenie przepływu (krócej przepływ) Q, będące stosunkiem objętości wody dV, przepływającej przez pewien przekrój w czasie dt, do tego czasu.

— pole przekroju poprzecznego S. Jeśli chodzi o przepływ w przewodach zamkniętych, jest ono z góry znane, natomiast w korytach otwartych zależy od napełnienia koryta. Jeśli kształt koryta jest znany, pole jest jednoznacznie określone przez rzędną swobodnej powierzchni z. Niekiedy, zwłaszcza gdy przekrój koryta jest regularny, wygodniej jest operować głębokością średnią, określoną jako $h_{śr} = S/B$, gdzie B jest szerokością strumienia.

— energia potencjalna płynącej wody, charakteryzowana przez $h_p = z+p/(\varrho g)$, tzw. wysokość energii potencjalnej.

Gdy kierunek ruchu jest niewiele odchylony od poziomu, a tak jest przecież w rzekach i kanałach, ciśnienie wzdłuż linii pionowych zmienia się niemal tak jak w wypadku wody stojącej i h_p wszystkich cząstek w przekroju jest wtedy identyczne i równe:

$$z_{pow}+\frac{p_0}{\varrho g}.$$

Ponieważ p_0 jest jednakowe we wszystkich przekrojach, do określenia energii potencjalnej strumienia w podanych wyżej warunkach wystarczy znajomość rzędnych zwierciadła wody z_{pow}.

Wielkościami pochodnymi od wyżej podanych są występujące często we wzorach i obliczeniach:
prędkość średnia w przekroju $v_{śr} = Q/S$,
promień hydrauliczny przekroju $R_h = S/O_z$. Symbolem O_z oznaczono obwód zwilżony, czyli długość tej części obwodu przekroju S, na której woda styka się z ciałami stałymi. Przyjmuje się, że w korytach naturalnych, szerokich i płytkich ($B > 30h_{śr}$) $R_h = h_{śr}$.

przepływ ustalony i nieustalony

Najogólniejszym rodzajem ruchu wody jest przepływ nieustalony (zmienny w czasie). W przepływie ustalonym parametry ruchu nie zależą od czasu, a natężenie przepływu, przynajmniej między dopływami lub rozgałęzieniami kanału, pozostaje stałe. Jeżeli kształt koryta jest niezmienny, jego kierunek prostoliniowy i głębokość wszędzie jednakowa, to mamy do czynienia z ruchem jednostajnym. Jest to najprostszy rodzaj ruchu.

W dynamice przepływów wód powierzchniowych zmiany gęstości, wywołane zmianami ciśnienia i temperatury, mają najczęściej znikome znaczenie i do rozważań można przyjąć równanie stanu w postaci:

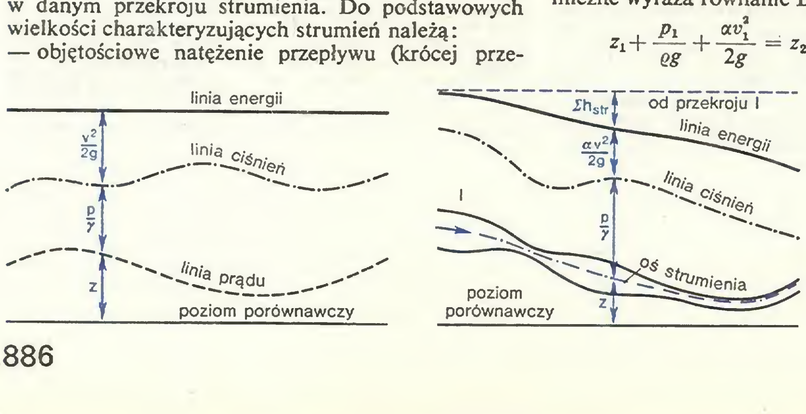
$$\varrho = \text{const}.$$

Dla koryt otwartych, przy rozpatrywaniu całościowym przekrojów i założeniu $\varrho = \text{const}$, równanie ciągłości ma postać $\partial Q/\partial l+\partial S/\partial t = 0$. W ruchu ustalonym (niezależnym od t) równanie to upraszcza się do warunku $Q = \text{const}$.

W ruchu jednostajnym oraz w zbliżonym do niego ustalonym ruchu wolnozmiennym zależności dynamiczne wyraża równanie Bernoulliego (rys. 1):

równanie Bernoulliego

$$z_1+\frac{p_1}{\varrho g}+\frac{\alpha v_1^2}{2g} = z_2+\frac{p_2}{\varrho g}+\frac{\alpha v_2^2}{2g}+\Delta h_e.$$



Rys. 1. Interpretacja równania Bernoulliego: a) ciecz doskonała, b) ciecz rzeczywista (w równaniu należy wówczas uwzględnić straty energii)

współczynnik Coriolisa

Wielkości z indeksami 1 i 2 dotyczą dwóch kolejnych przekrojów strumienia. Współczynnik

$$\alpha = \frac{1}{v_{\acute{s}r}^3} \int_S v^3 dS,$$

zwany współczynnikiem Coriolisa lub Saint-Venanta, pozwala na uwzględnienie nierównomiernego rozkładu energii kinetycznej w strumieniu. Każdy z wyrazów równania Bernoulliego ma wymiar liniowy i wyraża wysokość odpowiedniego rodzaju energii (potencjalnej i kinetycznej), tzn. stosunek wielkości energii do ciężaru cieczy. Ostatni symbol Δh_c przedstawia różnicę wysokości łącznej energii mechanicznej w obu przekrojach. Ten ubytek energii pochodzi stąd, że płynąca ciecz musi pokonać pewne opory, które są analogiczne do sił tarcia ale których wielkość zależy od prędkości ruchu.

Bezwymiarowa wielkość:

$$J = \frac{1}{mg} \frac{dE}{dl},$$

przedstawiająca straty energii mechanicznej E na jednostkę długości odniesione do ciężaru płynącej wody, nazywana jest spadkiem hydraulicznym.

Ruch jednostajny w korycie otwartym opisuje równanie Bernoulliego zredukowane do:

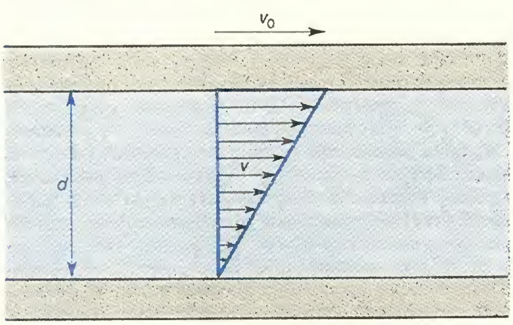
$$z_1 - z_2 = Jl.$$

Wielkość J jest więc wtedy równa spadkowi zwierciadła wody.

Określenie ilości energii zużytej na pokonanie oporów ruchu i wyjaśnienie mechanizmu powstawania tych strat stanowi jedno z najważniejszych zadań dynamiki przepływów. Jeśli prędkość jest bardzo mała, cząsteczki wody płyną wzdłuż regularnych torów, tworząc warstewki posuwające się jedne po drugich. Ruch taki nazwano uwarstwionym, czyli laminarnym.

ruch laminarny

Wyobraźmy sobie dwie bardzo duże płaskie płyty (rys. 2), między którymi znajduje się ciecz. Górna płyta przesuwa się względem dolnej równolegle i ze

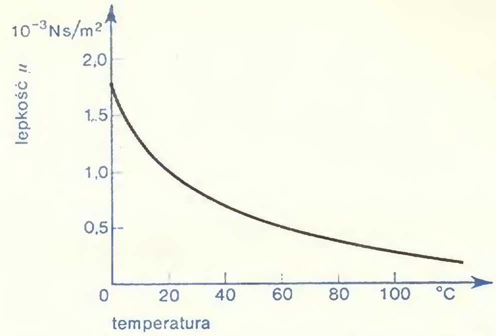


Rys. 2. Ruch cieczy lepkiej między płytami ruchomą i nieruchomą

stałą prędkością v_0. Cząsteczki wody przylegające do obu płyt mają prędkości takie jak płyty, tzn. w pobliżu płyty górnej mają prędkość v_0. Natomiast cząsteczki znajdujące się dalej płyną z coraz mniejszą prędkością; płyną one warstewkami (stąd nazwa uwarstwiony, laminarny) jedna po drugiej. Newton podał wzór wiążący występującą między sąsiednimi warstewkami siłę tarcia wewnętrznego T z prędkością ruchu:

$$\frac{T}{S} = \mu \frac{v_0}{d}$$

(S jest powierzchnią, na której działa siła styczna T). Występujący we wzorze współczynnik proporcjonalności μ zwany jest dynamicznym współczynnikiem lepkości. Jest to wielkość charakteryzująca dany płyn, np. dynamiczny współczynnik lepkości wody w temperaturze 20°C ma wartość 0,001 Ns/m^2, oliwy zaś 0,084 Ns/m^2. Maleje on szybko ze wzrostem temperatury; zależność dynamicznego współczynnika lepkości wody od temperatury jest przedstawiona na rys. 3.



Rys. 3. Wykres zależności lepkości wody od temperatury

ruch turbulentny

Przy nieco większych prędkościach obraz ruchu bardzo się komplikuje. Cząsteczki wody, poruszając się wzdłuż torów chaotycznych i splątanych, tworzą mniejsze i większe zawirowania. Nawet w ruchu makroskopowo ustalonym chwilowa prędkość i ciśnienie w każdym punkcie podlegają bardzo szybkim i przypadkowym zmianom. Ruch taki nazywa się burzliwym lub turbulentnym. W ruchu burzliwym tarcie wewnętrzne jest o wiele większe niżby to wynikało z wzoru Newtona przy założeniu, że v_0 jest prędkością przeciętną, tzn. średnią w czasie w każdym punkcie. Co więcej, z tego wzoru wynika, że straty powinny być proporcjonalne do prędkości, natomiast doświadczenia wykazują, że w ruchu burzliwym dwukrotny wzrost prędkości średniej powoduje prawie czterokrotny wzrost strat. Znaczy to, że szukana zależność jest zbliżona do parabolicznej. We wszystkich rodzajach przepływu wód powierzchniowych interesujących ze względu na znaczenie praktyczne występuje rozwinięty ruch turbulentny.

W praktyce do obliczeń prędkości średniej najczęściej stosuje się wzór Chezy'ego:

wzór Chezy'ego

$$v_{\acute{s}r} = c\sqrt{R_h \cdot J}.$$

Współczynnik c może być wyznaczony z różnych wzorów empirycznych, np. ze wzoru Manninga:

wzór Manninga

$$c = \frac{1}{n} R_h^{1/6},$$

gdzie n jest współczynnikiem szorstkości. Niektóre wartości tego współczynnika dotyczące przepływów w korytach otwartych podano w tabeli.

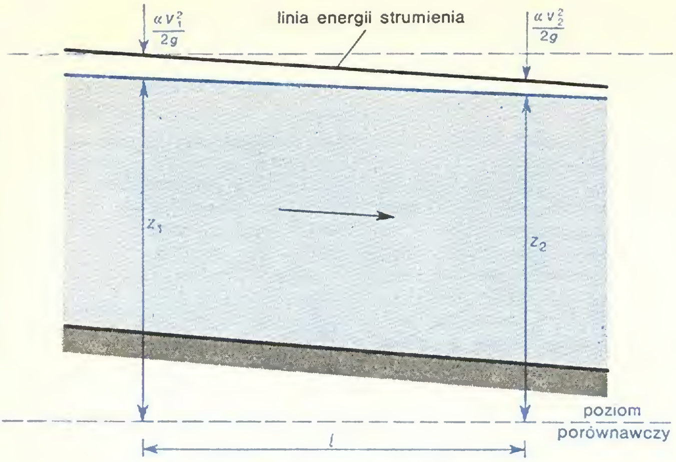
Wartości współczynnika szorstkości n do wzoru Manninga

Rodzaj koryta	n
Kanały betonowe	0,017
Kanały w zwartej ziemi	0,020
Małe kanały ziemne, rzeki o czystym i prostym korycie	0,025
Kanały w złym stanie, rzeki o dobrych warunkach przepływu	0,030
Rzeki o przeciętnych warunkach przepływu	0,035
Rzeki kręte, częściowo zarośnięte	0,040

W ruchu wolnozmiennym, tzn. takim, w którym zarówno przekroje S, jak i rozkłady prędkości wzdłuż strumienia zmieniają się płynnie i niezbyt szybko, równanie Bernoulliego dotyczące odcinka l ma postać:

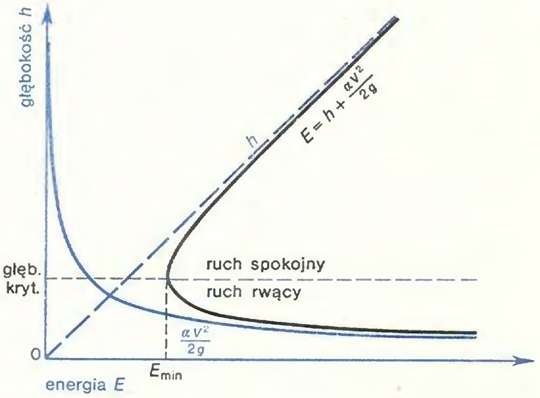
$$z_1 + \frac{\alpha v_1^2}{2g} = z_2 + \frac{\alpha v_2^2}{2g} + J_{\acute{s}r} \cdot l.$$

Z równania tego można wyznaczyć położenie zwierciadła w kolejnych przekrojach (rys. 4).



Rys. 4. Przepływ wolnozmienny w rzece — bilans energii

Bardzo istotną sprawą, zwłaszcza gdy chodzi o przepływy niejednostajne, jest określenie typu ruchu. Ta sama ilość wody może płynąć w danym korycie prędzej, mniejszym przekrojem, lub wolniej — korytem o większej powierzchni i głębokości. Zależnie od głębokości i prędkości zmienia się energia strumienia: gdy przekrój strumienia dąży do zera, prędkość i energia kinetyczna rośnie nieograniczenie; natomiast gdy głębokość i energia potencjalna dąży do nieskończoności, prędkość i energia kinetyczna maleje do zera (rys. 5). Pomiędzy tymi dwoma skrajnymi rodzajami przepływu mieszczą się przepływy rzeczy-



Rys. 5. Zależność energii strumienia od jego głębokości, przy stałym natężeniu przepływu

przepływ krytyczny

wiste, o ograniczonej energii, a wśród nich przepływ o energii minimalnej, nazywany krytycznym. Jest to taki przepływ, w którym przy danym Q energia całkowita (określona sumą wyrazów równania Bernoulliego) jest najmniejsza z możliwych. Parametry ruchu krytycznego są związane równaniem

$$\frac{S^3}{B} = \frac{Q^2}{g}.$$

Odpowiada to prędkości zwanej również krytyczną:

prędkość krytyczna

$$v_k = \sqrt{\frac{1}{\alpha} g \cdot h_{\acute{s}r}}\,.$$

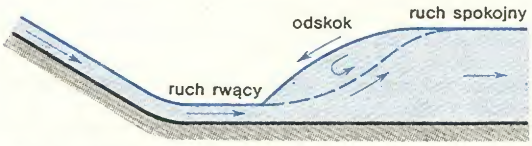
Prędkość ta jest niemal równa prędkości rozchodzenia się fal powierzchniowych. Ruch krytyczny jest istotny z dwóch względów: 1) Szereg zjawisk w przyrodzie zachodzi według zasady minimum energii, np. przypływy przez progi, przewężenia, jeżeli strumień przy tym ulega spiętrzeniu, odbywają się ruchem krytycznym. 2) Po przekroczeniu v_k żadne zaburzenia nie mogą się już przenosić pod prąd. Prędkość wody jest większa od prędkości fal biegnących po jej powierzchni w przeciwnym kierunku.

przepływy spokojne i rwące

Ze względu na tę ostatnią cechę dzielimy przepływy na dwa rodzaje: przepływy spokojne, $v < v_k$, w których zaburzenia przenoszą się w obie strony cieku (ruchem spokojnym płynie ogromna większość rzek, strumieni i potoków); przepływy rwące, $v > v_k$, w których górna część cieku jest izolowana od wpływu czynników występujących poniżej (ruch ten występuje w niektórych potokach górskich i u podnóża budowli piętrzących wodę).

odskok

Oprócz wymienionego już ustalonego ruchu wolnozmiennego spotyka się w naturze przepływy z gwałtownymi zmianami warunków ruchu. Jako ciekawy przykład takiego przepływu można podać przejście z ruchu rwącego do ruchu spokojnego. Strumień płynący ruchem rwącym o niewielkiej głębokości natrafia na wodę płynącą ruchem spokojnym, wolniejszym, o powierzchni wzniesionej dużo wyżej (rys. 6). Powstaje niepodparta ściana wody, z której spływa ona w kierunku pod prąd. Woda ta nie może jednak płynąć dalej, porywa ją bowiem i unosi z powrotem strumień rwący. Powstaje spłaszczony walec o poziomej osi, w którym krąży woda. Strumień przepływający pod



Rys. 6. „Odskok" — przejście strumienia z ruchu rwącego w ruch spokojny

spodem walca jest hamowany przez różnicę ciśnień do prędkości ruchu spokojnego — i dalej znów całym przekrojem płynie w kierunku zasadniczym. Zjawisko to nazywa się odskokiem i można je często obserwować, gdy woda spływa po stromej ścianie i trafia do zbiornika lub na płaski odcinek koryta.

W badaniach ruchów nieustalonych wyodrębnić należy przede wszystkim ruchy okresowe i nieokresowe. Ruchy okresowe, czyli falowanie, odgrywają w dynamice wód śródlądowych mniejszą rolę, natomiast stanowią istotny element dynamiki morza. Ruchy nieokresowe można podzielić na przepływy o zmianach powolnych i zmianach gwałtownych. Przykładem przepływu pierwszego rodzaju jest fala powodziowa; długość fali jest tysiące razy większa od jej wysokości, prędkość przesuwania się szczytu fali jest nieco większa niż prędkość przepływu, ale wznoszenie zwierciadła odbywa się bardzo powoli. Z gwałtownymi zmianami przepływu możemy mieć do czynienia np. w momencie przerwania zapory piętrzącej wodę. W dół rzeki biegnie wtedy potężna i stroma fala o wielkiej energii, ale dość szybko rozpływająca się, zwłaszcza gdy objętość zbiornika jest niewielka, a dolina — szeroka. Jednocześnie w samym zbiorniku powstaje gwałtowna fala obniżenia, idąca w górę cieku. Jest ona również niebezpieczna (chociaż w mniejszym stopniu niż fala przyboru), może bowiem spowodować obsunięcie zboczy zbiornika i zniszczyć jego brzegi.

badania współczesne

Prowadzone współcześnie badania w dziedzinie przepływów wód powierzchniowych skupiają się na następujących zagadnieniach:

a) zrozumieniu zjawisk zachodzących wewnątrz strumienia cieczy, przede wszystkim turbulencji i związanego z nią zjawiska dyfuzji;

b) badaniu zjawisk wzajemnego oddziaływania dwóch strumieni, np. wody słodkiej i słonej (ujścia rzek), ciepłej i zimnej (zrzuty wody chłodzącej), czystej i zanieczyszczonej:

c) bardzo istotnych i ciekawych doświadczeniach na modelach kanałów, rzeki budowli wodnych. Zja-

wiska zachodzące w strumieniu wody płynącym korytem o skomplikowanym kształcie są tak złożone, że nie znane są jeszcze metody opisu matematycznego tego ruchu ani sposoby rozwiązania. Jeżeli się chce wiedzieć, co się będzie działo po zbudowaniu zapory czy uregulowaniu rzeki, uciec się trzeba do doświadczeń. Oczywiście nie może to być doświadczenie w warunkach naturalnych, bo nie do pomyślenia jest wznoszenie na próbę kosztownych budowli. Odtwarza się je więc w laboratorium w zmniejszonej skali i na takim modelu wykonuje się doświadczenia. Zasady odtwarzania zjawisk dynamicznie podobnych, konstruowania modeli i wykonywania na nich badań stanowią dziś już odrębny dział mechaniki cieczy. Coraz częściej badania modelowe wspomaga się aparaturą elektronową i maszynami liczącymi, które rejestrują i opracowują wyniki pomiarów, a ponadto służą do wykonywania — jednocześnie z doświadczeniami fizycznymi — obliczeń tego samego zadania.

d) technice cyfrowej, która pozwoliła również na przeprowadzenie bardzo skomplikowanych obliczeń i rozwiązywanie metodami przybliżonymi złożonych równań (dotyczących np. ruchu nieustalonego).

Dynamika koryt

Woda płynąca korytem rzeki oddziałuje na nie powodując erozję (rozmycie), transport cząstek stałych i ich sedymentację (osadzanie). Energia przepływu zużywana jest nie tylko na pokonanie oporów ruchu, ale również na odspojenie materiału, jego transport i obróbkę (rozdrobnienie i wygładzenie). Wszystko to powoduje, że spadek hydrauliczny w korycie niosącym rumowisko jest większy niż w odpowiadającym mu korycie o dnie stałym. Istnieje zależność między ilością i jakością (wielkością i gęstością) materiału wleczonego i unoszonego a parametrami samego przepływu, a więc głębokością wody h i spadkiem hydraulicznym J. Każdy strumień posiada pewną zdolność transportową, którą stara się wykorzystać. Jeśli więc na jakimś odcinku cieku ilość materiału dostarczana przez przekrój początkowy jest mniejsza niż zdolność transportowa w dolnej części odcinka, następuje erozja i wynoszenie dodatkowych ilości materiału. Prowadzi to do silnych rozmyć, głównie brzegów, rzeka zaczyna meandrować, jej bieg się wydłuża, maleje spadek, a z nim i zdolność transportowa. Jeżeli brzegi są zabezpieczone przed rozmyciem, następuje silne pogłębienie koryta. Zjawisko to występuje wyraźnie np. po wybudowaniu zapory i stworzeniu zbiornika. Spadki powyżej zapory są bardzo małe i całe rumowisko osadza się w zbiorniku lub przy wlocie do niego. Następuje przy tym rozsortowanie materiału. Im bliżej zapory, tym mniejsza prędkość wody i tym drobniejszy materiał osiada na dnie. Przez zaporę przepływa już tylko czysta woda i strumień stara się jak najszybciej „nasycić" materiałem rozmywając i pogłębiając koryto poniżej budowli.

zdolność transportowa strumienia

Ilości transportowanego materiału mogą być bardzo duże (np. Wisła środkowa przenosi w przybliżeniu masę 3 mln ton rocznie). Sprawy określenia bilansu rumowiska, jego ilość i jakość są więc istotne, ale i bardzo skomplikowane. Powodem tego jest nie tylko niedostateczna jeszcze znajomość mechanizmu unoszenia i wleczenia, duża różnorodność kształtu i wielkości ziaren oraz form dna, ale i ciągła zmiana warunków przepływu: prędkości, głębokości, a nawet spadku. Otrzymane z doświadczeń liczne wzory określające ilość rumowiska mają wciąż jeszcze charakter empiryczny i dość wąskie zastosowanie.

Graniczna prędkość przepływu, poniżej której ziarna pozostają nieruchome, zależy od wielkości ziaren, z których zbudowane jest dno (od ich średnicy d) i wyznacza się ją doświadczalnie. Zależnie od prędkości przepływu materiał dna tworzy różne formy. Jeśli prędkość jest mała, ale przekracza prędkość graniczną, na płaskim dnie powstają drobne zmarszczki. Przy wzroście prędkości pojawiają się większe fałdy (diuny), pokryte często zmarszczkami. Ziarna wtaczają się po łagodnym stoku fałdy i przesypują się przez jej krawędź. Przy dalszym wzroście prędkości dno znowu staje się płaskie i wszystkie cząstki biorą jednocześnie udział w ruchu. Wreszcie przy bardzo dużej prędkości wystąpić mogą znowu fałdy (antydiuny). Stroną o łagodniejszym stoku są one zwrócone w kierunku przepływu i przesuwają się (fałdy, a nie tworzące je ziarna!) pod prąd. Zależnie od istniejących form dennych zmienia się szorstkość dna i opór stawiany przepływowi wody.

diuny i antydiuny

Mechanika przepływów wód podziemnych

Grunty sypkie składają się z ziaren różnej wielkości i kształtu. Między ziarnami, w porach gruntu, może się znajdować i przepływać woda. Jeżeli woda wypełnia wszystkie pory, mówimy, że grunt jest całkowicie nasycony wodą, jeżeli nie — nasycenie jest niepełne. Kanaliki, którymi przepływa woda w gruncie, mają bardzo skomplikowane i różnorodne formy. Nie jest więc możliwe ani konieczne określenie rzeczywistej prędkości cząstek wody. W rozważaniach zastępuje się prędkość rzeczywistą fikcyjną prędkością filtracji. Jest to iloraz przepływu i całkowitego pola przekroju warstwy, tzn. łącznie pola kanalików i ziaren: $v_f = Q/S_w$. Oczywiście prędkość filtracji jest mniejsza od prędkości rzeczywistej. Z wyjątkiem wypadków przepływu wody przez złoża rumoszu i bardzo grubego żwiru, gdzie się może pojawić turbulencja, przepływ w gruntach odbywa się ruchem laminarnym. Opory ruchu wyrażają się wtedy wzorem Darcy'ego:

prędkość filtracji

$$v_f = kJ = -k\frac{dh}{dl};$$

Współczynnik filtracji k zależy od rodzaju gruntu, tzn. od wielkości ziaren, ilości porów i temperatury wody (lepkości); współczynnik k żwirów wynosi ok. $3\cdot10^{-2}$ m/s, piasków średnioziarnistych 10^{-4} m/s, glin 10^{-8}m/s.

Ponieważ $h_p = z+p/(\varrho g)$ oznacza wysokość energii potencjalnej wody, a równanie ciągłości w wypadku nieściśliwego gruntu i wody wyraża wzór div $\vec{v_f} = 0$, złożenie obu tych wzorów, gdy grunt jest jednorodny (stałe k), prowadzi do równania Laplace'a:

$$\nabla^2 h_p = 0.$$

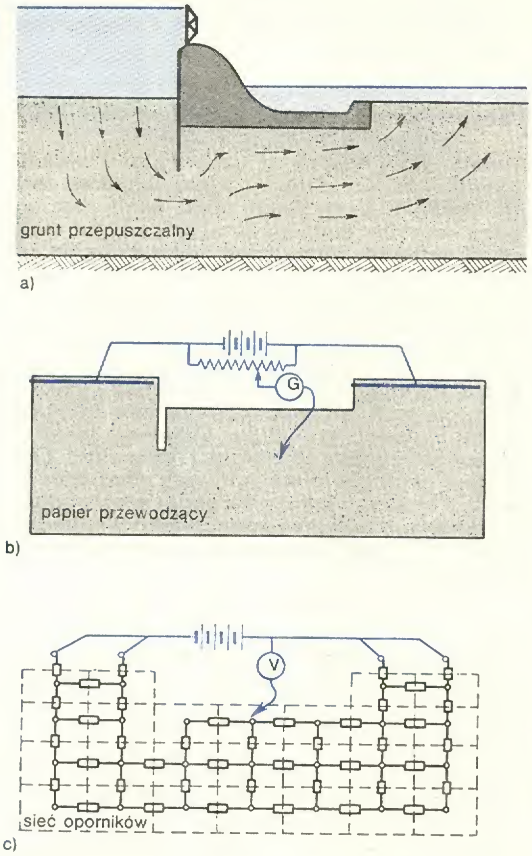
Jest to równanie identyczne co do postaci z równaniem określającym przepływ prądu elektrycznego. Można mówić o analogii dwóch zjawisk: przepływu wody w gruncie i prądu w przewodniku. Wysokość energii h_p odpowiada napięciu prądu, przepływ wody Q — natężeniu prądu, a odwrotność k — oporności przewodnika. Analogię tę wykorzystuje się dziś powszechnie przy badaniu skomplikowanych przepływów wód podziemnych — stosuje się mianowicie albo modele AEHD (modele analogii elektro-hydrodynamicznej), albo specjalne urządzenie zwane analizatorem pola.

W metodzie AEHD modelami gruntu są przewodniki o kształcie geometrycznie podobnym do badanej warstwy. Jeżeli model może być płaski, tzn. gdy przepływ odbywa się w równoległych płaszczyznach (np. pionowych), jako przewodnika używa się papieru elektroprzewodzącego; wycina się z niego kształt (rys. 7b) odpowiadający rzeczywistym warunkom przepływu (rys. 7a). Jeśli zadanie jest bardziej złożone, użyć można elektrolitu, wypełniającego wanienkę o odpowiednim kształcie. Sondą podłączoną przez galwanometr do dzielnika napięcia wyznacza się przebieg linii jednakowego napięcia, odpowiadających liniom jednakowej energii wody. Linie przepływu wody (linie prądu) są prostopadłe do linii jednakowej energii, a szybkość filtracji — odwrotnie proporcjonalna do

metoda analogii elektro-hydro-dynamicznej (AEHD)

odległości między dwoma kolejnymi liniami energii. Można więc określić zarówno przepływ, jak i ciśnienie w całym obszarze. Metoda AEHD została bardzo rozwinięta. Przy użyciu tzw. modeli odwrotnych można wyznaczać w doświadczeniach przebieg linii prądu; dołączając do modelu głównego dodatkowe oporniki, można rozwiązywać zadania bardziej złożone, a nawet — na modelach płaskich — można badać pewne przepływy trójwymiarowe.

modelowanie filtracji



Rys. 7. Filtracja wody pod budowlą: a) przekrój przez jaz piętrzący wodę, przepływ w warunkach rzeczywistych, b) model typu AEHD przepuszczalnej warstwy, na której zbudowany jest jaz, c) model warstwy przepuszczalnej zbudowany z oporników w analizatorze pola

analizator pola

Analizator pola jest urządzeniem umożliwiającym rozwiązywanie pewnego rodzaju równań różniczkowych, między innymi równania przepływu wód podziemnych. Warstwę wodonośną gruntu dzieli się na prostokątne pola, w środku których umieszcza się węzły. Przewodność gruntu między dwoma sąsiednimi polami zastępują na modelu odpowiednio dobrane oporniki, wstawione między węzły (rys. 7c). Warstwę wodonośną zastępuje siatka oporników, do której na brzegach przykłada się napięcie. Badania na analizatorze pola są szybsze i wygodniejsze niż na modelach AEHD, ale wymagają dalej idącej schematyzacji warunków przepływu.

ruch w strefie nienasyconej

Wiele ostatnio prowadzonych badań dotyczy ruchu wody w strefie niepełnego nasycenia. Siły powodujące ruch w tej strefie to ciężar wody, napięcie kapilarne, ewentualnie ciśnienie osmotyczne przy zmiennym zasoleniu wody w różnych miejscach warstwy. Jeżeli wyrazimy te siły przez wysokość energii h_p, a ilość płynącej wody — przez umowną średnią prędkość filtracji v, to i w strefie nienasyconej słuszny będzie wzór:

$$v = -k_0 \frac{dh}{dl}.$$

Współczynnik k_0 jest jednak zależny nie tylko od jakości gruntu, ale i od ilości wody zawartej w jednostce objętości gruntu θ. Również potencjał h zależy od θ. Badania przepływów w strefie nienasyconej wymaga więc ciągłych badań wilgotności gruntu. Najwygodniejszą metodą okazał się pomiar pochłaniania promieni gamma emitowanych przez izotop promieniotwórczy. Pozwala to na badanie próbek bez ich suszenia i naruszania struktury, a więc podczas doświadczeń. Ruch wody w strefie nie nasyconej ma duże znaczenie, woda bowiem wznosząca się w górę zasila korzenie roślin, zwilża grunt i zwiększa parowanie. Po deszczu natomiast woda wsiąkająca w grunt (infiltrująca) musi przebyć tę strefę, zanim osiągnie warstwę wodonośną.

Termika wód

Woda znajdująca się w zbiorniku lub między dwoma przekrojami cieku posiada w danej chwili ilość ciepła Q. Wskutek dopływu ciepła do zbiornika i odpływu ciepła do otoczenia ilość ta zmienia się o ΔQ, co powoduje zmiany temperatury wody. Ilości ciepła biorące udział w wymianie w pewnym czasie t ujmuje równanie bilansu cieplnego:

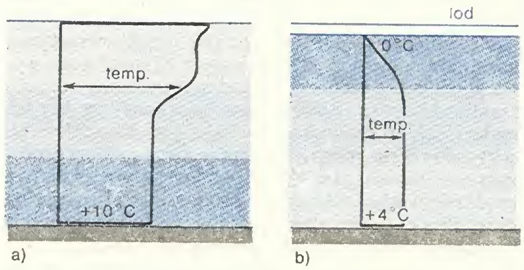
bilans cieplny

$$R+Q_e+Q_a+Q_v = \Delta Q.$$

R oznacza ilość ciepła przenoszoną przez promieniowanie (krótkie i długofalowe), Q_e — ciepło zużywane w związku z procesem parowania, Q_a — ciepło wymieniane z atmosferą przez konwekcję, Q_v — ilość ciepła transportowaną z wodą wpływającą i wypływającą z badanego obszaru. W okresie letnim woda ochładza się głównie przez parowanie, w okresie zimowym większą rolę odgrywa czynnik Q_a.

rozkład temperatury w jeziorze

W wodach płynących i w płytkich jeziorach woda ulega często wymieszaniu i znajduje się pod wpływem dobowych zmian temperatury. Natomiast w głębokich jeziorach wytwarza się pewien typowy rozkład temperatur, zasadniczo różny w okresie letnim i zimowym (rys. 8). W lecie temperatura wody w pobliżu dna wynosi ok. 10°C (tzn. nieco więcej niż średnia temperatura w Polsce); powyżej zalega warstwa wody cieplejszej, podlegającej wpływom atmosfery. W zimie



Rys. 8. Wykresy rozkładu temperatury w głębokich jeziorach: a) w lecie, b) w zimie

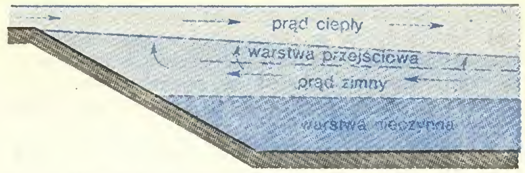
na powierzchni wody tworzy się powłoka lodowa, pod nią temperatura wody wynosi 0°C, niżej woda jest coraz cieplejsza — aż do 4°C w pobliżu dna. To ciekawe zjawisko odwrócenia rozkładu temperatur występuje dzięki anomalii, jaką posiada gęstość wody. Dwa razy do roku — na wiosnę i na jesieni — występuje przemieszczenie wód w jeziorze i wówczas temperatura w całej masie wody jest wyrównana.

zrzucenie podgrzanej wody

Poważnym problemem technicznym i naukowym jest wymiana ciepła w zbiornikach i rzekach, do których zrzuca się podgrzaną wodę z elektrowni lub innych zakładów przemysłowych. Z punktu widzenia sprawności chłodzenia aparatury i kosztów inwestycyjnych jest to rozwiązanie najwygodniejsze, wymaga jednak spełnienia dwóch warunków: woda w rzece czy jeziorze nie może być podgrzana powyżej pewnej

granicy, przekroczenie której stanow zagrożenie dla życia biologicznego (w Polsce ok. 26°C); w zbiornikach pobór i zrzut muszą być tak usytuowane względem siebie, aby do ujęcia dopływała woda odpowiednio ochłodzona.

Na ogół w rzekach następuje szybkie wymieszanie wody na całej głębokości, natomiast jest wyraźne zróżnicowanie temperatur w poprzek cieku. Całymi kilometrami woda przy jednym brzegu może być wyraźnie cieplejsza niż przy drugim (np. na odcinku ok. 40 km poniżej elektrowni w Kozienicach utrzymuje się różnica temperatur ok. 2°C pomiędzy wodą płynącą przy brzegu lewym i prawym). Podobnie wygląda sytuacja w płytkich jeziorach. W rzekach głębokich zjawisko przebiega inaczej, np. poniżej elektrowni „Dolna Odra" przy głębokości rzeki 10 m ciepła woda miesza się tylko w górnej 4-metrowej warstwie. Bardziej złożony układ powstaje w głębokich jeziorach. Woda ciepła tworzy strumień powoli rozlewający się po powierzchni i zmierzający do miejsca poboru wody (rys. 9). Strumień ten porywa ze sobą pewne ilości wody zimnej i miesza się z nią. Ochłodzenie następuje częściowo przez wymieszanie, a głównie przez parowanie i oddawanie ciepła do atmosfery. Na większej



Rys. 9. Mieszanie się ciepłego strumienia z chłodnymi wodami głębokiego zbiornika

głębokości, pod strumieniem ciepłym, tworzy się powrotny prąd wody zimnej, który uzupełnia wodę porywaną przez prąd ciepły. Zachodzące zjawiska są bardzo złożone i trudne do odtworzenia w warunkach laboratoryjnych, w skali zmniejszonej. W tym wypadku odtworzyć należy procesy dynamiki i termiki, które się rządzą różnymi prawami i wymagają spełnienia zupełnie różnych zależności przy odtwarzaniu warunków zewnętrznych powodujących przepływ i ochładzanie.

E. Czetwertyński, B. Utrysko *Hydraulika i hydromechanika*, Warszawa 1969; J. Gruat i in. *Teoria i praktyka badań hydraulicznych*, Wrocław 1970; Z. Mikulski *Zarys hydrografii Polski*, Warszawa 1965.

Geofizyka poszukiwawcza

Zbigniew Fajklewicz

Geofizyka poszukiwawcza (zwana też stosowaną lub geologiczną) jest nauką wyodrębnioną z geofizyki ogólnej. Przedmiotem jej są własności fizyczne i budowa ośrodka geologicznego przypowierzchniowej części skorupy ziemskiej oraz zależności między własnościami fizycznymi ośrodka skalnego a występowaniem w nim struktur geologicznych i złóż kopalin użytecznych. Metodami geofizycznymi wykryto i zbadano wiele złóż o znaczeniu światowym. Metody geofizyki poszukiwawczej są obecnie najważniejszymi metodami poszukiwawczymi, oddają one nieocenione usługi w wykrywaniu złóż ropy i gazu, rud, soli, siarki i innych kopalin użytecznych.

zadania geofizyki poszukiwawczej

Do geofizyki poszukiwawczej należy także badanie zmian parametrów fizycznych złóż oraz otaczających je skał, będących wynikiem prowadzonych w nich prac górniczych, jak również poznawanie własności fizykotechnicznych podłoża przeznaczonego pod budowę dużych budowli, a następnie wpływ takich budowli na własności ośrodka skalnego, Poszukiwanie i badanie wód podziemnych to jeszcze jedna dziedzina wchodząca w zakres tej nauki. Metody badawcze geofizyki poszukiwawczej stosowane są nie tylko w geologii i górnictwie, lecz również w budownictwie wodnym i lądowym, a nawet w archeologii.

Ze względu na zakres zastosowań rozróżnia się zespoły metod, które tworzą geofizykę: strukturalną, złożową, górniczą, inżynierską i techniczną. Specyfika pomiarów pozwala wyodrębnić geofizykę powierzchniową i wiertniczą, a także podziemną, lotniczą (aerogeofizykę) oraz morską. Natomiast przyjmując za kryterium podziału badane wielkości fizyczne rozróżnia się następujące metody: grawimetrię poszukiwawczą, magnetometrię poszukiwawczą, geoelektrykę poszukiwawczą, sejsmikę poszukiwawczą oraz radiometrię poszukiwawczą.

W zależności od postawionego do rozwiązania zadania dobiera się taki zestaw różnych uzupełniających się wzajemnie metod, który pozwala osiągnąć cel szybciej i przy mniejszych nakładach – wykonuje się w ten sposób kompleksowe badania geofizyczne. Kompleksowa interpretacja pomiarów geofizycznych polega na korelacji wszystkich otrzymanych wyników w celu uzyskania rozwiązania optymalnego, ograniczającego wieloznaczność rozwiązań metod geofizycznych.

Grawimetria poszukiwawcza

Niejednorodny rozkład mas w ośrodku geologicznym wpływa na wartość natężenia pola siły ciężkości na powierzchni Ziemi. W geofizyce natężenie pola siły ciężkości (stosunek siły do masy, na którą działa) nazywa się krótko polem siły ciężkości. Struktury geologiczne i antropogeniczne (będące wynikiem działalności człowieka) oraz złoża kopalin użytecznych nazywa się ciałami zaburzającymi lub anomalnymi. Rozkład pola siły przyciągania grawitacyjnego ciał zaburzających jest funkcją różnicy między ich gęstością i gęstością skał otaczających, a także zależy od ich wielkości, kształtu i głębokości występowania. Pola tych ciał wpływają na rozkład wartości ziemskiego pola siły ciężkości. Wykrywanie tych ciał oraz określenie wyżej podanych parametrów należy do grawimetrii poszukiwawczej. Podstawę metody stanowią pomiary grawimetryczne, tj. pomiary względne (wartości) siły ciężkości (wyznaczenie wartości różnicy siły ciężkości między punktami pomiarowymi) wykonywane w punktach położonych na powierzchni terenu lub na innym poziomie pomiarowym.

pomiary grawimetryczne

Do pomiarów względnych siły ciężkości służą grawimetry (il. 209, tabl. 56). Zasadniczym elementem systemu pomiarowego grawimetru jest ramię ze skupioną na końcu masą. W czasie pomiaru siłę ciężkości działającą na masę równoważy najczęściej siła sprężystości powstała wskutek wydłużenia, skręcenia lub ugięcia sprężyny lub układu sprężyn utrzymujących ramię. Deformacja sprężyny jest miarą zmiany siły ciężkości.

grawimetry

Dokładność nowoczesnych grawimetrów jest bardzo duża — dochodzi do jednej stumilionowej części wartości siły ciężkości na powierzchni Ziemi. Oznacza to, że dokładność pomiaru dochodzi do 0,01 mGal (jednostką natężenia siły ciężkości stosowaną w grawimetrii jest gal, 1 Gal = 10^{-2} N/kg).

Punkty pomiarowe na powierzchni terenu są z reguły lokalizowane na różnych względem siebie wysokościach i położone są na kompleksie skalnym o niejednakowej grubości, często bardzo zróżnicowanym, odmienna jest również rzeźba terenu otaczającego te punkty. Z tych względów zmierzone wartości nie są ze sobą porównywalne i nie mogą stanowić podstawy do

wyciągnięcia wniosków na temat budowy geologicznej. Należy zatem wyeliminować wpływ wspomnianych czynników na zmierzone wartości i sprowadzić je, czyli zredukować, do jednego poziomu odniesienia. Zazwyczaj poziomem takim jest poziom morza (powierzchnia geoidy).

redukcja pomiarów

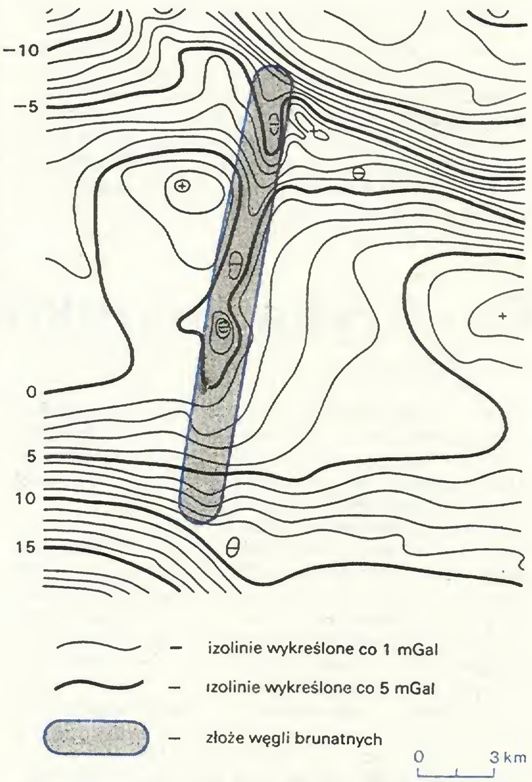
Wyniki pomiarów grawimetrycznych wykonywanych dla celów poszukiwawczych sprowadza się do poziomu odniesienia, stosując redukcję Bouguera, która uwzględnia wpływ wysokości punktu pomiarowego nad przyjęty poziom odniesienia oraz wpływ grawitacyjny mas występujących nad tym poziomem. Redukcja ta zazwyczaj jest uzupełniona poprawką topograficzną do siły ciężkości, która eliminuje składową pionową siły przyciągania pochodzącą od nierówności terenowych.

anomalie siły ciężkości

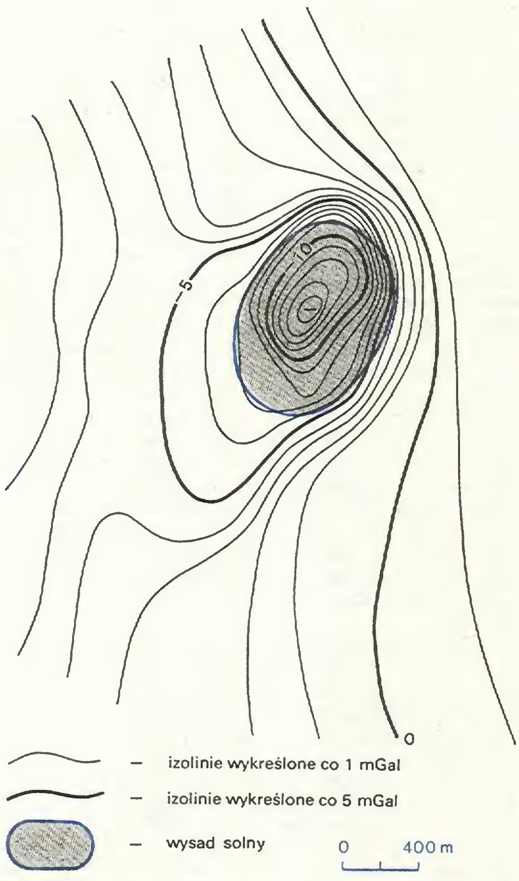
Wnioski dotyczące budowy geologicznej badanego obszaru wyciąga się na podstawie wartości anomalii siły ciężkości, czyli różnicy między zredukowaną wartością siły ciężkości w danym punkcie i odpowiadającą jej wartością normalną w tym samym punkcie. Jako wartości normalne przyjęto wartości obliczone dla wyidealizowanej Ziemi w kształcie elipsoidy z jednorodnym rozkładem masy. Anomalie siły ciężkości wskazują na stopień niejednorodności rzeczywistego rozkładu mas w ośrodku skalnym. Ponieważ sposób redukcji pomiarów może być różny, mówi się więc np. o anomalii siły ciężkości w redukcji Bouguera. Stosuje się również podział anomalii ze względu na obszar, w którym występują; można wyodrębnić anomalie lokalne i regionalne siły ciężkości. Zmierzony rozkład jest zwykle superpozycją obu wymienionych typów anomalii. Pierwsze z nich związane są z lokalnymi, a drugie — z regionalnymi ciałami zaburzającymi a więc np. strukturami geologicznymi, złożami kopalin użytecznych itp.

mapy grawimetryczne

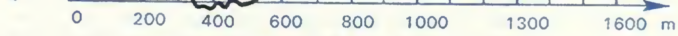
W celu graficznego przedstawienia rozkładu anomalii siły ciężkości wykreśla się na mapie izolinie, tj. linie łączące punkty o tej samej wartości anomalii. Mapy grawimetryczne są syntezą materiału pomiarowego i stanowią podstawę wnioskowania geologicznego. Dla przykładu na rys. 1 został przedstawiony rozkład anomalii siły ciężkości nad jednym z wykrytych w Polsce wysadów solnych. Pasmo ujemnych anomalii siły ciężkości występujące w centrum zdjęcia grawimetrycznego (rys. 2) jest odbiciem jednego z wykrytych w naszym kraju złóż węgli brunatnych wypełniającego rów erozyjny w utworach mezozoiku.



Rys. 2. Rozkład anomalii siły ciężkości nad złożem węgli brunatnych (wg Przedsiębiorstwa Badań Geofizycznych w Warszawie)



Rys. 1. Rozkład anomalii siły ciężkości nad wysadem solnym



Rys. 3. Wyniki podziemnych pomiarów grawimetrycznych wykonanych w chodniku górniczym (opracował Z. Fajklewicz)

Podziemne pomiary grawimetryczne pozwalają poznać szczegóły budowy geologicznej eksploatowanego górotworu. Na rys. 3 jest przedstawiony rozkład anomalii siły ciężkości zmierzony w chodniku górniczym. Na jego podstawie wykryto uskoki przecinające pokład węgla kamiennego w jednej z kopalń na Górnym Śląsku.

W zakres grawimetrii wchodzi również rozwijana w ostatnich latach metoda gradientu pionowego siły

metoda gradientu pionowego

ciężkości, polegająca na pomiarze za pomocą grawimetru i specjalnie do tego celu skonstruowanej wieży pomiarowej (il. 208, tabl. 56), różnicy siły ciężkości między dwoma punktami leżącymi na pewnej znanej odległości pionowej nad sobą. Służy ona do wykrywania małych form tektonicznych, erozyjnych i antropogenicznych, jak np. stare wyrobiska górnicze. Na il. 208 (tabl. 56) pokazano rozkład anomalii gradientu pionowego siły ciężkości nad takim wyrobiskiem wykrytym na terenie jednego z miast w Polsce.

Magnetometria poszukiwawcza

Przy poszukiwaniu i eksploatacji niektórych złóż kopalin użytecznych cennych informacji dostarcza badanie pól magnetycznych indukowanych w skałach przez stałe pole magnetyczne Ziemi. Obserwowane w przestrzeni otaczającej Ziemię pole magnetyczne jest sumą jej stałego pola naturalnego, które nazywa się normalnym, pól zmiennych oraz pól indukowanych w skałach (→ Magnetyzm ziemski). Różnicę między polem mierzonym, z którego zostały wyeliminowane zmiany dobowe, a normalnym w danym punkcie nazywa się anomalią magnetyczną. Rozróżnia się anomalie natężenia całkowitego oraz anomalie elementów i składowych pola magnetycznego Ziemi. Ze względu na obszar występowania, w magnetometrii poszukiwawczej wyróżnia się anomalie regionalne i lokalne.

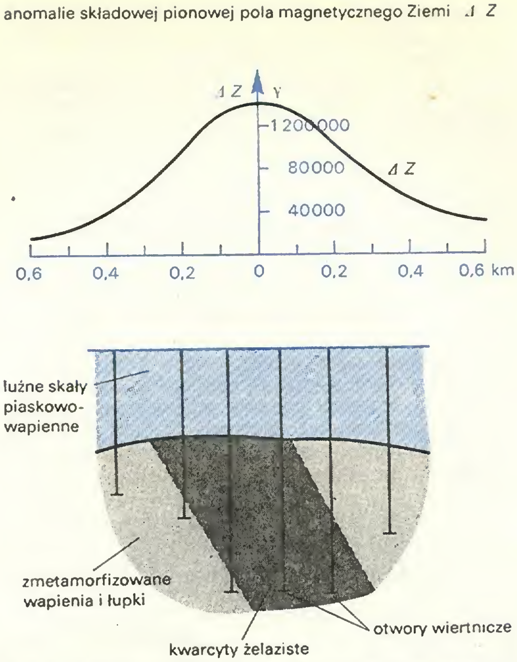
anomalie magnetyczne

Na podstawie map anomalii magnetycznych, wykreślanych analogicznie do map grawimetrycznych, można scharakteryzować ciała namagnesowane, podając ich głębokość występowania, rozmiary i kształt. Skonstruowanie precyzyjnych przyrządów mierzących pole magnetyczne z dokładnością do 1 γ (gamma jest stosowaną w geofizyce jednostką indukcji magnetycznej, $1\ \gamma = 10^{-9}$ T), znacznie poszerzyło zakres zastosowań metody magnetycznej, która nie tylko stała się nieodzowna przy poszukiwaniu złóż żelaza czy innych ferromagnetyków, ale z powodzeniem można ją również stosować w poszukiwaniu złóż metali i surowców skalnych o słabych własnościach magnetycznych. Metody magnetyczne stosuje się obecnie również przy badaniu zagadnień tektonicznych, np. do wyznaczania kontaktów skał osadowych i magmowych.

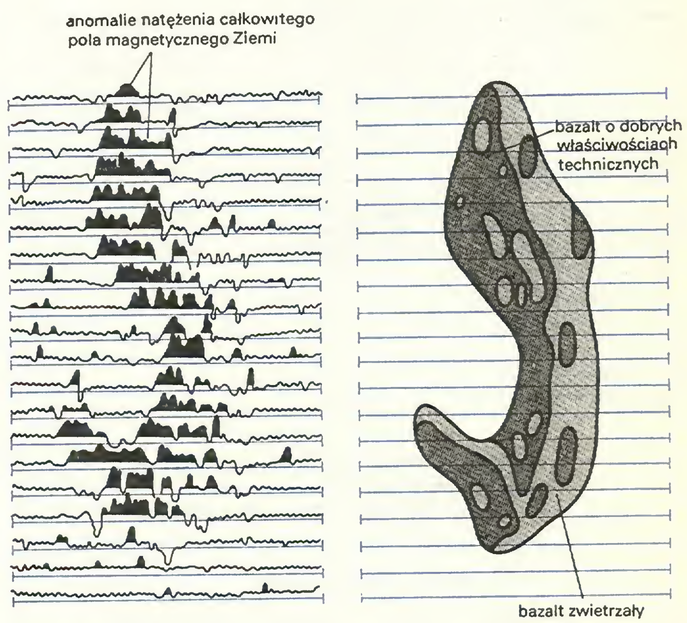
Rozwój techniki aparaturowej umożliwił wykonywanie pomiarów aeromagnetycznych i hydromagnetycznych, w których odpowiednia aparatura pomiarowa jest umieszczona w samolocie lub na statku.

Powszechnie stosowanymi przyrządami do wykonywania pomiarów magnetycznych są wagi magnetyczne służące do względnych pomiarów składowych: pionowej *Z* i poziomej *H* pola magnetycznego Ziemi oraz magnetometr protonowy. Magnetometr protonowy jest najnowocześniejszym przyrządem używanym do pomiarów bezwzględnych wartości całkowitego natężenia pola magnetycznego Ziemi *T*, a także jego poziomej i pionowej składowej. Zasadniczą częścią przyrządu jest sonda — puszka z wodą destylowaną umieszczona wewnątrz solenoidu tak ustawionego, że stały prąd elektryczny płynący przez solenoid wytwarza pole magnetyczne o kierunku w przybliżeniu prostopadłym do kierunku pola magnetycznego Ziemi. Pole solenoidu polaryzuje swobodne protony występujące w wodzie (protony orientują się w ten sposób, że osie ich momentów magnetycznych są równoległe do linii tego pola). Po wyłączeniu pola polaryzującego protony poddane działaniu ziemskiego pola magnetycznego zaczynają wykonywać ruch precesyjny wokół jego kierunku. Częstość precesji jest proporcjonalna do natężenia pola magnetycznego Ziemi. Mierząc częstość zmian siły elektromagnetycznej powstałej w uzwojeniu solenoidu można wyznaczyć wartość tego natężenia. Ilustracja 210 (tabl. 56) przedstawia magnetometr protonowy do pomiarów naziemnych polskiej konstrukcji.

magnetometr protonowy



Rys. 4. Rozkład anomalii składowej pionowej pola magnetycznego Ziemi nad złożem kwarcytów żelazistych — Kurska anomalia magnetyczna (opracował K. P. Sokołow)



Rys. 5. a) Rozkład anomalii natężenia całkowitego pola magnetycznego Ziemi nad złożem bazaltu, b) złoże bazaltu (opracował St. Małoszewski)

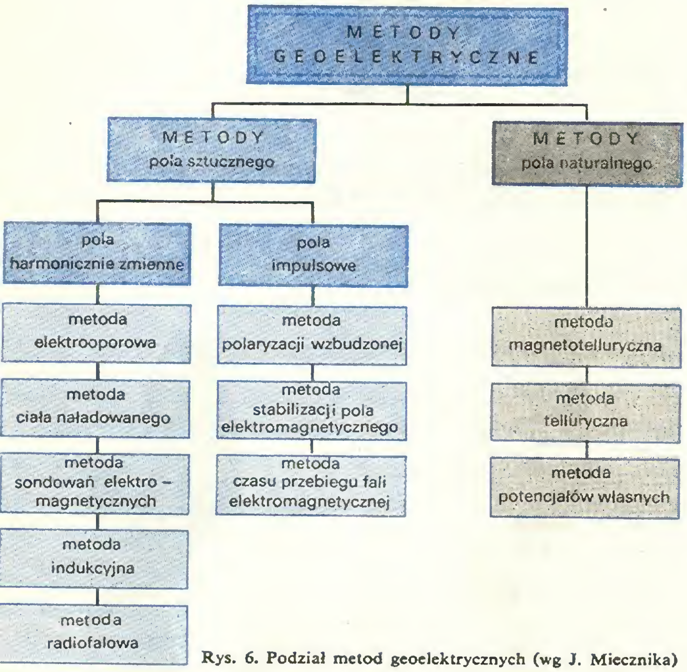
Zmierzone anomalie magnetyczne mogą wskazywać na miejsce występowania złóż rud żelaza (magnetytu, tytanomagnetytu, kwarcytów żelazistych, itp.), miedzi, ołowiu, glinu, manganu i innych metali, w których to rudach mogą występować w postaci domieszek minerały ferromagnetyczne. Przykładowo na rys. 4 jest pokazany przekrój geologiczny jednego ze złóż kwarcytów żelazistych wykrytych w Związku Radzieckim w wyniku wykonanych pomiarów anomalii składowej pionowej pola magnetycznego Ziemi.

Metodę magnetyczną można również z dużym powodzeniem stosować do wykrywania złóż surowców skalnych. Zapis anomalii natężenia całkowitego pola

magnetycznego w pewnym obszarze na Dolnym Śląsku oraz zarys wykrytego złoża bazaltu, ukazuje rys. 5. Interpretacja anomalii doprowadziła nie tylko do zlokalizowania złoża, ale również do wyróżnienia w nim stref występowania surowca o pożądanych własnościach technicznych oraz bazaltu zwietrzałego.

Geoelektryka poszukiwawcza

Do geoelektryki poszukiwawczej należy badanie naturalnych i sztucznie wzbudzonych pól elektrycznych i elektromagnetycznych w ośrodku geologicznym. Istnieją różne metody elektryczne (rys. 6), których efektywność w poszukiwaniu złóż kopalin użytecznych i badaniu budowy geologicznej ośrodka skalnego zależy od zróżnicowania skał pod względem oporu właściwego, przenikalności elektrycznej i magnetycznej oraz zdolności do polaryzowania się pod



Rys. 6. Podział metod geoelektrycznych (wg J. Miecznika)

wpływem przepływającego przez nie prądu. Większość tych metod można stosować nie tylko w badaniach naziemnych, ale również lotniczych i podziemnych.

Metoda elektrooporowa

W praktyce poszukiwawczej najszersze zastosowanie znalazła metoda elektrooporowa. Polega ona na wyznaczaniu elektrycznego oporu właściwego kompleksów skalnych; stosuje się ją do poszukiwania surowców skalnych, wód podziemnych i badania podłoża skalnego będącego przedmiotem zainteresowania budownictwa lądowego i wodnego.

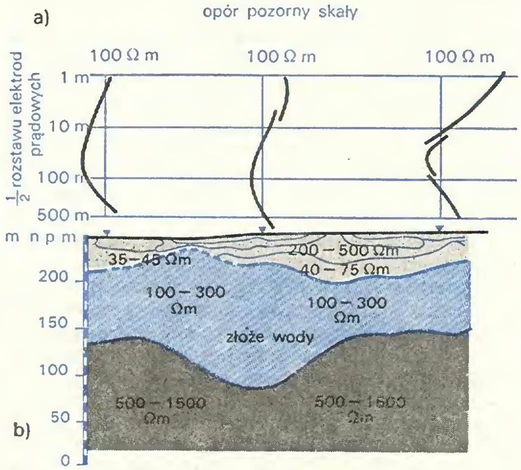
Warstwy skalne są zwykle niejednorodne pod względem własności elektrycznych, toteż w rzeczywistości wyznacza się tzw. opór pozorny, czyli opór właściwy warstwy traktowanej jako jednorodna, zależny przy tym od geometrii układu pomiarowego. Opór pozorny warstwy skalnej wyznacza się przepuszczając stały prąd elektryczny o znanym natężeniu między dwiema elektrodami prądowymi wbitymi w grunt i mierząc, za pomocą dwu elektrod pomiarowych i miernika geoelektrycznego, wartość różnicy potencjału wywołanej przepływem prądu przez warstwę skalną.

opór pozorny

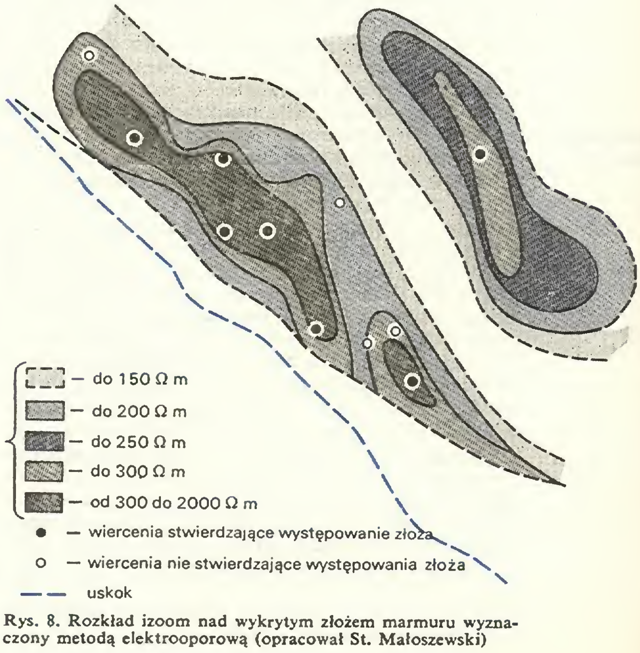
Serie pomiarów, w których rozstaw elektrod prądowych zmienia się tak, że znajdują się stale na jednej linii, a punkt leżący pośrodku między nimi (tzw. środek sondowania) pozostaje nieruchomy, stanowią pionowe sondowanie elektrooporowe. Ten system pomiarowy pozwala, stosując zmianę głębokości penetracji (przez zmianę rozstawu elektrod), wyodrębnić warstwy skalne różniące się oporem elektrycznym.

pionowe sondowanie elektrooporowe

Stosując sondowanie elektrooporowe, np. w poszukiwaniu wód podziemnych, można wyodrębnić w badanym ośrodku skalnym skały wodonośne, określić warunki ich występowania oraz formę geometryczną złoża, co pozwala ocenić jego zasoby. Pozorny opór właściwy strefy występowania wód podziemnych zawiera się w przedziale 100–300 Ωm, a wapieni kredowych podścielających skały wodonośne — w przedziale 500–1500 Ωm. Przykład zastosowania metody sondowania elektrooporowego przedstawia rys. 7. Należy tutaj nadmienić, że ok. 90% wszystkich kosztów poszukiwania wód na całym świecie przypada właśnie na tę metodę.



Rys. 7. Wyniki zastosowania metody elektrooporowej do poszukiwań wód podziemnych w skrasowiałych wapieniach jurajskich w jednym z rejonów w Polsce: a) krzywe pionowych sondowań elektrooporowych, b) przekrój geologiczny opracowany na podstawie wykonanych sondowań (opracował A. Iciek)



Rys. 8. Rozkład izoom nad wykrytym złożem marmuru wyznaczony metodą elektrooporową (opracował St. Małoszewski)

Pomiary, w których odległość wzajemna elektrod prądowych pozostaje stała, zmienia się zaś ich położenie, stanowią profilowanie elektrooporowe. System ten (przy ustalonym zasięgu głębokości) umożliwia dokładniejsze poznanie budowy geologicznej ośrodka lub ustalenie położenia granic warstw geologicznych o różnych oporach elektrycznych. W ten sposób jest możliwe wykrywanie złóż surowców skalnych występujących płytko pod powierzchnią terenu. Wyniki poszukiwań i okonturowanie złoża marmurów metodą profilowania elektrooporowego są przedstawione na rys. 8. Widać na nim że rozkład izoom jest ściśle związany z odkrytym złożem.

profilowanie elektrooporowe

Inne metody elektryczne

metoda potencjałów własnych

Górne części złóż rud i kruszców znajdują się — w wyniku działania wód podziemnych bogatych w tlen — w warunkach utleniających, a części dolne — w warunkach redukcyjnych. Powoduje to przeciwstawne naelektryzowanie stropu i spągu, tak że złoże staje się naturalnym źródłem prądu elektrycznego. Podobny efekt polaryzacji powoduje ruch roztworów wodnych w skałach. Z tego względu metoda potencjałów własnych — posługująca się pomiarami potencjałów własnych — służy głównie do poszukiwania przypowierzchniowych złóż rud oraz do wyznaczania kierunku i szybkości spływu wód podziemnych i miejsc filtracji wody. Pomiary potencjału własnego na powierzchni terenu wykonuje się za pomocą układu składającego się z miernika potencjału i dwu elektrod pomiarowych, z których jedna jest nieruchoma, a druga przemieszczana w terenie.

metoda polaryzacji wzbudzonej

Zjawisko chwilowej polaryzacji skał i złóż, wywołanej przepływem impulsu elektrycznego wykorzystuje metoda polaryzacji wzbudzonej. Pozwala ona na bezpośrednie poszukiwania złóż będących dobrymi przewodnikami elektrycznymi, a więc siarczków żelaza, miedzi, niklu, ołowiu, srebra, itp.

metoda ciała naładowanego

Zależność kształtu izopowierzchni potencjału elektrycznego od kształtu naelektryzowanego sztucznie złoża kopalin użytecznych będących dobrym przewodnikiem elektrycznym, a występujących pośród skał o dużym oporze właściwym, wykorzystuje metoda ciała naładowanego. Metoda ta służy głównie do badania form występowania wykrytych złóż i wyznaczania kierunku i prędkości przepływu wód podziemnych w otworze wiertniczym.

metoda magnetotelluryczna

metoda telluryczna

Fale elektromagnetyczne, których źródłem jest jonosfera, rozprzestrzeniają się w ośrodku geologicznym w sposób zależny od oporu właściwego skał i częstości drgań. Zależność tę wykorzystuje metoda magnetotelluryczna. Wielkościami podlegającymi bezpośredniemu pomiarowi są składowe natężenia pola elektrycznego i magnetycznego oraz kąt przesunięcia fazowego między tymi składowymi. Za wariant tej metody można uznać metodę telluryczną, w której pomiar ogranicza się tylko do składowej elektrycznej. Obie metody stosuje się do wykrywania struktur roponośnych oraz badania wgłębnej tektoniki i morfologii podłoża krystalicznego.

Fala elektromagnetyczna może być również sztucznie wzbudzona w ośrodku geologicznym, co wykorzystuje się między innymi w metodzie sondowań elektromagnetycznych.

Wspomnimy jeszcze o stosowanej do rozwiązywania problemów geologii strukturalnej metodzie stabilizacji pola, w której wykorzystuje się zależność rozprzestrzeniania się impulsu prądowego od oporu elektrycznego warstw skalnych.

Do bezpośrednich poszukiwań złóż kopalin użytecznych będących dobrymi przewodnikami elektrycznymi jest jeszcze stosowana metoda indukcyjna, polegająca na badaniu sztucznie wzbudzonego indukowanego pola elektromagnetycznego, oraz metoda radiofalowa, polegająca na badaniu pochłaniania lub odbicia fal radiowych o częstości powyżej 10^5 Hz.

Sejsmika poszukiwawcza

O budowie geologicznej i własnościach sprężystych ośrodka skalnego można wnioskować na podstawie rozchodzenia się w nim fal sprężystych (sejsmicznych). Do sejsmiki poszukiwawczej należy badanie rozchodzenia się sztucznie wzbudzonych fal sejsmicznych. Źródłem fal jest detonacja materiału wybuchowego umieszczonego w specjalnie przygotowanym płytkim otworze lub w otworze wiertniczym. Niekiedy do przekazywania drgań cząsteczkom gruntu czy skał stosuje się także wibratory elektromagnetyczne lub hydrauliczne, udary mechaniczne lub eksplozje sprężonych gazów.

granice sejsmiczne

Na każdej granicy sejsmicznej rozdzielającej dwie warstwy o różnym oporze akustycznym właściwym (zw. w geofizyce twardością akustyczną; jest to iloczyn gęstości ośrodka i prędkości rozchodzenia się w nim fali sprężystej) powstaje fala odbita, czyli refleksyjna. Jeżeli prędkość fal w warstwie nad granicą jest mniejsza niż w warstwie pod granicą, w tej pierwszej może powstać fala refrakcyjna. Pojawiają się one wówczas, gdy fala sejsmiczna pada na granicę dwóch ośrodków pod kątem krytycznym i fala załamana pod kątem 90° biegnie wzdłuż granicy dwóch ośrodków. Inicjuje ona falę sejsmiczną refrakcyjną biegnącą od granicy ku powierzchni terenu.

hodograf fali sejsmicznej

Fale sejsmiczne są odbierane na powierzchni terenu przez specjalne odbiorniki drgań zwane geofonami i rejestrowane wraz z ich czasem przyjścia w postaci sejsmogramu czyli obrazu falowego. Rejestrację można wykonać w postaci analogowej na papierze światłoczułym lub na taśmie magnetycznej, stosuje się też zapis cyfrowy. W sejsmicznych badaniach poszukiwawczych używa się też wielu geofonów położonych na powierzchni terenu w różnych odległościach od punktu wzbudzenia fal sejsmicznych, zwykle na jednej linii prostej zwanej profilem sejsmicznym. Analiza sejsmogramu pozwala wyznaczyć czas przyjścia fal sejsmicznych do geofonów oraz umożliwia porównanie amplitud zarejestrowanych drgań. Zależność czasu przyjścia fali sejsmicznej od odległości między punktem wzbudzenia a geofonami jest zwana hodografem fali sejsmicznej. Czas przebiegu fal sejsmicznych zależy od geometrii układu pomiarowego, położenia i kształtu granic sejsmicznych w ośrodku skalnym oraz prędkości rozchodzenia się fal sejsmicznych w poszczególnych warstwach skalnych.

W badaniach sejsmicznych rozróżnia się dwie podstawowe metody: refrakcyjną i refleksyjną.

metoda refrakcyjna

Metoda refrakcyjna polega na rejestrowaniu fal refrakcyjnych oraz czasu ich dojścia do geofonów. Rejestruje się albo tylko pierwsze impulsy, albo rejestruje się, a następnie koreluje impulsy dalsze (korelacyjna metoda refrakcyjna). Wyniki pomiarów przedstawia się w postaci hodografów fal refrakcyjnych, na podstawie których, używając specjalnie opracowanych metod obliczeń, wyznacza się w profilu głębokościowym bieg granic sejsmicznych refrakcyjnych. Metoda ta jest najczęściej stosowana do wyznaczania granic warstw geologicznych znacznie różniących się prędkością rozchodzenia się w nich fal sejsmicznych, stąd służy do wyznaczania głębokości i kształtu stropu podłoża krystalicznego.

metoda refleksyjna

Metoda refleksyjna polega na rejestracji fal odbitych od granic sejsmicznych oraz ich czasu dojścia do geofonów. Analiza zarejestrowanych sejsmogramów pozwala na odtworzenie rzeczywistego położenia i kształtu granic sejsmicznych refleksyjnych. W tym celu dokonuje się licznych operacji, które przekształcają zarejestrowany sejsmogram w przekrój czasowy. Odzwierciedla on rzeczywiste położenie tych granic w układzie współrzędnych kartezjańskich: długość profilu, wzdłuż którego umieszczone są geofony i czas przyjścia fal do geofonów, mierzony wzdłuż drogi prostopadłej do granicy odbijającej. Gdy tę ostatnią współrzędną zastąpić głęboko-

ścią położenia granic refleksyjnych, otrzymuje się przekrój głębokościowy. Wykreślenie tych przekrojów wymaga znajomości prędkości średniej fal sejsmicznych w badanym ośrodku skalnym czyli stosunku długości całkowitej drogi promienia fali sejsmicznej do czasu przebiegu fali wzdłuż tej drogi.

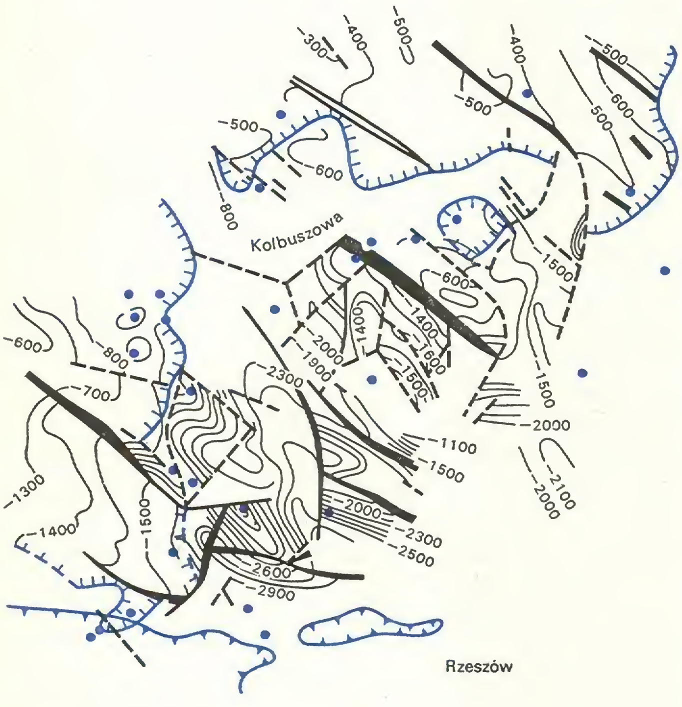
Określenie położenia i kształtu sejsmicznych granic odbijających ma zasadnicze znaczenie przy wykrywaniu struktur geologicznych. Stąd metoda refleksyjna jest najważniejszą metodą stosowaną do poszukiwań struktur gazo- i roponośnych. Charakteryzuje się ona dużym stopniem rozdzielczości, dzięki czemu można rozróżnić granice refleksyjne występujące blisko siebie.

Na uwagę zasługuje też duży zasięg głębokościowy metody, wynosi on bowiem kilka do kilkunastu km. Najlepsze wyniki w wyznaczaniu granic osiąga się, gdy nachylenie warstw nie przekracza 10–20°.

Na il. 207 (tabl. 56) jest przedstawiony przekrój czasowy odpowiadający jednemu z profilów sejsmicznych, w którym zostały wykonane badania refleksyjne w synklinorium łódzkim koło Uniejowa. Dostarcza on bardzo istotnych informacji o budowie geologicznej tego rejonu. Wykryte granice sejsmiczne wskazują na narastanie utworów cechsztynu w kierunku mezozoicznej struktury Uniejowa oraz na niezgodność zalegania granicy podsolnej cechsztynu w stosunku do kompleksu cechsztyńsko-mezozoicznego.

mapy strukturalne

W celu lepszego rozpoznania budowy geologicznej badania sejsmiczne prowadzi się w taki sposób, aby można było na ich podstawie wykonać mapy głębokościowe granic sejsmicznych — tzw. mapy strukturalne. Rysunek 9, to właśnie mapa strukturalna stropu podłoża miocenu w zachodniej części zapadliska rzeszowskiego).



— izorytmy stropu podłoża miocenu

— uskoki i strefy dyslokacyjne

— orograficzny brzeg Karpat

— zasięg występowania ewaporytów w spągu miocenu

— wiercenia

Rys. 9. Mapa strukturalna stropu podłoża miocenu w zachodniej części zapadliska rzeszowskiego (ewaporyty są to osady wytrącone z wody morskiej lub jeziornej, przy jej wyparowywaniu). Opracował A. Łapinkiewicz, M. Nowicki

zastosowanie komputerów

Zasadniczym przełomem w rozwoju badań sejsmicznych było wprowadzenie zapisu cyfrowego do pomiarów polowych oraz do procesów przetwarzania wyników otrzymywanych z terenu. Umożliwił on zastosowanie komputerów do skomplikowanych i czasochłonnych obliczeń mających na celu uzyskanie jak największej ilości danych o budowie geologicznej ośrodka. Przekrój czasowy przedstawiony na il. 207 (tabl. 56) opracowany został za pomocą komputera.

Badania sejsmiczne mogą być prowadzone nie tylko na lądzie, lecz również na morzu, w celu poznania budowy geologicznej dna morskiego. Dobre rezultaty przynosi również zastosowanie metod sejsmicznych w górnictwie do badania warunków występowania złóż stałych kopalin użytecznych.

Coraz powszechniejsze jest zastosowanie badań sejsmicznych w geologii inżynierskiej. Mała przenośna aparatura rejestrująca równocześnie przebieg refrakcyjnych fal podłużnych i poprzecznych pozwala obliczyć parametry sprężyste i wytrzymałościowe skał i gruntów, dostarczając tym samym niezbędnych informacji przy projektowaniu lokalizacji budowli i zapór wodnych.

Radiometria poszukiwawcza

Do radiometrii poszukiwawczej (zwanej też geofizyką jądrową lub jądrowymi metodami poszukiwawczymi) należy badanie własności promieniotwórczych skał oraz efektów działania promieniowania jądrowego na badany ośrodek. Na podstawie tych badań wyznacza się miejsca występowania złóż pierwiastków promieniotwórczych, złóż metali oraz innych kopalin, takich jak beryl, bar, siarka, itp. Wyznacza się również strefy zaburzeń tektonicznych oraz zróżnicowanie skał. Radiometria służy także do wyznaczania gęstości i porowatości skał oraz stopnia ich nasycenia wodą lub ropą naftową. Metodą atomów znaczonych można wyznaczyć kierunki i prędkości przemieszczania się wód w ośrodku skalnym.

Pomiary radiometryczne wykonuje się na powierzchni terenu, w kopalniach, na dnie mórz, z pokładów samolotów oraz w otworach wiertniczych (geofizyka wiertnicza). W badaniach radiometrycznych stosuje się metody pasywne, aktywne i otwartych źródeł promieniowania.

metody pasywne

Metody pasywne polegają na pomiarze promieniotwórczości naturalnej skał. Podczas badań terenowych mierzy się przede wszystkim natężenie promieniowania γ (metoda γ) lub promieniowania α izotopów radonu (lotnych) powstających z rozpadu radu, toru lub aktynu (metoda emanacyjna).

Wyodrębnienie obszarów o podwyższonej promieniotwórczości γ jest szczególnie przydatne przy sporządzaniu map geologicznych. Pomiary na dużych obszarach zwykle wykonuje się z samolotu. Wariant tej metody przystosowany do pomiarów natężenia promieniowania γ w otworach wiertniczych nosi nazwę profilowania γ.

Pomiar promieniowania radonu wykonuje się zwykle na powierzchni terenu. Metody emanacyjne umożliwiają wykrywanie rud oraz zaburzeń tektonicznych na podstawie wielkości stężenia radonu w próbkach powietrza glebowego.

metody aktywne

Metody aktywne polegają na wprowadzeniu do ośrodka skalnego (otworu wiertniczego) zamkniętego źródła promieniowania γ lub neutronów i na pomiarze produktów reakcji tego promieniowania z ośrodkiem skalnym (promieniowanie γ, neutrony lub inne produkty). Odpowiednio do tego rozróżnia się profilowanie γ–γ (PGG), γ–neutron (PGN), neutron–neutron (PNN), neutron–γ (PNG) oraz metody aktywacyjne. Metody PGG stosuje się do wyznaczania gęstości skał oraz do wydzielania w profilu geologicznym otworu wiertniczego złóż węgli kamiennych, brunatnych, złóż rud metali, itp.

Metody PNN służą do określania porowatości przewiercanych skał oraz do poszukiwania złóż manganu, boru i innych pierwiastków silnie absorbujących neutrony. Metody PNG są stosowane głównie w celu wykrywania zbiorników ropo-, gazo- i wodonośnych, określenia budowy przewiercanych skał oraz do wykrywania złóż rud żelaza, niklu itp.

Zastosowanie impulsowych źródeł neutronów prędkich wprowadzanych do otworów wiertniczych (tzw. odwiertowe generatory neutronów) i pomiar czasowych charakterystyk produktów reakcji tych neutronów ze skałami umożliwia wyznaczenie wielkości charakteryzujących ich zbiornikowe własności, a zatem wykrywanie kontaktów woda–ropa, gaz ziemny–ropa oraz wyznaczenia roponasycenia. Impulsowe źródła neutronów służą również do badania zmienności składu przewiercanych skał oraz do wykrywania złóż uranu.

metody otwartego źródła

Metody, w których stosuje się otwarte źródła promieniowania polegają na pomiarze kierunku i szybkości przemieszczenia się wód, w których zostały rozpuszczone związki pewnych izotopów promieniotwórczych. Służą one do ustalenia kierunku i prędkości przepływu wód gruntowych, wyznaczania współczynnika filtracji, określania wieku i pochodzenia wód, badania szczelności zapór wodnych itp.

Geofizyka wiertnicza

Geofizyka wiertnicza, inaczej — geofizyka otworowa, jest działem geofizyki poszukiwawczej obejmującym metody geofizyczne przystosowane do pomiarów w otworach wiertniczych, służące do badania własności fizycznych skał, a zwłaszcza ich własności zbiornikowych (porowatość, przepuszczalność, nasycenie itp.) jako kolektorów złożowych. Na podstawie kompleksu zastosowanych metod prowadzi się także badania zmienności składu i wieku przewiercanych skał, wykrywa się i bada złoża kopalin użytecznych. Do geofizyki wiertniczej zalicza się również metody fizyczno-chemiczne mające na celu lokalizowanie na podstawie profilu otworu wiertniczego złóż gazu i ropy naftowej a także metody badania stanu technicznego otworów wiertniczych.

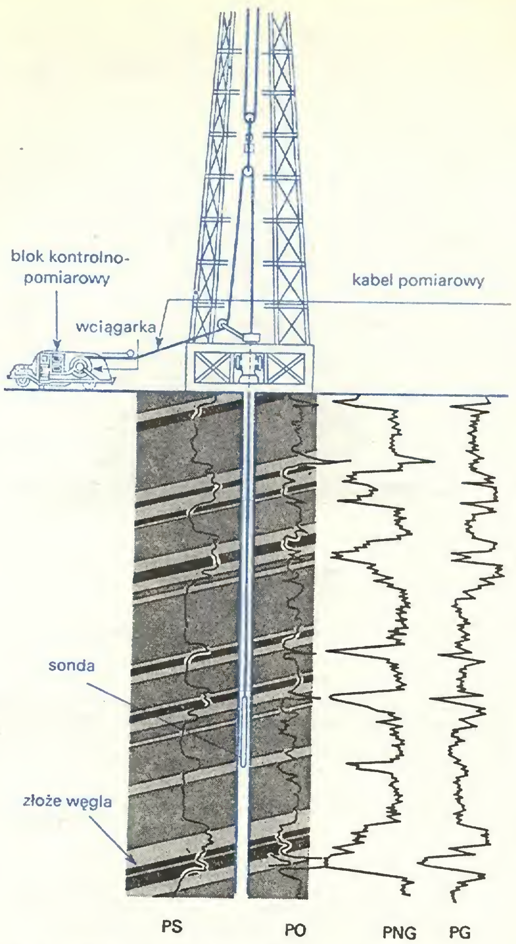
profilowanie

Pomiary zmian parametru fizycznego wzdłuż otworu wiertniczego nazywa się profilowaniem (karotażem). W odróżnieniu od instrumentów pomiarowych stosowanych w geofizyce powierzchniowej w tym przypadku aparatura składa się z dwóch oddzielnych elementów: urządzenia pomiarowo-nadawczego, czyli sondy wprowadzanej do otworu wiertniczego, oraz z bloku kontrolno-rejestrującego pozostającego na powierzchni terenu. Oba te urządzenia są ze sobą połączone wielożyłowym przewodem elektrycznym przeznaczonym do zawieszenia sondy i przesyłania informacji między sondą a aparaturą kontrolno-rejestrującą. Sposób wykonywania pomiarów geofizycznych w otworze wiertniczym jest przedstawiony na rys. 10.

sondy

Sonda jako element pomiarowo-nadawczy jest wyposażona w detektor wykrywający zmiany naturalnych lub sztucznie wzbudzonych pól fizycznych (w drugim wypadku sonda zawiera również nadajnik, czyli źródło pola pierwotnego). Sonda kompleksowa ma kilka odbiorników, co pozwala na równoczesny pomiar wielu parametrów. W zależności od potrzeb (badanie warstw o różnej grubości i oporze stosuje się różne wzajemne położenia elektrod w sondzie pojedynczej, jak również zamienia się ich biegunowość oraz funkcje. Używa się też sondy o różnych długościach.

Najprostsza pojedyncza sonda elektryczna, tzw. dwubiegunowa jest wyposażona w trzy elektrody — dwie prądowe i jedną pomiarową. Druga elektroda pomiarowa jest umieszczona na powierzchni terenu,



Rys. 10. Zasada geofizycznego profilowania w otworach wiertniczych wraz z wynikami profilowań stosowanych do wykrywania pokładów węglowych PS krzywa profilowania potencjałów polaryzacji naturalnej, PO krzywa profilowania oporności, PG krzywa profilowania γ, PNG krzywa profilowania neutron–γ (opracował St. Plewa)

a więc nie wchodzi w skład sondy. Sonda taka najczęściej jest stosowana do profilowania oporu przewiercanych skał (profilowanie elektrooporowe).

elektrometria wiertnicza

Po wyłączeniu elektrod zasilających sondę, można wykonywać profilowanie potencjałów polaryzacji naturalnej skał (metoda potencjałów własnych w geoelektryce poszukiwawczej). W tym wypadku mierzy się różnicę potencjału pola elektrycznego samoistnie powstałego w przewiercanych przez otwór warstwach. Potencjał pojawia się w wyniku procesów dyfuzyjno-absorpcyjnych w ścianach otworu, przez które przenikają wodne roztwory soli, ośrodka skalnego i płuczki, jako rezultat filtracji tych roztworów oraz wskutek procesów utleniająco-redukcyjnych zachodzących w pokładach węgli, złóż rud siarczkowych i w strukturach zawierających tlenki metali.

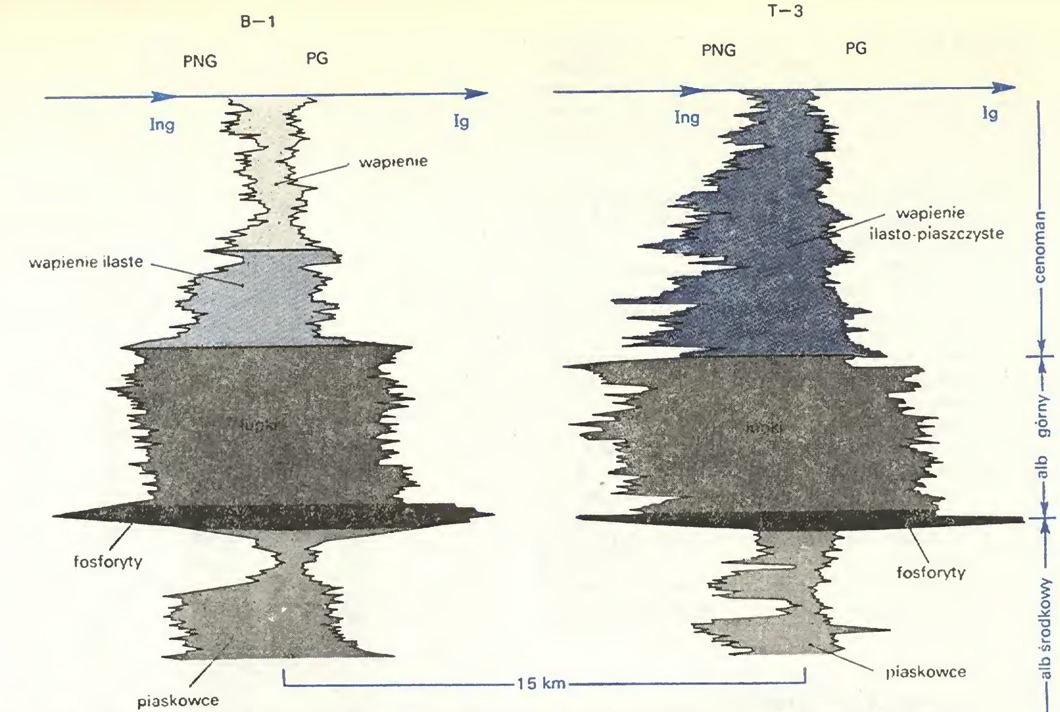
W zakres elektrometrii wiertniczej wchodzą różne warianty profilowania oporu (PO), profilowanie potencjałów polaryzacji naturalnej (spontanicznej, PS), profilowanie potencjałów polaryzacji wzbudzonej (PW) i profilowanie indukcyjne (PI).

radiometria wiertnicza

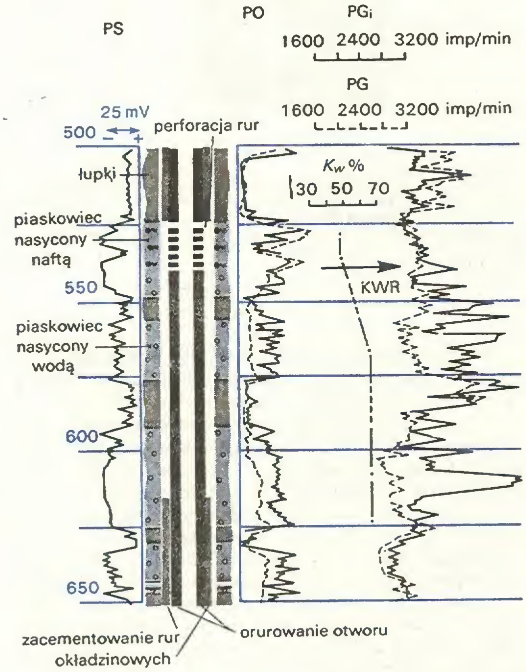
W radiometrii wiertniczej do profilowania wykorzystuje się sondy zawierające detektory promieniowania γ lub neutronów, ewentualnie także źródło promieniowania.

termometria wiertnicza

W termometrii wiertniczej wyróżnia się profilowanie cieplnego oporu skał (PTu) i profilowanie przewodzenia temperatury (PTn). Metodę PTu stosuje się w warunkach ustalonej równowagi cieplnej



Rys. 11. Przykład zastosowania radiometrycznego profilowania otworów do określenia litologii przewierconych skał oraz do ich korelacji międzyotworowej; PG krzywa profilowania γ, PNG krzywa profilowania neutron-γ, Ing — natężenie wtórnego promieniowania γ, Ig natężenie naturalnego profilowania γ, B-1, T-3 nazwy otworów wiertniczych (opracował St. Plewa)



Rys. 12. Przykład wykrywania horyzontów roponośnych na podstawie profilowania potencjałów polaryzacji naturalnej PS i profilowania oporności PO oraz kontroli stanu zacementowania otworu metodą profilowania γ (PG) i γ izotopowego (PG_i), K_w współczynnik nasycenia skał wodą, KWR kontakt woda–ropa (opracował St. Plewa)

w otworze wiertniczym; służy ona do pomiaru stopnia geotermicznego i gęstości strumienia cieplnego Ziemi. Metoda PTn polega na pomiarze temperatury w warunkach nieustalonej równowagi cieplnej; wartość temperatury w otworze jest wówczas proporcjonalna do współczynnika przewodzenia temperatury przewierconych skał. PTu wykonuje się w celu wyodrębnienia różnego typu skał i warstw gazonośnych oraz określania stanu zacementowania odwiertu.

sejsmometria wiertnicza

Do sejsmometrii wiertniczej należą pomiary prędkości rozchodzenia się fal sejsmicznych w przewierconych skałach. Wykonuje się np. profilowanie akustyczne prędkości (PAP) na potrzeby sejsmiki poszukiwawczej oraz w celu skorelowania warstw przewierconych w różnych miejscach i obliczenia ich współczynnika porowatości.

magnetometria wiertnicza

W zakres magnetometrii wiertniczej wchodzi profilowanie podatności magnetycznej (PPM) skał, którego celem jest wykrywanie i lokalizowanie ciał o dużej podatności magnetycznej, np. niektórych typów złóż rud żelaza. Jeżeli tego rodzaju złoże nie zostało przewiercone, a może występować w pobliżu otworu, to w celu wykrycia stosuje się profilowanie pola magnetycznego (PZM), na podstawie którego wyznacza się w otworach wiertniczych nachylonych zmiany wartości natężenia i kierunku pola.

kompleksowa interpretacja profilowań

W celu ograniczenia wieloznaczności interpretacji pojedynczych krzywych pomiarowych, w otworach wiertniczych wykonuje się komplet profilowań dobrany odpowiednio do podstawowego zadania geologicznego. Analiza otrzymanych w ten sposób wyników w powiązaniu z dostępnymi danymi geologicznymi i górniczymi nosi nazwę kompleksowej interpretacji profilowań. Dla przykładu na rys. 10 przedstawiony jest kompleks profilowań stosowany do wykrywania pokładów węglowych, a na rys. 11 są podane wyniki zastosowania metod radiometrycznych do określania składu i wykształcenia przewierconych skał oraz do ich korelacji międzyotworowej. Na rys. 11 zwraca uwagę wykryte złoże fosforytów.

Komplet profilowań wykonany na potrzeby górnictwa naftowego pokazano na rys. 12, podane dane stanowią pełną informację o wykrytym złożu ropy naftowej.

G. Dohr, *Applied Geophysics. Geology of Petroleum*, vol. 1, Stuttgard 1974; Z. Fajklewicz, *Grawimetria poszukiwawcza*, Warszawa 1973; St. Plewa, *Geofizyka wietrnicza*, Katowice 1972; *Przewodnik pracownika służby geologicznej*, Warszawa 1971; *Sostojanije i zadaczi razwiedocznoj gieofiziki*, Moskwa 1971; Z. Fajklewicz (red.), *Zarys geofizyki stosowanej*, Warszawa 1972.

ASTROFIZYKA I KOSMOCHEMIA

Kosmologia · Antymateria we Wszechświecie · Radioastronomia · Astronomia promieni X i γ · Astronomia w podczerwieni · Ewolucja gwiazd · Galaktyki · Gwiazdy zmienne pulsujące · Kwazary · Pulsary · Czarne dziury i zapadanie grawitacyjne · Fale grawitacyjne · Promieniowanie kosmiczne · Rozpowszechnienie pierwiastków chemicznych i molekuł we Wszechświecie · Reakcje jądrowe w gwiazdach · Powstawanie pierwiastków chemicznych ·

Kosmologia

Michał Heller

Obserwowalny Wszechświat

Podstawowe znaczenie dla kosmologii współczesnej ma doświadczalny fakt, że wszystkie informacje we Wszechświecie mogą być przenoszone tylko ze skończoną prędkością. Istnieje górna, nieprzekraczalna granica prędkości przenoszenia informacji, zwana prędkością światła i oznaczana przez c. Prędkość światła nie zależy — to również fakt doświadczalny — od wyboru układu odniesienia. Jeżeli przyjmiemy, że Wszechświat jest największym możliwym zbiorem zdarzeń, z których każde scharakteryzujemy przez podanie miejsca (x, y, z) i chwili (t), w której ono zaszło (rys. 1), to z faktu istnienia nieprzekraczalnej prędkości c rozchodzenia się informacji natychmiast wynika, że muszą istnieć zdarzenia dla nas nieobserwowalne.

Wszechświat — zbiór zdarzeń

Jeżeli bowiem dwa zdarzenia $A(x_A, y_A, z_A, t_A)$ i $B(x_B, y_B, z_B, t_B)$ są tak rozmieszczone, że informacja wysłana w chwili t_A z miejsca (x_A, y_A, z_A), przenosząca się z prędkością c, nie zdąży dojść do miejsca (x_B, y_B, z_B) przed chwilą t_B, czyli:

$$(x_B-x_A)^2+(y_B-y_A)^2+(z_B-z_A)^2 > c^2(t_B-t_A)^2,$$

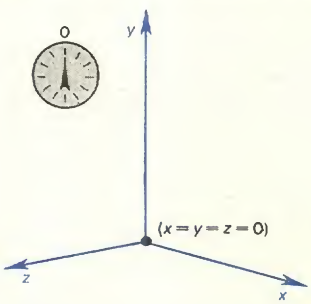
to zdarzenie B jest nieobserwowalne dla A. Mówimy, że zdarzenia A i B są oddzielone od siebie w czasoprzestrzeni przedziałem przestrzennopodobnym (rys. 2).

zdarzenia obserwowalne i nieobserwowalne

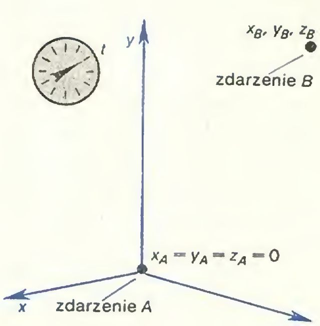
Dla zdarzeń, które mogą wymieniać ze sobą informację, mamy:

$$(x_B-x_A)^2+(y_B-y_A)^2+(z_B-z_A)^2 \leqslant c^2(t_B-t_A).$$

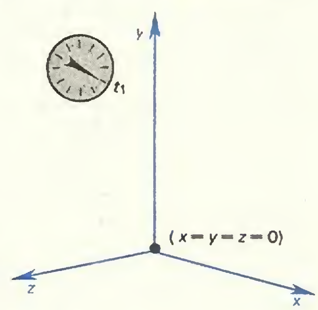
Gdy obowiązuje znak $<$, zdarzenia mogą sobie przekazywać informacje za pomocą sygnałów wolniejszych od c; mówimy wówczas, że zdarzenia są oddzielone od siebie w czasoprzestrzeni przedziałem czasopodobnym (rys. 3). Gdy obowiązuje znak $=$, zdarzenia mogą się kontaktować tylko za pomocą sygnałów rozchodzących się z prędkością c (sygnałów świetlnych); mówimy wówczas, że zdarzenia są oddzielone od siebie przedziałem typu zerowego (zerowym, świetlnym).



Rys. 1. Zdarzenie, którego wszystkie współrzędne są równe zeru: $x = y = z = t = 0$



Rys. 2. Dwa zdarzenia o różnych współrzędnych przestrzennych, lecz zachodzące w tej samej chwili. Przedział s między tymi zdarzeniami wynosi: $s^2 = (x_B-x_A)^2 + (y_B-y_A)^2 + (z_B-z_A)^2 > 0$. Przedziały, dla których $s^2 > 0$ nazywamy przedziałami przestrzennopodobnymi



Rys. 3. Zdarzenie, dla którego: $x = y = z = 0, t = t_1$. Przedział s między tym zdarzeniem a zdarzeniem z rys. 1 wynosi: $s^2 = -c^2t_1^2 < 0$. Przedziały, dla których $s^2 < 0$, nazywamy przedziałami czasopodobnymi

Rozważmy zbiór zdarzeń, z którymi zdarzenie O (określone przez: $x_O = y_O = z_O = 0$) może kontaktować się za pomocą sygnałów świetlnych, tzn.:

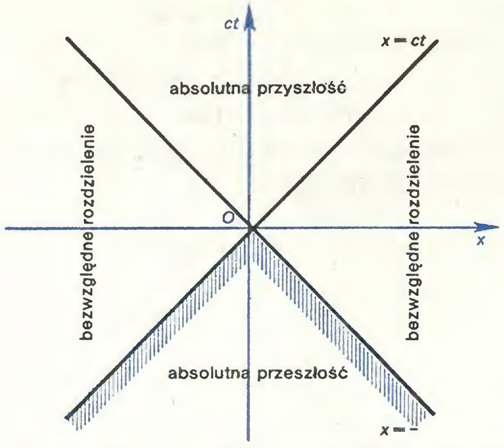
$$ct = \pm\sqrt{x^2+y^2+z^2}\,.$$

Równanie to opisuje powierzchnię dwóch czterowymiarowych stożków stykających się ze sobą wierzchołkami w początku układu odniesienia (rys. 4). Nazywamy je stożkami świetlnymi zdarzenia O. Na powierzchni stożków leżą zdarzenia, z którymi O może się kontaktować używając sygnałów świetlnych; wewnątrz dolnego i górnego stożka leżą zdarzenia, z którymi O może się kontaktować za pośrednictwem sygnałów wolniejszych niż światło. Poza tymi stoż-

stożki świetlne zdarzenia

stożek przeszłości i stożek przyszłości

kami znajdują się zdarzenia, z którymi O nie może się kontaktować. Ze stożka dolnego zdarzenie O może odbierać informacje, nazywamy go stożkiem świetlnym przeszłości zdarzenia O; do stożka górnego O może przesyłać informacje, nazywamy go stożkiem świetlnym przyszłości zdarzenia O.



Rys. 4. Stożki świetlne obserwatora O (ziemskiego astronoma). Dla jaśniejszego zobrazowania stosunków czasoprzestrzennych pominięto dwa wymiary przestrzenne (y i z), zachowując tylko wymiar ct i jeden wymiar przestrzenny x

Ponieważ za O można przyjąć dowolny punkt w czasoprzestrzeni (w dowolnym miejscu można wybrać początek układu odniesienia), powiadamy, że czasoprzestrzeń ma strukturę stożkową. Dotychczas milcząco zakładaliśmy, że czasoprzestrzeń jest płaska — tak czyni się w szczególnej teorii względności — ale, jak wiemy z ogólnej teorii względności, tak wcale być nie musi. Czasoprzestrzeń może być zakrzywiona lub nawet nieregularnie pofałdowana. Wówczas, oczywiście, deformują się także stożki świetlne. Możemy jednak zawsze rozważać tak małe obszary czasoprzestrzeni, żeby w dobrym przybliżeniu można je było uważać za płaskie; w obrębie takich obszarów spełnione są wszystkie powyższe wzory. Od dokładności pomiarów, którymi dysponujemy badając dane zagadnienie, zależy, jak duże obszary czasoprzestrzeni można uważać jeszcze za płaskie. Tak np. R. V. Pound i G. A. Rebka (1960 r.) wykorzystując zjawisko rezonansu jądrowego (→ Jądrowy rezonans magnetyczny) i efekt Mössbauera (→ Zjawisko Mössbauera), zmierzyli krzywiznę czasoprzestrzeni w obszarach przestrzennych zaledwie kilkunastu metrów i w przedziale czasu potrzebnym do pokonania tej odległości przez promień świetlny. Natomiast w zagadnieniach astronomicznych, przy obecnej technice obserwacyjnej, poprawki wynikające z ogólnej teorii względności można mierzyć dopiero na granicy odległości, do których w ogóle sięgamy.

czym dysponuje badacz Wszechświata

Przyjmijmy teraz, że O jest ziemskim astronomem, który bada obecnie Wszechświat. Dla niego obserwowalnym Wszechświatem może być tylko jego stożek przeszłości. Astronom, badając Wszechświat, posługuje się: a) znajomością fizyki sprawdzonej eksperymentalnie na Ziemi i w jej bezpośrednim sąsiedztwie (loty kosmiczne) oraz b) obserwacjami polegającymi na rejestrowaniu fal elektromagnetycznych (w zakresie optycznym i radiowym) przychodzących z przestworzy kosmicznych. A zatem w praktyce nie penetrujemy całego naszego stożka świetlnego przeszłości, lecz tylko jego obszary zakreślone na rys. 4 niebieskim kolorem.

Chcąc na tak małej bazie obserwacyjnej odtworzyć strukturę-ewolucję Wszechświata, musimy przyjąć pewne uogólnienia, a mianowicie: a) prawa fizyki sprawdzone na Ziemi i w jej najbliższym sąsiedztwie obowiązują w całym Wszechświecie, b) obserwacje astronomiczne wykonane na Ziemi są typowe dla całego Wszechświata. Przyjęcie położenia Ziemi we Wszechświecie za typowe wiąże się zwykle z nazwiskiem Kopernika. Założenia a) i b) łącznie nazywa się zasadą kosmologiczną.

zasada kosmologiczna

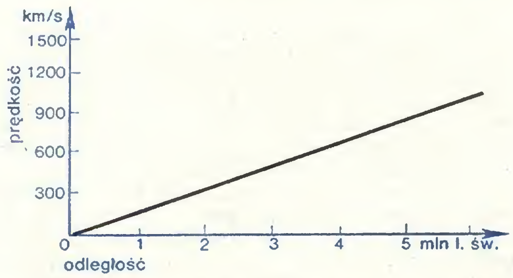
Obserwacyjne podstawy kosmologii

Obserwowany Wszechświat wypełniony jest galaktykami i gromadami galaktyk (→ Galaktyki). Oszacowania średniej gęstości materii we Wszechświecie dają wartości pomiędzy 10^{-28} a 10^{-31} g/cm^3. Zbiór galaktyk nie jest zbiorem statycznym. Systematyczne przesunięcia ku czerwieni w widmach galaktyk, zinterpretowane jako zjawisko Dopplera oznaczają, że galaktyki oddalają się od obserwatora jednakowo we wszystkich kierunkach (czyli izotropowo), z prędkościami $\vec{v}$ wprost proporcjonalnymi do odległości $\vec{r}$ (od danej galaktyki do obserwatora), tzn. według prawa:

$$\vec{v} = H\vec{r}, \qquad (1)$$

zw. prawem Hubble'a, gdzie H — stała Hubble'a (rys. 5). Rozszerzanie się przestrzeni 2-wymiarowej można przedstawić poglądowo w przestrzeni 3-wy-

prawo Hubble'a



Rys. 5. Zależność prędkości ucieczki galaktyk od odległości określona przez Hubble'a



Rys. 6. Powierzchnia balonika zakrzywiona i zamknięta stanowi przykład przestrzeni 2-wymiarowej. Przy równomiernym nadmuchiwaniu balonika kropki na jego powierzchni oddalają się od siebie zgodnie z prawem Hubble'a. Żadna z nich nie zajmuje wyróżnionego położenia. Hubbleowskie rozszerzanie nie łamie zasady Kopernika

Wartość stałej H otrzymana przez różnych badaczy

Nazwisko	Rok	H, (km/s)/Mps*
E. Hubble	1936	526
W. Baade	1950	200
A. R. Sandage	1968	$75{,}3^{+19}_{-15}$
S. Van den Bergh	1970	95^{+15}_{-12}
G. De Vaucouleurs	1970	50
A. R. Sandage, G. A. Tammann	1975	55 ± 5
P. Teerikorpi	1976	41 ± 3
L. Bottinelli, L. Gouguenheim	1976	76 ± 8

* 1 Mps = 10^6 ps ≈ $3\cdot10^{22}$ m.

miarowej (rys. 6). Poglądowe przedstawienie rozszerzania się przestrzeni 3-wymiarowej nie jest możliwe.

Wyznaczenie wartości stałej H z danych obserwacyjnych ma zasadnicze znaczenie dla kosmologii. Hubble przyjmował pierwotnie, że wartość stałej H równa się ok. 500 (km/s)/Mps. Jego oszacowanie oparte na błędnej ocenie odległości do najbliższych galaktyk kilkakrotnie poprawiano. Niektóre wyniki przedstawia tabela na str. 900.

Przyjmujemy za Sandagem, że prawdziwa wartość stałej H znajduje się pomiędzy 50 a 100 (km/s)/Mps. Zgodnie z tym mamy:

$$H = h \cdot 100 \text{ (km/s)/Mps},$$

przy czym zakładamy, że h jest zawarte pomiędzy 0,5 a 1,0.

Wielkość przesunięcia ku czerwieni w widmie danej galaktyki z definiuje się następująco:

$$z = \frac{\lambda_{obs} - \lambda_G}{\lambda_G},$$

gdzie λ_{obs} — długość fali zmierzonej przez obserwatora, λ_G — długość fali emitowanej przez galaktykę.

Ponieważ prawo Hubble'a przypisuje jednoznacznie galaktyce znajdującej się w określonej odległości $\vec{r}$ przesunięcie z, to w praktyce można przyjąć z jako miarę odległości. Z kolei w astronomii „głębia odległości" odpowiada „głębi czasu" (galaktykę odległą o 10 mln lat świetlnych oglądamy taką, jaka była 10 mln lat temu), zatem w kosmologii przyjęto wyrażać czas poprzez wartości z.

promieniowanie reliktowe

Obserwacje radioastronomiczne (wykonane po raz pierwszy przez A. A. Penziasa i R. W. Wilsona w 1965 r.) wykazały, że przestrzeń Wszechświata jest izotropowo wypełniona promieniowaniem elektromagnetycznym, zwanym promieniowaniem reliktowym lub promieniowaniem tła. W promieniowaniu tła przeważają fale o długościach milimetrowych. Jego temperatura wynosi $T_{prom} \approx 2{,}7$ K, a odpowiadająca jej liczba fotonów na jednostkę objętości $n_{prom} \approx 10^3/\text{cm}^3$ i gęstość energii $\varepsilon_{prom} \approx 4 \cdot 10^{-14}$ J/m³. Mierząc temperaturę promieniowania tła w różnych kierunkach otrzymano jednakowe wyniki (odchylenia są rzędu 10^{-4} K); zatem pole promieniowania jest izotropowe.

Zgodnie z zasadą kosmologiczną izotropowość ucieczki galaktyk i izotropowość promieniowania tła możemy ekstrapolować na cały Wszechświat. Zakładamy więc, że obserwator umieszczony w dowolnym punkcie przestrzeni widzi wokół siebie materię rozmieszczoną izotropowo (gromady galaktyk i promieniowanie). Jeżeli rozkład materii jest izotropowy względem każdego punktu przestrzeni, to jest on również jednorodny (prawo Schura). Izotropowość (brak wyróżnionych kierunków w przestrzeni) i jednorodność (brak wyróżnionych punktów w przestrzeni) są zasadniczą treścią zasady kosmologicznej. Innymi słowy, zasada kosmologiczna postuluje sferyczną symetrię rozkładu materii względem każdego obserwatora w przestrzeni. Z obserwacji zaś wynika, że średnia gęstość materii w przestrzeni jest stała, jeśli uśrednienia dokonywać po obszarach obejmujących wiele gromad galaktyk (po obszarach o liniowych rozmiarach większych niż 100 Mps). Byłoby to dodatkowym obserwacyjnym potwierdzeniem zasady kosmologicznej i wskazywałoby, że „elementarną cegiełką" w kosmologii jest gromada galaktyk.

Konstrukcja modeli kosmologicznych

Ze wszystkich oddziaływań (→ Cząstki elementarne i ich oddziaływanie) znanych współczesnej fizyce (silne, słabe, elektromagnetyczne, grawitacyjne) oddziaływania grawitacyjne, choć najsłabsze, mają największy (teoretycznie nieograniczony) zasięg (jeśli pominąć oddziaływania elektromagnetyczne, których efektywny zasięg jest zmniejszony przez to, że ładunki dodatnie znoszą się z ujemnymi). Przede wszystkim więc one są odpowiedzialne za strukturę Wszechświata. Dlatego też współczesna kosmologia teoretyczna opiera się na einsteinowskiej teorii grawitacji, czyli na ogólnej teorii względności; jest więc kosmologią relatywistyczną.

W ogólnej teorii względności pole grawitacyjne jest reprezentowane przez zakrzywienie czasoprzestrzeni. Ilościowy związek między rozkładem materii będącej źródłem pola grawitacyjnego a geometrią czasoprzestrzeni wyrażają równania pola grawitacyjnego:

równania pola grawitacyjnego

$$R_{ab} \quad \tfrac{1}{2}Rg_{ab} = - \varkappa T_{ab}. \qquad (2)$$

Lewa strona tego równania tensorowego (równoważnego dziesięciu równaniom skalarnym) jest zbudowana z wielkości geometrycznych; reprezentują one geometryczną strukturę czasoprzestrzeni. Prawa strona to tensor energii — pędu T_{ab}, opisujący rozkład gęstości, ciśnień, pędów i energii, jest on pomnożony przez stałą grawitacji Einsteina $\varkappa = 8\pi G/c^4$, gdzie G — newtonowska stała grawitacji. Tak więc równania (2) ukazują, w jaki sposób rozkład materii określa geometrię czasoprzestrzeni; a czytane w drugą stronę informują o tym, jak geometria czasoprzestrzeni determinuje ruchy materii: swobodne cząstki materialne (tzn. takie, na które nie działają żadne siły z wyjątkiem grawitacyjnych) poruszają się po liniach najprostszych (tzw. liniach geodezyjnych lub geodetykach) w czasoprzestrzeni.

Z równań (2) nie wynika statyczny model Wszechświata, co wydawało się być ich wadą. Bezpośrednio po powstaniu ogólnej teorii względności (1917 r.) Einstein nie podejrzewał nawet, że Wszechświat się rozszerza, sądził, że świat jest statyczny. Aby otrzymać statyczny model Wszechświata dodał do równań pola tzw. człon kosmologiczny Λg_{ab}, gdzie Λ — pewna nowa stała o wymiarze [długość]$^{-2}$, zwana stałą kosmologiczną. Zmodyfikowane równania pola miały postać:

$$R_{ab} - \tfrac{1}{2}Rg_{ab} + \Lambda g_{ab} = -\varkappa T_{ab}. \qquad (2a)$$

stała kosmologiczna

Obecnie, na podstawie danych obserwacyjnych, sądzi się, że wartość stałej kosmologicznej jest równa zeru, lub mało od niego różna. Dlatego w dalszym ciągu bardziej szczegółowo omówimy modele Wszechświata bez stałej kosmologicznej.

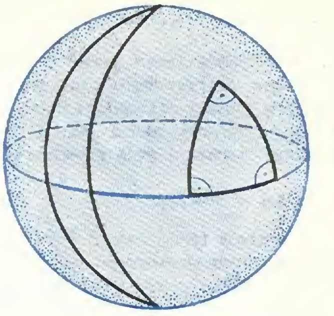
Podstawowe zagadnienie kosmologiczne polega na tym, by na podstawie danych obserwacyjnych o rozmieszczeniu materii we Wszechświecie i równań pola ogólnej teorii względności odpowiedzieć na pytanie, jaka jest globalna (w największej skali) geometryczna struktura Wszechświata. Informacje zawarte w zasadzie kosmologicznej znacznie ułatwiają znalezienie odpowiedzi na to pytanie. Geometryczna treść zasady kosmologicznej sprowadza się bowiem do następujących stwierdzeń: a) Można wybrać taki układ odniesienia, względem którego materia wypełniająca Wszechświat, średnio rzecz biorąc, spoczywa; inaczej: ponieważ cząstki materialne (galaktyki, gromady galaktyk) mogą oddalać się od siebie (lub zbliżać), układ odniesienia nie jest sztywny, lecz rozszerza się (lub kurczy) razem z materią. Nazywamy go współporuszającym się układem odniesienia. Jego matematycznym odpowiednikiem jest współporuszający się układ współrzędnych (współrzędne są jakby przyczepione do cząstek materialnych i poruszają się razem z nimi).

układ współporuszający się

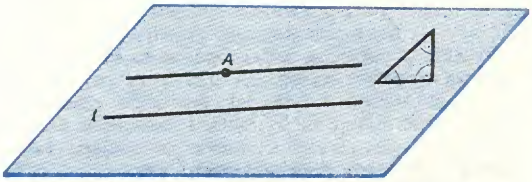
b) We współporuszającym się układzie odniesienia czasoprzestrzeń daje się w każdym punkcie rozłożyć na uniwersalny czas kosmiczny t i prostopadłą do niego 3-wymiarową przestrzeń równoczesności, t = const. Tak więc z upływem czasu (uniwersalnego) zmieniają się w sposób ciągły różne obserwowane własności geometryczne 3-wymiarowej przestrzeni. Przestrzenie równoczesności są przestrzeniami o stałej krzywiźnie: dodatniej (dwuwymiarowym przykładem przestrzeni o stałej dodatniej krzywiźnie

przestrzenie równoczesności

jest powierzchnia kuli, rys. 7), zerowej (przestrzenie płaskie, rys. 8) lub ujemnej (dwuwymiarowym przykładem jest powierzchnia wklęsłego siodła, rysunek 9).



Rys. 7. Powierzchnia kuli jest przykładem dwuwymiarowej przestrzeni o dodatniej krzywiźnie. Jest to przestrzeń zamknięta: promień światła po skończonym czasie wraca do punktu wyjścia. Wszystkie linie geodezyjne (południki) ogniskują się na biegunach. Suma kątów w trójkącie jest większa od 180°



Rys. 8. Przykład dwuwymiarowej przestrzeni płaskiej — płaszczyzna euklidesowa. Jest to przestrzeń otwarta: promień światła wysłany w przestrzeń nie powróci do punktu wyjścia. Do danej prostej *l* istnieje tylko jedna prosta równoległa przechodząca przez punkt *A* nie leżący na prostej *l* (piąty postulat Euklidesa). Suma kątów w trójkącie jest równa 180°



Rys. 9. Przykład dwuwymiarowej przestrzeni o krzywiźnie ujemnej. Jest to przestrzeń otwarta. Do danej prostej (geodetyki) *l* istnieje nieskończenie wiele równoległych przechodzących przez punkt *A* nie leżący na *l*. Suma kątów w trójkącie jest mniejsza od 180

H. P. Robertson i A. G. Walker (niezależnie od siebie) pokazali, że każda czasoprzestrzeń spełniająca zasadę kosmologiczną musi mieć następującą metrykę (czyli wzór na przedział między dwoma dowolnie sobie bliskimi zdarzeniami; metryka jest jednym z najbardziej podstawowych wyrażeń geometrycznych charakteryzujących daną przestrzeń):

$$ds^2 = c^2dt^2 - \mathcal{R}^2(t)\frac{dx^2+dy^2+dz^2}{[1+\frac{1}{4}k(x^2+y^2+z^2)]^2}, \quad (3)$$

czynnik skali $\mathcal{R}$

gdzie t — czas kosmiczny, $\mathcal{R}(t)$ — czynnik skali (zwany też promieniem Wszechświata). Czynnik skali charakteryzuje zmianę odległości między dowolnymi punktami przestrzeni: jeśli w pewnej chwili odległość między dwoma obiektami wynosi L, to po upływie czasu t odległość ta będzie równa $\mathcal{R}(t)\cdot L$. Stała k może przybierać jedną z trzech wartości: 0, ± 1. Jeśli $k = +1$, przestrzenie stałego czasu mają stałą dodatnią krzywiznę; jeśli $k = 0$, przestrzenie te są płaskie, i jeśli $k = -1$ — są to przestrzenie o ujemnej krzywiźnie.

Dziesięć równań pola (2a), po uwzględnieniu metryki (3), sprowadza się do następujących dwóch wyrażeń:

$$\varkappa\varrho c^2 = \Lambda + 3\frac{kc^2+(d\mathcal{R}/dt)^2}{c^2\mathcal{R}^2}, \quad (4a)$$

$$\varkappa p = \Lambda - \frac{2\mathcal{R}(d^2\mathcal{R}/dt^2)+(d\mathcal{R}/dt)^2+kc^2}{c^2\mathcal{R}^2}, \quad (4b)$$

gdzie ϱ — gęstość materii, a p — ciśnienie gazu galaktyk.

Musimy teraz przyjąć pewne upraszczające założenia dotyczące materii we Wszechświecie. Jeśli wyobrazimy sobie zbiór galaktyk (ściślej gromad galaktyk) jako gaz o stałej gęstości przestrzennej (gęstość gazu w czasie może ulegać zmianom wskutek ekspansji), to okazuje się, że ciśnienie tego gazu jest pomijalnie małe w porównaniu z jego gęstością. Przyjęcie zasady kosmologicznej i założenie, że ciśnienie gazu galaktyk jest równe zeru, pozwala z równań pola otrzymać jedno równanie, zwane równaniem Friedmana.

W rzeczywistym Wszechświecie stosunek $p/c^2\varrho$ wynosi ok. 10^{-6}. Możemy więc przyjąć, że ciśnienie znika:

$$p = 0. \quad (5)$$

Związek (5) nazwano równaniem stanu dla nieoddziałującej ze sobą materii pyłowej (gaz galaktyk fizycznie jest podobny do cieczy doskonałej bez ciśnienia). Wstawiając warunek (5) do równania (4b) otrzymujemy, po prostych, ale dość żmudnych przekształceniach, równanie Friedmana:

równanie Friedmana

$$\left(\frac{d\mathcal{R}}{dt}\right)^2 = c^2\left(\frac{\varkappa E}{3\mathcal{R}} + \frac{\Lambda\mathcal{R}^2}{3} - k\right), \quad (6)$$

gdzie występuje stała całkowania E:

$$E = \varrho\mathcal{R}^3c^2, \quad (7)$$

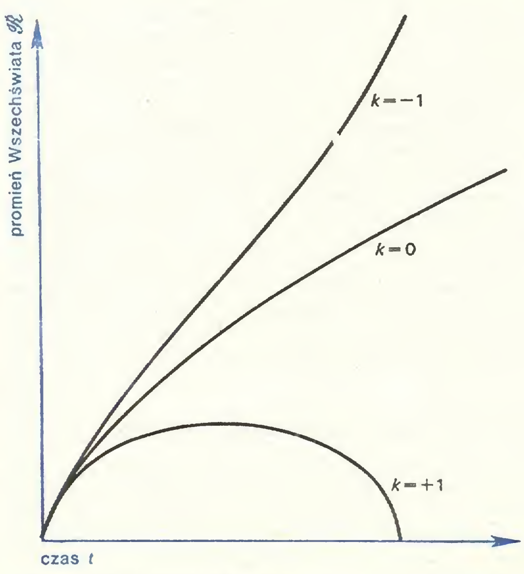
odgrywająca rolę całkowitej, zachowującej się energii pyłu (suma energii spoczynkowych wszystkich cząstek w określonej początkowo, współporuszającej się objętości).

Równanie (6) określa zmianę $\mathcal{R}$ w czasie. Każde jego rozwiązanie przedstawia ewolucję określonego modelu kosmologicznego.

Modele Friedmana

Równanie (6) dla $\Lambda = 0$ ma trzy rozwiązania. Będziemy je nazywać modelami kosmologicznymi Friedmana. Modele te przedstawia rys. 10.

Ewolucja trzech modeli Friedmana rozpoczyna się od stanu „osobliwego", który charakteryzują następ-



Rys. 10. Modele kosmologiczne z $\Lambda = 0$

pujące zależności: gdy $t \to 0$, wówczas $\mathcal{R} \to 0$ i $\varrho \to \infty$. W pobliżu początkowej osobliwości wszystkie trzy rozwiązania przebiegają jednakowo. Płaski model Friedmana (określony przez: $k = \Lambda = p = 0$), zwany również kosmologicznym modelem Einsteina–de Sittera (model E–S), rozszerza się z prędkością ucieczki. Jeżeli tempo ekspansji jest mniejsze od prędkości ucieczki, jak dla zamkniętego ($k = +1$) modelu Friedmana, model po osiągnięciu maksymalnych rozmiarów kurczy się do punktu stanowiącego końcową osobliwość. Jeżeli tempo ekspansji równa się (jak w

model ***E–S***

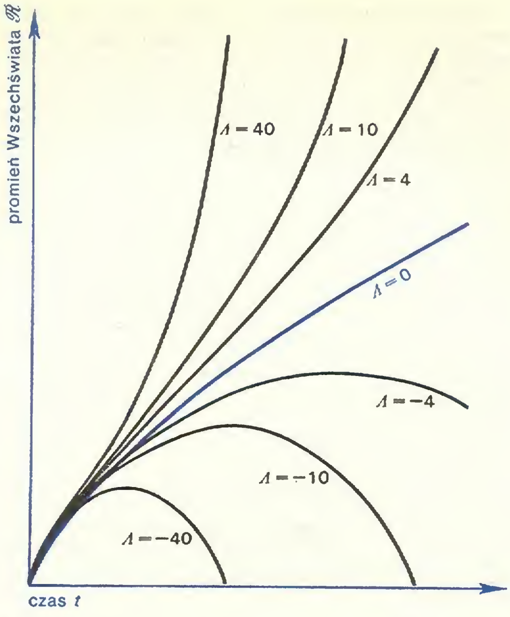
modelu E–S) lub przewyższa (jak w modelu z $k = -1$) prędkość ucieczki, ekspansja trwa nieograniczenie długo i po nieskończenie długim czasie, gdy $t \to \infty$, gęstość materii wypełniającej Wszechświat spada do zera ($\varrho \to 0$).

Jak można tego oczekiwać, prędkość ucieczki jest zależna od gęstości materii. W modelu Einsteina–de Sittera gęstość ϱ_{E-S} jest równa:

$$\varrho_{E-S} = 1{,}9 \cdot 10^{-29} h^2 \text{ g/cm}^3$$

(wielkość h wprowadziliśmy na str. 901).

W modelu zamkniętym mamy: $\varrho > \varrho_{E-S}$, a w modelu otwartym $\varrho < \varrho_{E-S}$. Zależności te są prostym testem obserwacyjnym mogącym rozstrzygnąć, który z trzech modeli Friedmana odpowiada rzeczywistości. Niestety, obecne oszacowania średniej gęstości materii we Wszechświecie są obarczone tak dużym błędem, że test ten nie daje rozstrzygających wyników.



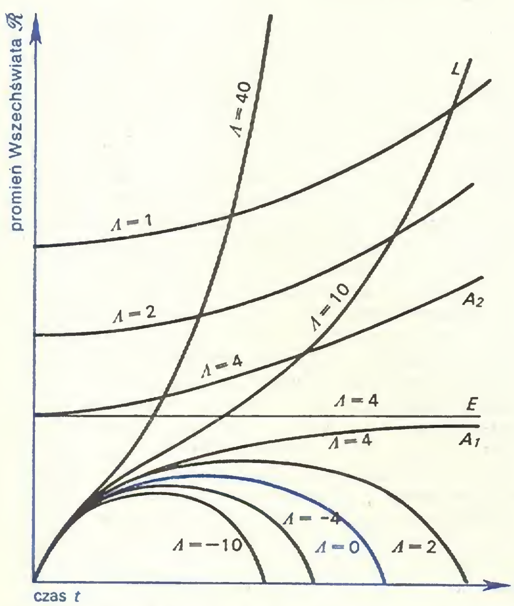
Rys. 12. Modele kosmologiczne otwarte płaskie, z $k = 0$, oraz $\Lambda \neq 0$. Model z $\Lambda = 0$ wyróżniono niebieskim kolorem

Modele Friedmana–Lemaître'a

Jeżeli przyjmiemy, że stała kosmologiczna Λ może być różna od zera, to otrzymujemy wiele modeli kosmologicznych. Będziemy je nazywać modelami Friedmana–Lemaître'a. Przedstawia je rys. 11 i 12.

model Einsteina

Wśród modeli zamkniętych (rys. 11) znajduje się jeden model statyczny (bez ewolucji), tzw. statyczny model Einsteina (model E). Jest to historycznie pierwszy spośród relatywistycznych modeli Wszechświata, podany przez Einsteina w 1917 r. Rok ten można uważać za datę powstania kosmologii relatywistycznej. Zauważmy, że w modelu E stała kosmologiczna $\Lambda \neq 0$. Właśnie po to Einstein wprowadził do teorii stałą kosmologiczną, aby uzyskać model statyczny.



Rys. 11. Modele kosmologiczne zamknięte, z $k = +1$, oraz $\Lambda \neq 0$. Model z $\Lambda = 0$ wyróżniono niebieskim kolorem. Na rys. 11 i 12 przy każdej krzywej podano wartości stałej kosmologicznej Λ w pewnych umownych jednostkach, aby zobrazować, w jaki sposób jej wartość wpływa na kształt ewolucji modelu. Zwróćmy uwagę, że dla $\Lambda = 4$, oprócz statycznego modelu Einsteina, możliwe są jeszcze dwa rozwiązania, tzw. modele asymptotyczne A_1 i A_2

model de Sittera

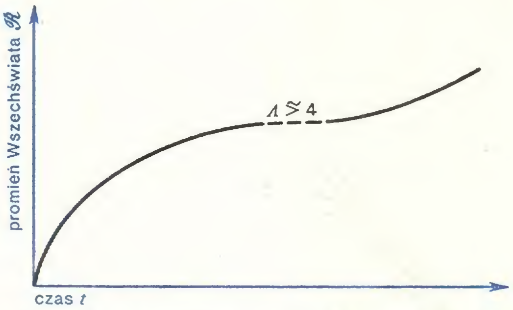
Historycznie drugim modelem kosmologicznym był model opisywany przez tzw. rozwiązanie de Sittera (model S). Swoimi, zdawałoby się paradoksalnymi, własnościami model ten w pewnym czasie sprawił teoretykom sporo kłopotów. Świat de Sittera jest pusty, a mimo to się rozszerza i jest stacjonarny. Pojęcie stacjonarności jest, w pewnym sensie, uogólnieniem pojęcia statyczności: układ jest statyczny, jeśli nic się w nim nie zmienia; układ stacjonarny może ulegać zmianom, ale jako całość średnio pozostaje zawsze taki sam. Liczba mieszkańców zamarłego w rozwoju miasteczka może być wielkością stacjonarną — mimo że poszczególni ludzie rosną i starzeją się — jeśli średnio tyle samo osób rodzi się, ile umiera. Stacjonarność świata de Sittera wynika z jego pustki, nie ulega on rozrzedzaniu w trakcie ekspansji, gdyż średnia gęstość zawartej w nim materii równa się zeru. Ale jak może się rozszerzać będąc pustym? Otóż z równań opisujących modele kosmologiczne wynika, że każdy rozszerzający się nieograniczenie model Wszechświata, w miarę gdy czas dąży do plus nieskończoności, przechodzi w świat de Sittera. Fizycznie jest to zrozumiałe: w trakcie rozszerzania ilość mas zawartych we Wszechświecie pozostaje stała, a objętość rośnie, świat staje się rzadszy, gęstość materii maleje, w granicy — gdy czas trwania ekspansji dąży do nieskończoności — gęstość materii dąży do zera. Pusty świat de Sittera jest granicą, ku której zmierza kosmiczna ekspansja.

model Lemaître'a

W historii kosmologii ważną rolę odegrał model Lemaître'a (L na rys. 11). Ewolucja tego modelu przebiega wyraźnie w trzech etapach: w pierwszym i ostatnim tempo ekspansji jest duże, w środkowym — wolniejsze. Model ten ma ciekawą własność. Im bardziej zbliżamy się z wartością Λ do wartości Λ charakterystycznej dla modelu E, tym bardziej wydłuża się środkowy etap ewolucji świata Lemaître'a (rys. 13). Dzięki temu można dowolnie wydłużać wiek świata Lemaître'a (tzn. okres, jaki upłynął od początkowej osobliwości).

Modele otwarte z ujemną krzywizną ($k = -1$) pod względem ewolucji czasowej nie różnią się istotnie od modeli płaskich ($k = 0$). Przedstawiają je krzywe jakościowo takie same (rys. 12).

modele wielkiego wybuchu Modele kosmologiczne mające początkową osobliwość nazywa się niekiedy modelami wielkiego wybuchu („Big Bang"). Zwróćmy uwagę na fakt, że wszystkie modele Friedmana-Lemaître'a, z wyjątkiem stosunkowo nielicznych odznaczających się dużą dodatnią wartością stałej kosmologicznej ($\Lambda \geqslant 4$), są modelami wielkiego wybuchu.



Rys. 13. Model kosmologiczny Lemaître'a z Λ niewiele większym od 4 ($\Lambda = 4$ w naszych jednostkach charakteryzuje statyczny model Einsteina). Zmierzając z wartością lambda do 4, okres prawie statyczny, zaznaczony linią przerywaną, można dowolnie wydłużać

Światy promieniste

gaz fotonów Istotną składową materii we Wszechświecie oprócz gazu galaktyk jest gaz fotonów. Dlatego z teoretycznego punktu widzenia ważną klasą modeli kosmologicznych są jednorodne i izotropowe modele wypełnione samym promieniowaniem, tzw. światy promieniste.

Gaz fotonów można uważać za ciecz doskonałą opisaną równaniem stanu:

$$p = {}^1/_3\varrho. \tag{8}$$

Konstrukcja modeli promienistych jest zasadniczo taka sama jak światów Friedmana-Lemaître'a, z tym, że w miejsce równości (5) przyjmuje się wzór (8).

Dla dalszych rozważań istotne okaże się porównanie gęstości energii otrzymanych na drodze teoretycznych rozważań dla świata wypełnionego materią (gazem galaktyk) i dla świata wypełnionego promieniowaniem (gazem fotonów).

W rozważanych modelach kosmologicznych objętość jest proporcjonalna do $\mathcal{R}^{-3}$, dlatego dla gęstości materii pyłowej można napisać:

$$\varrho_{\mathrm{mat}} \sim \frac{1}{\mathcal{R}^3}. \tag{9}$$

Związek ten wynika także bezpośrednio z wzoru (7). Liczba fotonów w jednostce objętości jest oczywiście również proporcjonalna do $\mathcal{R}^{-3}$, ale obliczając gęstość promieniowania należy uwzględnić poprawkę wynikającą stąd, że energia każdego fotonu zmienia się wraz z ekspansją (lub kontrakcją) Wszechświata wskutek efektu Dopplera. Z teorii wynika, że energia każdego fotonu wskutek tego efektu zmienia się proporcjonalnie do R^{-1}. A zatem gęstość energii promieniowania:

$$\varrho_{\mathrm{prom}} \sim \frac{1}{R^4}. \tag{10}$$

Inne modele kosmologiczne

Omówione modele Wszechświata, spełniające zasadę kosmologiczną, są najprostszymi modelami relatywistycznymi. Z teoretycznego punktu widzenia jest rzeczą bardzo interesującą studiowanie modeli z odchyleniami od izotropowości lub jednorodności. Tym bardziej, że w małej skali świat na pewno nie jest jednorodny (materia skupia się w galaktyki i gromady galaktyk), a w dawniejszych epokach mógł być nawet silnie anizotropowy.

model Gödla W 1949 r. Kurt Gödel zbudował model obracającego się Wszechświata. Model ten jest jednorodny: w każdym punkcie można umieścić początek lokalnie inercjalnego układu odniesienia, względem którego Wszechświat wykonuje obrót „absolutny", ale anizotropowy: rotacja wyróżnia pewien kierunek — oś obrotu Wszechświata. Model Gödla nie tłumaczy poczerwienienia w widmach galaktyk, świat gödlowski się nie rozszerza. Ponadto, we Wszechświecie Gödla nie można czasoprzestrzeni jednoznacznie rozłożyć na czas kosmiczny i przestrzenie chwilowe, co więcej — w czasoprzestrzeni Gödla istnieją zamknięte krzywe czasowe. Podróżując wzdłuż takiej krzywej można by wrócić do swojej przeszłości.

Dziś znamy wiele modeli kosmologicznych obracającego się Wszechświata. Okazuje się, że w kosmologii relatywistycznej wszystkie modele rotujące są anizotropowe, ale niekoniecznie odwrotnie — istnieją modele anizotropowe bez obrotu.

model niejednorodny Zadaniem znacznie bardziej skomplikowanym jest zbudowanie modelu niejednorodnego. Badania w tej dziedzinie są raczej mało zaawansowane. Zwykle zagadnienie upraszcza się w ten sposób, że w modelu jednorodnym, wypełnionym materią o stałej gęstości, rozważa się małe zaburzenia gęstości (odchylenia od stałej gęstości). Tego rodzaju rozważania wiążą się z zagadnieniem powstawania galaktyk (które są przecież lokalnymi zaburzeniami gęstości).

Wszechświat stacjonarny W 1948 r. powstała nowa teoria kosmologiczna, zwana teorią Wszechświata w stanie stacjonarnym, której twórcami byli H. Bondi i T. Gold. Jej podstawą jest tzw. idealna (zwana również doskonałą lub mocną) zasada kosmologiczna, która do postulatów jednorodności i izotropowości wymaganych przez zwykłą zasadę kosmologiczną dodaje postulat stacjonarności Wszechświata: obraz Wszechświata jest niezależny nie tylko od położenia obserwatora w przestrzeni, lecz także od chwili, w jakiej dokonuje on obserwacji. Jeżeli świat się rozszerza (co autorzy teorii uznali za fakt obserwacyjny) a nie jest pusty, to jego obraz jako całości może się nie zmieniać w czasie tylko wtedy, gdy gęstość materii jest stale uzupełniana przez powstającą materię. Materia, zdaniem Bondiego i Golda, powstaje w przestrzeni z niczego w ilości średnio: masa równoważna masie jednego atomu wodoru na jeden litr objętości na $5 \cdot 10^{11}$ lat. W ten sposób kosztem odstępstwa od zasady zachowania energii teoria Wszechświata stacjonarnego unika kłopotów kosmologii relatywistycznej związanych z występowaniem początkowej osobliwości.

Relatywistyczną wersję teorii Wszechświata w stanie stacjonarnym stworzył F. Hoyle. Do einsteinowskich równań pola wprowadził on wyrażenie opisujące tzw pole tworzenia materii. Rozwiązania tych równań dają modele kosmologiczne przedstawiające zasadniczo ten sam obraz Wszechświata, co model zaproponowany przez Bondiego i Golda. Wszystkie rozwiązania Hoyle'a zmierzają asymptotycznie do rozwiązania de Sittera, które — jak pamiętamy — przedstawia świat stacjonarny. Wprowadzone do równań pole tworzenia materii sprawia, że rozwiązanie de Sittera nie opisuje teraz świata pustego.

Odkrycie izotropowego promieniowania tła podważyło teorię Wszechświata stacjonarnego, która — w swojej oryginalnej wersji — nie potrafiła wyjaśnić genezy tego promieniowania. Mimo prób uratowania tej teorii, czynionych zwłaszcza przez Hoyle'a, opinia naukowców jest jej nadal nieprzychylna.

Historia Wszechświata

Dane obserwacyjne przedstawione w paragrafie: Obserwacyjne podstawy kosmologii (jednostajny rozkład gromad galaktyk, izotropowe promieniowanie tła

o temperaturze 2,7 K) świadczą, że Wszechświat w obecnej chwili jest poprawnie opisywany przez któryś z modeli Friedmana.

Wiele danych obserwacyjnych wskazuje na to, że świat jako całość podlega ewolucji; prześledźmy jego dzieje. W tym celu wykorzystamy modele Friedmana (ze stałą kosmologiczną równą zeru). Jak już wspomniano, różnice między trzema modelami Friedmana są istotne dopiero w późniejszych stadiach ewolucji. Przy cofaniu się wstecz w czasie modele te zmierzają do rozwiązania Einsteina–de Sittera. A zatem historia Wszechświata jest w przybliżeniu jednakowa dla wszystkich trzech modeli Friedmana. Najczęściej przeprowadza się obliczenia dla modelu Einsteina–de Sittera, gdyż jest on matematycznie najmniej skomplikowany (ciśnienie, stała kosmologiczna oraz krzywizna przestrzenna są równe zeru).

Obecna teoria mówi, że gdy promień Wszechświata $\mathcal{R}$ zmierza do zera, to temperatura, ciśnienie i gęstość dążą do nieskończoności. Co było przedtem — pozostaje wielką niewiadomą. Sądzi się, że przy pewnej krytycznej nadgęstości einsteinowska teoria grawitacji przestaje obowiązywać, gdyż zaczynają odgrywać rolę kwantowe efekty grawitacji. W obecnych teoriach fizycznych istotną rolę odgrywają tzw. podstawowe stałe fizyczne, są to: c — prędkość światła, $\hbar$ — stała Plancka i G — newtonowska stała grawitacji, z których można zbudować wielkość o wymiarze gęstości:

$$\varrho_{\text{kryt}} = \frac{c^5}{\hbar G^2} = 5 \cdot 10^{93} \text{ g/cm}^3.$$

Przypuszcza się, że powyżej tej krytycznej gęstości świat opisuje kosmologia kwantowa, której dotychczas nie znamy.

era Plancka

Okres, w którym we Wszechświecie panowała gęstość krytyczna nazywamy erą Plancka lub epoką progu (tabela: Ewolucja Wszechświata). Począwszy od tego progu słuszne są znane nam dziś prawa fizyki, a szczególnie einsteinowska teoria grawitacji. Wszechświat był wówczas wypełniony cząstkami elementarnymi, niezwykle stłoczonymi, odznaczającymi się ogromnymi energiami kinetycznymi. A ponieważ miarą średniej energii kinetycznej jest temperatura, więc i temperatury panujące wówczas we Wszechświecie były niewyobrażalnie wielkie. Tak np. można obliczyć, że przy planckowskiej gęstości krytycznej musiały panować temperatury rzędu 10^{33} K.

Chwile następujące bezpośrednio po erze Plancka giną w mroku domysłów. Grawitony, czyli kwanty pola grawitacyjnego, bardzo słabo oddziałują z innymi cząstkami elementarnymi. Jednakże w ekstremalnych warunkach gęstości i temperatur oddziaływania takie musiały być znaczące. Czas ustalenia się równowagi termodynamicznej między grawitonami a innymi cząstkami elementarnymi był bardzo krótki i odpowiadał mniej więcej erze Plancka. Oznacza to, że zaraz po erze Plancka grawitony oddzieliły się od reszty gorącej materii i przestały z nią oddziaływać. Jeśli hipoteza ta jest słuszna, to we Wszechświecie do dziś powinno znajdować się tzw. tło grawitonowe. Detekcja grawitonów tła mogłaby, w zasadzie, przynieść empiryczne informacje o warunkach fizycznych panujących w pobliżu ery Plancka. Niestety, praktyczne przeprowadzenie tego obserwacyjnego testu wykracza daleko poza nasze obecne możliwości techniczne.

tło grawitonowe

Przy ogromnych gęstościach (o rzędy wyższych niż gęstość 10^{14} g/cm^3 charakterystyczna dla materii wewnątrz jądra atomowego) i temperaturze ponad 10^{12} K o stanie materii decydowały przede wszystkim oddziaływania silne. Najważniejszym składnikiem materii na tym etapie były cząstki silnie oddziałujące — hadrony (bariony i mezony). Najrozmaitsze odmiany hadronów znajdowały się w równowadze termodynamicznej ze sobą, nie tylko te najbardziej trwałe, znane nam z wielu laboratoriów ziemskich, jak nukleony, hiperony, piony, kaony, ale i wiele tak krótko żyjących (w warunkach laboratoryjnych), że być może jeszcze ich nie zdołaliśmy zaobserwować. Poza cząstkami istniały w dużych ilościach antycząstki oraz promieniowanie. Nieustannie powstawały pary cząstka-antycząstka w wyniku oddziaływań promieniowania z innymi cząstkami, jednocześnie przebiegały procesy anihilacji par.

Wraz z rozszerzaniem się Wszechświata zmniejszała się gęstość materii i temperatura oraz malała energia promieniowania. Warunki równowagi między wysokoenergetycznym promieniowaniem a parami ulegały stopniowo zmianie na niekorzyść par, które powstawały coraz rzadziej z uwagi na malejącą energię kwantów promieniowania. Równowaga dynamiczna przesuwała się wyraźnie na korzyść procesów anihilacji. W końcu wszystkie pary barion-antybarion uległy anihilacji. Pozostały tylko te bariony, które nie miały z czym zanihilować. Erę powyższą, zwaną erą hadronową, trwającą ok. 10^{-4} s, aż do ustalenia się gęstości rzędu 10^{14} g/cm^3 i temperatury rzędu 10^{12} K, przeżyć mogły tylko najtrwalsze hadrony: proton p i neutron n. Obok nich pozostały jeszcze leptony, które poprzednio stanowiły nic nie znaczącą domieszkę, obecnie wysunęły się na pierwsze miejsce.

era hadronowa

Ewolucja Wszechświata

Czas	Gęstość g/cm^3	Temperatura, K	Procesy fizyczne	Co pozostało do dziś?
?	?	?	Kosmologia kwantowa	
Era Planoka				
10^{-44} s	10^{93}	10^{33}		
Era hadronowa			Oddzielenie się grawitonów (?)	tło grawitonowe (?)
			plazma w równowadze termodynamicznej (nukleony-antynukleony, mezony-antymezony, elektrony-pozytony, neutrina-antyneutrina, kwarki-antykwarki (?))	nukleony
			anihilacja pionów-antypionów (mezonów π)	
10^{-4} s	10^{14}	10^{12}		
Era leptonowa			anihilacja mionów-antymionów (mezonów μ)	
			elektrony-pozytony, neutrina-antyneutrina w równowadze	
			oddzielenie się neutrin	tło neutrinowe
			anihilacja elektronów-pozytonów	
10 s	10^{4}	10^{10}		
Era promienista			plazma i promieniowanie w równowadze	
			synteza helu i lekkich pierwiastków aż do litu	hel, lekkie pierwiastki
			rekombinacja wodoru i oddzielenie się promieniowania (fotonów)	wodór, promieniowanie tła
10^{6} lat	10^{-21}	3000		
Era galaktyczna			powstanie galaktyk	galaktyki
			powstanie gwiazd	gwiazdy
			synteza reszty pierwiastków chemicznych we wnętrzach masywnych gwiazd	pierwiastki chemiczne
			powstanie planet i Ziemi	planety, Ziemia
10^{10} lat	10^{-31}–10^{-28}	2,7	powstanie życia na Ziemi — kosmiczne „teraz”	życie, człowiek

Stąd też nazwa następnej ery: era leptonowa. W temperaturach między 10^{12} i 10^{10} K (podajemy je w kolejności, w jakiej następowały w czasie po sobie) materia obok nukleonów (p i n) składała się z leptonów i ich antycząstek (obecnie znamy ich cztery, ale być może istnieje ich sześć lub więcej). Były więc pary elektron-pozyton, μ^-–μ^+, obok nich dwa rodzaje neutrin i dwa rodzaje antyneutrin oraz fotony. Wraz z postępującą ekspansją Wszechświata spadała temperatura i gęstość, pary ulegały anihilacji, w końcu erę leptonową przetrwały tylko elektrony oraz promieniowanie. (Przypominamy, że promieniowaniem zwykło się nazywać cząstki o zerowej masie spoczynkowej; zalicza się więc do niego promieniowanie elektromagnetyczne, jak również neutrina ν_e, ν_μ i ich antycząstki $\bar{\nu}_e$, $\bar{\nu}_\mu$). Miony, ze średnim czasem zaniku rzędu 10^{-6}, nie miały większej szansy na przeżycie ery leptonowej, trwającej zaledwie ok. 10 sekund.

era leptonowa

Podczas ery leptonowej — w chwili, którą oznaczmy t^{++} — neutrina przestały niemal oddziaływać z resztą materii; jako wartość krytyczną temperatury poniżej której nastąpiło oddzielenie się neutrin od reszty materii przyjmuje się $2 \cdot 10^{10}$ K. Począwszy od chwili, gdy temperatura ta została osiągnięta, plazma (złożona przede wszystkim z elektronów, protonów i neutronów) jest dla neutrin całkiem przezroczysta. Gdyby się nam udało zaobserwować te pierwotne neutrina, tworzące tzw. promieniowanie neutrinowe tła, to otrzymalibyśmy informacje z chwili ostatniego ich oddziaływania z materią, tj. o stanie Wszechświata w chwili t^{++}. Chociaż detekcja tła neutrinowego jest teoretycznie łatwiejsza od obserwacyjnego wykrycia tła grawitonowego, i ona obecnie leży poza naszymi praktycznymi możliwościami.

era promienista

Gdy zakończyła się przewaga leptonów, podobnie jak przewaga barionów w poprzedniej erze, zaczyna się następna era, zwana erą promienistą. Największy wkład do gęstości materii (i energii) na początku tej ery wnoszą cząstki promieniowania: fotony, neutrina, antyneutrina. Ale zgodnie z teorią względności gęstość energii związanej z cząstkami o zerowej masie spoczynkowej maleje ze wzrostem czynnika skali $\mathcal{R}$ jak $\mathcal{R}^{-4}$, gęstość zaś energii związanej z cząstkami o niezerowej masie spoczynkowej (takich jak e, p, n) maleje jak $\mathcal{R}^{-3}$ (por. wzory (9) i (10)). Stąd też z upływem czasu gęstość promieniowania zmniejsza się szybciej niż gęstość materii korpuskularnej. Wreszcie — pod koniec ery promienistej — wkład od promieniowania do całkowitej gęstości energii jest już do pominięcia. Wraz z ekspansją Wszechświata oraz jego stygnięciem, podczas ery promienistej, maleje średnia energia fotonów. Dopóki była ona większa od energii wiązania najlżejszego jądra atomowego, deuteronu ($p+n$), równej 2,2 MeV, nie mogły tworzyć się oddzielne jądra atomowe z gęstego gazu nukleonów. Gdy tylko powstał jakiś deuteron, natychmiast trafiał weń foton i rozbijał go z powrotem na neutron i proton. Dopiero gdy średnia energia fotonu spadła poniżej 1 MeV, deuterony stały się trwałe. Mogły one jednocześnie reagować z sobą, tworząc dalsze jądra, np. $2\,^2D \rightarrow {}^4He+\gamma$, a także wychwytywać obecne wciąż jeszcze protony i neutrony. Tworzyły się pierwsze jądra atomowe, aż do litu włącznie.

era galaktyczna

Gdy gęstość materii promienistej stała się równa gęstości materii korpuskularnej, temperatura ośrodka wypełniającego Wszechświat wynosiła 3000–30 000 K (trudno nam dziś jeszcze podać jej wartość z większą dokładnością); było to jeszcze podczas ery promienistej. Wciąż jeszcze nie istniały atomy, jedynie jądra atomowe (wodoru, helu i litu) oraz swobodne elektrony. Dopiero w temperaturach poniżej 3000 K wodór ulega rekombinacji, kończy się okres, podczas którego promieniowanie elektromagnetyczne ulegało łatwo rozpraszaniu thompsonowskiemu na elektronach swobodnych. Materia staje się plazmą nieprzezroczystą dla pozostającego z nią w równowadze (do tej chwili!) promieniowania elektromagnetycznego. Promieniowanie to oddziela się od materii korpuskularnej i niemal przestaje z nią oddziaływać. Ma to doniosłe konsekwencje obserwacyjne. Promieniowanie tła, jakie dziś obserwujemy (porównaj paragraf: Obserwacyjne podstawy kosmologii), jest źródłem informacji o stanie świata w okresie, gdy fotony tego promieniowania po raz ostatni oddziaływały z materią korpuskularną, a więc o stanie świata w pewnej chwili t^+. Obserwatorzy nazywają stan świata w chwili t^+ powierzchnią ostatniego oddziaływania (promieniowania z materią korpuskularną).

Dopiero gdy gęstość energii promieniowania jest znacznie mniejsza od gęstości materii korpuskularnej, mogą się tworzyć lokalne zagęszczenia materii. Ciśnienie promieniowania nie spowoduje już ich natychmiastowego rozbicia. Z zagęszczeń tych mogą się tworzyć galaktyki, z mniejszych zagęszczeń materii w tych ostatnich — gwiazdy. Nastaje era gwiazdowa, zwana też erą galaktyczną. Era ta trwa do dziś.

Fizyka w pobliżu osobliwości

Współczesna kosmologia daje teoretycznie zadowalający obraz ewolucji Wszechświata począwszy od ery Plancka. Ale co było przedtem? Początkowa osobliwość nakłada ograniczenia na nasze możliwości badawcze. Dlatego też podejmuje się liczne próby zbudowania modelu kosmologicznego bez początkowej osobliwości. Teoretycznie można to zrobić wprowadzając do równań pewne nowe elementy (np. tzw. spin makroskopowy, lepkość objętościową), jednakże oznacza to odejście od pierwotnych — logicznie najprostszych — równań Einsteina i powoduje nowe trudności interpretacyjne. S. W. Hawking, R. Penrose i R. Geroch udowodnili szereg twierdzeń, z których wynika, że przy realistycznych założeniach fizycznych równania Einsteina z reguły prowadzą do osobliwości, takich jak osobliwość początkowa w kosmologii. Ponadto, przyjmowana obecnie interpretacja promieniowania tła jest silnym argumentem przemawiającym za tym, że świat u początku swojej obecnej ewolucji przechodził przez stan nadgęsty, który z punktu widzenia naszej dzisiejszej wiedzy fizycznej na pewno zasługuje na miano „osobliwego".

osobliwość początkowa

Znamy dziś wiele modeli kosmologicznych, które mogą opisywać historię Wszechświata w nadgęstych stanach bliskich osobliwości. Wszystkie one — z wyjątkiem tylko „nielicznych" — przedstawiają światy anizotropowe. Kosmologowie zwykli mawiać, że w zbiorze wszystkich możliwych modeli kosmologicznych modele jednorodne i izotropowe są podzbiorem miary zero. Powstaje pytanie: dlaczego właśnie któryś z tych nielicznych, a więc mało prawdopodobnych modeli, w dobrym przybliżeniu opisuje rzeczywisty świat? Wydaje się, że istnieją dwie możliwe odpowiedzi na to pytanie, pierwsza: świat od początku jest opisywany przez jednorodny i izotropowy model i nie należy zadawać pytania, dlaczego akurat przez taki model a nie inny; druga: w stanach bliskich osobliwości świat był chaotyczny, niejednorodny i nieizotropowy, dopiero wskutek pewnych procesów fizycznych stosunkowo szybko nastąpił proces wyrównywania się niejednorodności i izotropizacji świata (w związku z tym przyjęło się mówić o friedmanizacji świata).

friedmanizacja świata

Z obserwacji promieniowania tła wiadomo, że świat jest izotropowy przynajmniej od chwili t^+. Można obliczyć, że jeżeli chwili tej odpowiada $z \geqslant 7$, to fotony promieniowania tła, jakie rejestrujemy obecnie z kierunków na sferze niebieskiej różnych co najmniej o 30°, w chwili t^+ znajdowały się w obszarach czasoprzestrzeni fizycznie całkowicie odizolowanych od siebie. Aby jakikolwiek sygnał mógł przeniknąć z jednego takiego obszaru do drugiego, musiałby się poruszać z prędkością większą od prędkości światła. A zatem żaden proces fizyczny nie mógł doprowadzić do wyrównania się temperatur w obu tych obszarach,

co z kolei jest niezbędne, by Wszechświat mógł stać się w przybliżeniu friedmanowski.

model wielkiego mieszania

Aby pokonać tę trudność, C. W. Misner skonstruował modele kosmologiczne z tzw. wielkim mieszaniem. Są to anizotropowe modele kosmologiczne mające zamknięte przestrzenie chwilowe. Proces mieszania polega na tym, że we wczesnych stadiach ewolucji promień światła zdąży kilkakrotnie obiec świat dookoła, dzięki temu nie ma izolowanych od siebie obszarów czasoprzestrzeni i występuje zjawisko wygładzania się Wszechświata.

Okazało się jednak, że w zbiorze wszystkich modeli kosmologicznych podzbiór modeli z wielkim mieszaniem jest także bardzo mały i pytanie, dlaczego świat jest jednorodny i izotropowy, pozostaje otwarte. Misner zaproponował więc inny mechanizm wygładzania się nierównomierności Wszechświata.

procesy wygładzające anizotropię

Anizotropowa ekspansja Wszechświata powoduje występowanie zjawiska lepkości (tarcia sąsiednich warstw materii rozszerzającej się z różnymi prędkościami), które z kolei działa wygładzająco. Duży wkład do tego procesu mogła dawać lepkość powodowana przez oddziaływanie neutrin z elektronami w temperaturze ponad 10^{10} K, w stanach bardzo bliskich osobliwości.

Modele anizotropowe jeszcze z innego powodu okazały się przydatne do opisu świata w pobliżu osobliwości. J. B. Zeldowicz i A. A. Starobiński wykazali, że wskutek działania efektów kwantowych w bardzo wczesnych stadiach ewolucji anizotropowych modeli kosmologicznych następuje intensywne tworzenie się cząstek elementarnych z krzywizny czasoprzestrzeni. Mówiąc obrazowo, krzywizna czasoprzestrzeni odpowiada pewnej energii pola grawitacyjnego. Przy bardzo wielkich krzywiznach w pobliżu osobliwości energia ta może przetwarzać się w cząstki. W modelach izotropowych, w zasadzie, proces ten nie występuje. Powstawanie cząstek z silnie zakrzywionej czasoprzestrzeni jest również źródłem pewnego rodzaju lepkości, bardzo silnie wygładzającej pierwotną anizotropię Wszechświata. W pobliżu osobliwości gęstość energii pochodząca od cząstek wypełniających Wszechświat nie ma żadnego wpływu na przebieg jego ewolucji, dominującą rolę odgrywa krzywizna czasoprzestrzeni (okres ten nazywa się stadium próżniowym); z czasem jednak gęstość energii, pochodząca od cząstek powstających z próżni, wzrasta i całkowicie determinuje dynamikę ekspansji Wszechświata. Jak piszą C. W. Misner, K. S. Thorne i J. A. Wheeler, „być może anizotropia stworzyła materialną zawartość naszego Wszechświata, sama przy tym ulegając wygładzeniu".

Testowanie modeli kosmologicznych

Testowanie modeli kosmologicznych jest odrębnym działem kosmologii, który bardzo silnie wiąże ją z astronomią pozagalaktyczną. Przedstawimy tu tylko samą zasadę testowania. Dla uproszczenia ograniczymy się do modeli Friedmana, przyjmując, że stała kosmologiczna jest równa zeru.

W testowaniu modeli ważną rolę odgrywają następujące wielkości: a) średnia gęstość materii we Wszechświecie ϱ; b) stała Hubble'a $H = (d\mathcal{R}/dt)/\mathcal{R}$; c) parametr hamowania (deceleracji) $q = \frac{\mathcal{R}(d^2\mathcal{R}/dt^2)}{(d\mathcal{R}/dt)^2}$. Wartości tych parametrów dla obecnej epoki można wyznaczyć obserwacyjnie (oznaczać je będziemy indeksem zero).

wiek Wszechświata

Gdyby Wszechświat zawsze rozszerzał się dokładnie według prawa Hubble'a (równanie 1), to odwrotność stałej Hubble'a $T_0 = 1/H_0$ określałaby wiek Wszechświata, tzn. okres czasu, jaki upłynął od początkowej osobliwości do chwili obecnej. W ogólniejszym przypadku, gdy rozszerzanie się Wszechświata nie odbywa się dokładnie według prawa Hubble'a, wiek Wszechświata może być mniejszy od T_0 (rys. 14).

Parametr q określa krzywiznę wykresu $\mathcal{R} = \mathcal{R}(t)$. Jeżeli q jest dodatnie, tempo ekspansji maleje z czasem. Wyznaczenie q_0 jednoznacznie wybiera jeden z modeli Friedmana. Przedstawia to tabela:

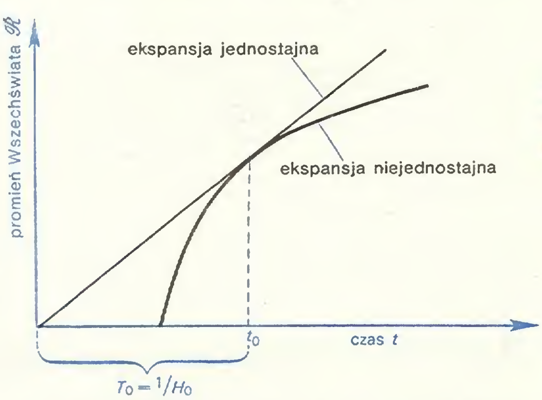
Modele Friedmana

q_0	Model Friedmana wypełniony materią pyłową ($p = 0$)
$< 1/2$	otwarty, $k = -1$
$= 1/2$	płaski, $k = 0$
$> 1/2$	zamknięty, $k = +1$

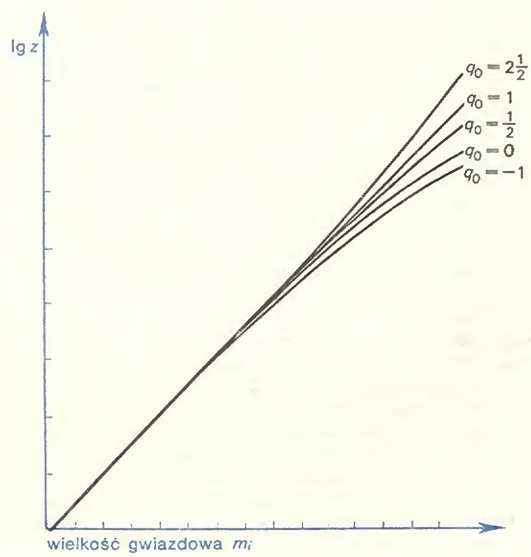
Na podstawie każdego modelu kosmologicznego można przewidywać wartości liczbowe różnych wielkości obserwowanych. Wymienimy dla przykładu kilka takich wielkości:

(*a*) przesunięcie ku czerwieni w widmach galaktyk;
(*b*) liczba obiektów (galaktyk, radioźródeł) widzialnych aż do pewnej odległości;
(*c*) średnice kątowe obiektów astronomicznych;
(*d*) jasność obiektu (mierzona w tzw. wielkościach

wyznaczanie wieku Wszechświata

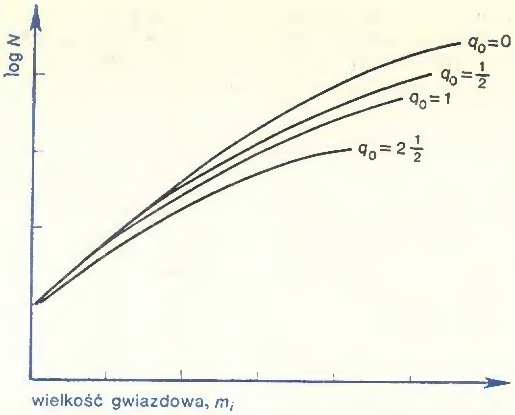


Rys. 14. Odwrotność stałej Hubble'a T_0 oznacza wiek Wszechświata, czyli okres od początku ekspansji $t = 0$ aż do chwili obecnej t_0, przy założeniu że ekspansja zawsze odbywała się z jednostajną prędkością. W rzeczywistości ekspansja nie musiała być jednostajna

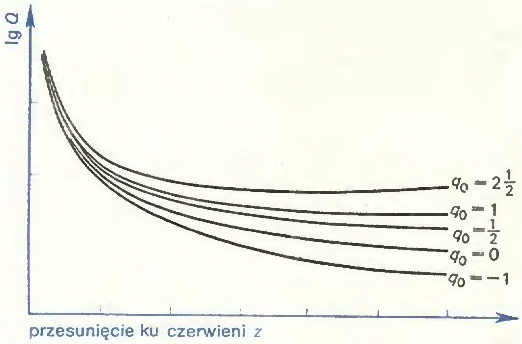


Rys. 15. Zależność przesunięcia ku czerwieni (z) od wielkości gwiazdowej (m_i); wykres: $a = f(d)$

przykłady zależności funkcyjnych



Rys. 16. Zależność liczby galaktyk $N(m_i)$ widzialnych aż do wielkości gwiazdowej m_i od danej wielkości gwiazdowej m_i (wykres: $b = h(d)$)



Rys. 17. Zależność średnic kątowych Q galaktyk od przesunięcia ku czerwieni z (wykres: $c = g(a)$).

gwiazdowych) przyjęta jest za miarę jego odległości: im obiekt jest bardziej odległy, tym jego jasność widzialna jest mniejsza.

Ponieważ wielkości te nie są na ogół niezależne, między wieloma ich parami można określić zależności funkcyjne: $a = f(d)$, $c = g(a)$ itp. Różne modele prowadzą do różnych postaci takich funkcji — porównanie krzywych teoretycznych z wynikami obserwacji daje możliwość wyboru tego modelu (lub tej klasy modeli), który najlepiej zgadza się z danymi obserwacyjnymi. Przykłady takich „wykresów testujących" (w uproszczeniu) przedstawiają rys. 15, 16 i 17.

Obserwacyjnych danych potrzebnych do testowania dostarcza radioastronomia i astronomia pozagalaktyczna. Precyzja współczesnych obserwacji znajduje się na granicy dokładności potrzebnej do testowania. W najbliższych dziesiątkach lat — zwłaszcza dzięki możliwości przeprowadzania obserwacji poza atmosferą ziemską — należy się spodziewać napływu ciekawych informacji. Niezbędny jest także postęp teorii. Dziś znacznie więcej wiemy o zależnościach między wielkościami dającymi się obserwować, niż o ewolucji obiektów, na których przeprowadza się testowanie (galaktyk, ich gromad, radioźródeł, kwazarów). Tymczasem uwzględnienie efektów ewolucyjnych tych obiektów może istotnie zmienić „krzywe testujące" w różnych modelach kosmologicznych.

Dotychczas najskuteczniejszym testem obserwacyjnym okazało się odkrycie i zbadanie własności izotropowego promieniowania tła. Istnienie tego promieniowania — przy obecnej jego interpretacji — dowodzi, że nasz świat przynajmniej od epoki t^+ jest światem Friedmana — lub prawie friedmanowskim (tzn. z dobrą dokładnością jednorodnym i izotropowym) — ze stanem nadgęstym u początków swojej ewolucji.

H. Bondi *Kosmologia*, Warszawa 1965; H. Bondi *Wszechświat nieznany*, Warszawa 1964; M. Heller *Początek świata*, Kraków 1976; M. Heller *Wobec Wszechświata*, Kraków 1971; D. W. Sciama *Kosmologia współczesna*, Warszawa 1975; J. B. Zeldowicz, I. D. Nowikow *Strojenije i ewolucyja wsielennoj*, Moskwa 1975; W. Zonn *Kosmologia współczesna*, Warszawa 1968.

Antymateria we Wszechświecie

Marcin Kubiak

Z punktu widzenia fizyki współczesnej nie ma zasadniczej różnicy między cząstkami i antycząstkami. Antycząstki doskonale mieszczą się w schematach teoretycznych i obficie występują w laboratoriach. Istnienie antycząstek jest przejawem spełnienia przez przyrodę pewnych ogólnych praw symetrii (→ Cząstki elementarne i ich oddziaływania), które w wypadku cząstek trwałych, takich jak elektrony, protony i neutrony, sprowadzają się do prawa symetrii ładunkowej. Zgodnie z tym prawem, równie prawdopodobne jest istnienie elektronu o ładunku ujemnym jak i elektronu o ładunku dodatnim (pozytonu). Podobnie proton może być nośnikiem zarówno ładunku dodatniego (zwykły proton) jak i ujemnego (antyproton). Makroskopowe własności materii zbudowanej z antycząstek, czyli tzw. antymaterii powinny być identyczne z własnościami materii zbudowanej ze zwykłych cząstek elementarnych (dla której szwedzki fizyk H. Alfvén zaproponował nazwę koinomaterii, od greckiego słowa *koinos* oznaczającego rzecz dobrze znaną). Atomy antymaterii różniłyby się od atomów materii tylko tym, że ich jądra miałyby ujemne ładunki, zobojętnione przez powłokę dodatnio naładowanych pozytonów. Poza tym wszystkie ich własności fizyczne i chemiczne byłyby takie same jak zwykłych atomów; antyatomy wysyłałyby np. dokładnie takie same linie widmowe i wchodziłyby w identyczne reakcje chemiczne jak zwykłe atomy.

koinomateria a antymateria

Lokalnie w sposób trwały może istnieć tylko jeden rodzaj materii: spotykające się ze sobą cząstki i antycząstki tego samego rodzaju ulegają niemal natychmiastowej anihilacji, polegającej na zamianie całej masy obu cząstek (oraz ich łącznej energii kinetycznej przed spotkaniem) w energię fotonów promieniowania elektromagnetycznego. Materia i antymateria nie mogą istnieć obok siebie. Na Ziemi antycząstki pojawiają się tylko w bardzo szczególnych okolicznościach i żyją do pierwszego spotkania z odpowiednią cząstką materii, wraz z którą ulegają anihilacji. Podobny los byłby udziałem cząstki naszej materii, która znalazłaby się w otoczeniu zbudowanym z antymaterii.

anihilacja materii i antymaterii

Stwierdzenie pełnej symetrii właściwości materii i antymaterii w skali mikroskopowej nasuwa w sposób naturalny pytanie, czy symetria ta jest również spełniona w skali całego Wszechświata, tzn. czy oba jej rodzaje występują w nim w jednakowych ilościach, czy też jest on zdominowany przez jeden tylko rodzaj materii. Gdyby okazało się, że cały Wszechświat jest zbudowany z jednego rodzaju materii, wówczas należałoby wyjaśnić, dlaczego przyroda — mając do wyboru dwa identyczne rodzaje materii — wybrała tylko jeden. Uznanie tego za przypadek byłoby oczywiście tylko pozornym wyjaśnieniem; wybór jednego rodzaju materii oznaczałby raczej, że własności materii i antymaterii nie są tak dokładnie symetryczne jak to sobie obecnie wyobrażamy.

zasada symetrii materii i antymaterii

Przy teraźniejszym stanie naszej wiedzy o materii, założenie pełnej równoważności materii i antymaterii wydaje się bardziej atrakcyjne, zwłaszcza z filozoficznego punktu widzenia, i dlatego warto jest zastanowić

się przez chwilę nad płynącymi z niego konsekwencjami kosmologicznymi. Warunek symetrii materii i antymaterii we Wszechświecie może być spełniony w różny sposób. Wprawdzie materia i antymateria nie mogą istnieć w bezpośrednim sąsiedztwie, ale możemy przyjąć, że:

a) antymateria istnieje w odległych obszarach Wszechświata, których nie możemy obserwować, natomiast nasza Metagalaktyka (‣ Galaktyki), czyli część Wszechświata dostępna obserwacjom, jest zbudowana z materii,

b) antymateria istnieje również w Metagalaktyce, np. średnio co druga galaktyka jest zbudowana z antymaterii,

c) antymateria istnieje w obrębie każdej galaktyki, tzn. średnio co druga gwiazda jest zbudowana z antymaterii.

metody rozróżniania materii od antymaterii

Obecnie brak jest jakichkolwiek danych obserwacyjnych na temat ewentualnego istnienia lub braku antymaterii w obserwowanej przez nas części Wszechświata. Wyda się to może zaskakujące, ale nie jesteśmy w stanie jednoznacznie stwierdzić nawet tego, czy najbliższa nas gwiazda, Proxima Centauri, jest zbudowana z materii czy z antymaterii. Zastanówmy się bowiem w jaki sposób moglibyśmy na odległość odróżnić gwiazdę zbudowaną z materii od gwiazdy zbudowanej z antymaterii. Teoretycznie możliwe są dwa sposoby. Po pierwsze, moglibyśmy starać się wykorzystać zjawisko Zeemana w widmach gwiazd mających silne pole magnetyczne: atomy materii i antymaterii, których ładunki są nawzajem wymienione, różnie reagują na obecność sił magnetycznych (i elektrycznych). Mówiąc dokładniej, obraz rozszczepienia zeemanowskiego linii widmowych wysyłanych przez antyatomy znajdujące się w polu magnetycznym o danym kierunku jest taki, jak obraz rozszczepienia zeemanowskiego linii wysyłanych przez zwykłe atomy znajdujące się w polu magnetycznym o przeciwnym kierunku. Gdybyśmy więc znali kierunek pola magnetycznego, moglibyśmy bez trudu stwierdzić, czy promieniujące atomy są zbudowane z materii, czy z antymaterii. Niestety, obecnie nie znamy żadnych metod określania kierunku pola magnetycznego gwiazd, a tym samym nie potrafimy odróżnić namagnetyzowanej antymaterii od materii namagnetyzowanej w kierunku przeciwnym.

promieniowanie neutrinowe

Inną możliwość odróżnienia gwiazd zbudowanych z materii i antymaterii mogłyby dać obserwacje strumienia neutrin opuszczających — jak się powszechnie sądzi — wnętrza gwiazd, w których przebiegają reakcje jądrowe przemiany wodoru w hel (‣ Ewolucja gwiazd). Reakcjom tym towarzyszy emisja pewnej liczby neutrin, które ze względu na bardzo mały przekrój czynny na oddziaływanie z materią bez trudu opuszczają miejsce swego powstania i swobodnie przebiegają ogromne przestrzenie Wszechświata. Oczywiście, reakcje takie przebiegające między atomami antymaterii byłyby źródłem strumienia antyneutrin. Chociaż laboratoryjne odróżnienie neutrina od antyneutrina nie nastręcza obecnie większych trudności, to jednak obie te cząstki są tak bardzo „nieuchwytne", że dotychczas nie udało się nawet uzyskać jednoznacznych informacji na temat strumienia neutrin słonecznych; tym mniejsze są też nadzieje na zbudowanie w niedalekiej przyszłości znacznie czulszego „neutrinowego teleskopu gwiazdowego".

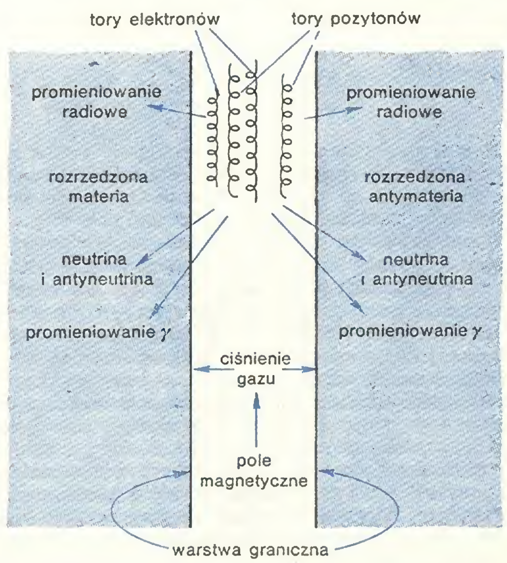
anihilacja

Tak więc, obecność antymaterii możemy stwierdzić tylko wówczas, gdy wchodzi ona w bezpośredni kontakt z materią i ulega anihilacji. Dzięki tej własności antymaterii możemy wykluczyć istnienie większych jej skupisk w Układzie Słonecznym, którego niemal cała objętość jest wypełniona rozrzedzonym obłokiem materii opuszczającej Słońce w postaci tzw. wiatru słonecznego (→ Promieniowanie kosmiczne). Gdyby cząstki wiatru słonecznego były innego rodzaju niż cząstki tworzące planety, wówczas obszar spotkania się materii słonecznej z materią planetarną byłby silnym źródłem promieniowania (‣ Astronomia promieni X i γ) pochodzącego z anihilacji. W Układzie Słonecznym źródeł takich jednak nie obserwujemy.

Z podobnych powodów możemy uznać za mało prawdopodobne, by nasza Galaktyka zawierała dużą liczbę gwiazd zbudowanych z antymaterii. W tym wypadku należałoby oczekiwać anihilacji materii gwiazdowej z otaczającą materią rozproszoną, chociaż samo zjawisko przebiegałoby w sposób nieco bardziej złożony i nasze wnioski nie byłyby tak jednoznaczne. W warunkach panujących w Galaktyce najprawdopodobniejsze jest oddziaływanie na siebie materii w stanie rozproszonym. Zetknięcie się ze sobą dwu gwiazd (prowadzące w przypadku spotkania się gwiazdy zbudowanej z materii z gwiazdą zbudowaną z antymaterii do gwałtownej eksplozji o niespotykanej sile) jest wydarzeniem niezwykle mało prawdopodobnym ze względu na małe rozmiary gwiazd w porównaniu z ich średnimi odległościami. Znacznie bardziej prawdopodobne jest zetknięcie się ze sobą dwóch obłoków materii rozproszonej, przy czym jeden z nich może być związany z gwiazdą, np. jako obłok stanowiący pozostałość po obłoku materii, z którego powstała gwiazda, lub jako obłok materii wyrzuconej z gwiazdy w postaci wiatru gwiazdowego. Z reguły, obłoki materii międzygwiazdowej, jak również rozrzedzone obłoki materii okołogwiazdowej, są przeniknięte słabym polem magnetycznym (np. wiatr słoneczny unosi ze sobą nie tylko zjonizowaną materię słoneczną, ale również „wmrożone" w nią pole magnetyczne). Nie tracąc więc nic z ogólności rozważań, możemy wziąć pod uwagę zderzenie dwóch obłoków rozrzedzonej materii w obecności pola magnetycznego.

zderzenie obłoków materii i antymaterii

Zderzenie, o którym mowa, jest przedstawione schematycznie na rys. 1. W miarę zbliżania się obłoków ku sobie, w obszarze ich zetknięcia utworzy się warstwa graniczna, w której będą przebiegać procesy



Rys. 1. Warstwa graniczna powstająca podczas zderzenia się obłoków rozrzedzonej materii i antymaterii jest źródłem promieniowania neutrinowego, promieniowania γ i promieniowania radiowego

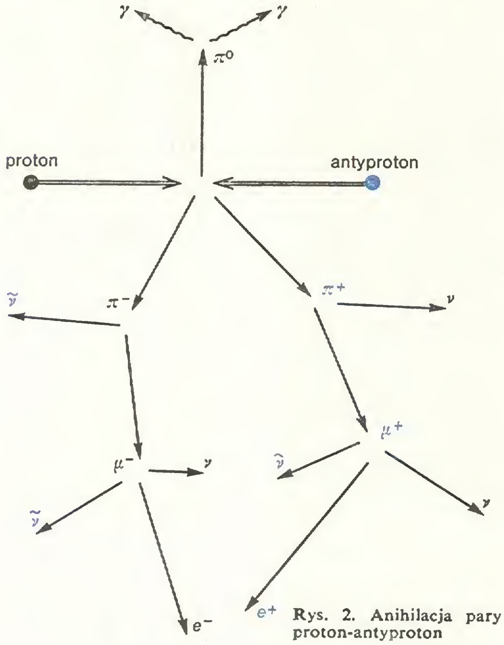
anihilacji. Ponieważ materia we Wszechświecie składa się na ogół z wodoru, wystarczy rozważyć tylko anihilację par elektron-pozyton i proton-antyproton.

anihilacja pary elektron-pozyton

Anihilacja pary elektron-pozyton polega na bezpośredniej zamianie masy spoczynkowej (i energii kinetycznej) obu cząstek w dwa lub trzy fotony γ o łącznej energii ok. 1 MeV (w wypadku zderzenia dwóch obłoków materii międzygwiazdowej możemy pominąć energię kinetyczną cząstek przed zderzeniem). Fotony te bez przeszkód opuszczają miejsce powstania i dodają się do ogólnego tła promieniowania elektromagnetycznego we Wszechświecie.

anihilacja pary proton-antyproton

Anihilacja pary proton-antyproton jest zjawiskiem nieco bardziej złożonym (rys. 2). Pierwszym produktem anihilacji są głównie mezony π, które za pośrednictwem mezonów μ rozpadają się ostatecznie na elektrony i pozytony. Bilans energetyczny całej reakcji przedstawia się następująco: z 1800 MeV energii odpowiadającej masie spoczynkowej pary proton-antyproton, 900 MeV unoszą neutrina i antyneutrina, 600 MeV unoszą kwanty γ, a 300 MeV przypada na energię kinetyczną elektronów i pozytonów (ich masa spoczynkowa odpowiadająca energii ok. 1 MeV jest w tym wypadku całkowicie do pominięcia).



Rys. 2. Anihilacja pary proton-antyproton

Neutrina, antyneutrina oraz kwanty γ opuszczają bez przeszkód miejsce anihilacji, natomiast elektrony i pozytony, jako cząstki obdarzone ładunkiem elektrycznym, zakreślają spirale wokół linii pola magnetycznego, które nie pozwala im oddalić się zbytnio od miejsca powstania. Przy energiach rzędu 150 MeV i przy natężeniach pola rzędu 10^{-9} T (typowa wartość natężenia kosmicznych pól magnetycznych) rozmiary spirali są rzędu 10^{11} m, czyli ok. 10^{-5} roku świetlnego, co w skali kosmicznej jest odległością bardzo małą. Pojawienie się cząstek o dużych energiach kinetycznych prowadzi do szybkiego wzrostu temperatury gazu w warstwie granicznej. Wzrost temperatury pociąga za sobą wzrost ciśnienia, które doprowadza do rozdzielenia obszarów zajmowanych przez materię i antymaterię, a tym samym do zmniejszenia tempa anihilacji. Zderzenie obłoku materii z obłokiem antymaterii jest więc zjawiskiem przebiegającym stosunkowo spokojnie i wbrew oczekiwaniom nie prowadzi do wydzielania się ogromnych — przynajmniej w skali kosmicznej — ilości energii.

promieniowanie warstwy granicznej

Warstwa graniczna, chociaż cienka (jej rozmiary są określone przez promienie torów zakreślanych przez elektrony i pozytony w polu magnetycznym), jest jednak źródłem promieniowania, które moglibyśmy starać się wykryć na drodze obserwacji. Jak wynika z podanego wyżej bilansu energii, ok. połowa energii anihilacji par proton-antyproton jest unoszona przez neutrina, jedna trzecia — przez promieniowanie γ i co najwyżej jedna szósta pozostaje na miejscu w postaci energii kinetycznej elektronów i pozytonów. Elektrony i pozytony poruszające się w polu magnetycznym wysyłają tzw. promieniowanie synchrotronowe, tak więc pewna część ich energii kinetycznej może również opuścić obszar anihilacji w postaci promieniowania radiowego. Przy obecnym stanie techniki obserwacyjnej wykrycie strumienia neutrin pochodzących z przeciętnej warstwy granicznej w naszej Galaktyce jest zupełnie niemożliwe. Podjęte stosunkowo niedawno obserwacje kosmicznego promieniowania γ pozwoliły, co prawda, na wykrycie ogólnego tła galaktycznego, ale są jeszcze zbyt mało dokładne, by dostarczyć jakichkolwiek informacji na temat ewentualnych warstw granicznych w naszej Galaktyce. Jedyne nadzieje można by w tej sytuacji wiązać z obserwacjami radiowymi, jednak z bardziej szczegółowych obliczeń natężenia promieniowania radiowego typowych warstw granicznych wynika, że warstwy te są zbyt słabymi radioźródłami, by mogły być wykryte przez dzisiejsze radioteleskopy. Zresztą samo zaobserwowanie źródła promieniowania synchrotronowego (a źródeł takich znamy dość dużo) nie jest jeszcze dowodem, że promieniowanie to jest związane z anihilacją materii i antymaterii.

skład pierwotnego promieniowania kosmicznego

Przeciwko istnieniu antymaterii w Galaktyce przemawia również fakt, że w pierwotnym promieniowaniu kosmicznym nie udało się wykryć złożonych antyjąder. Domieszka takich antycząstek jak antyprotony, a zwłaszcza pozytony nie stanowi żadnego dowodu istnienia antymaterii, ponieważ są one produktami oddziaływań wysokoenergetycznego promieniowania kosmicznego przechodzącego przez ośrodek międzygwiazdowy. (W celu uniknięcia nieporozumień należy podkreślić, że chodzi tu o promieniowanie kosmiczne w takiej postaci, w jakiej dobiega ono do nas z przestrzeni kosmicznej). Badania promieniowania kosmicznego przyczyniły się w istotny sposób do poznania właściwości antycząstek, jednak obserwowane antycząstki były cząstkami wtórnymi, powstającymi w wyniku zderzeń szybkich cząstek pierwotnego promieniowania kosmicznego z materią ziemską. Wprawdzie nie znamy jeszcze dokładnie natury źródeł promieniowania kosmicznego, jednak wydaje się bardziej prawdopodobne, że źródłami promieniowania kosmicznego złożonego ze zwykłych cząstek są obiekty zbudowane ze zwykłej materii aniżeli obiekty zbudowane z antymaterii.

Jak więc widzimy, obserwacje nie dają nam żadnej zdecydowanej odpowiedzi co do istnienia wielkoskalowych skupisk antymaterii we Wszechświecie i co najwyżej pozwalają przypuszczać, że w obrębie naszej Galaktyki antymateria nie występuje w większych ilościach. Innymi słowy, obserwacje nie zaprzeczają wyraźnie uczynionemu na wstępie założeniu o istnieniu symetrii między materią i antymaterią w skali kosmicznej. Jeżeli chcemy pozostać przy tym założeniu, to musimy znaleźć odpowiedź na dwa zasadnicze pytania: kiedy powstała materia i antymateria, oraz w jaki sposób nastąpiło ich rozdzielenie? Nie trzeba chyba specjalnie podkreślać, że oba problemy są jeszcze dalekie od rozwiązania i można co najwyżej wskazać ogólne kierunki, w których idą rozważania. Rozważania te muszą być oczywiście zgodne z tym co wiemy (lub spodziewamy się, że wiemy) na temat powstania i ewolucji Wszechświata.

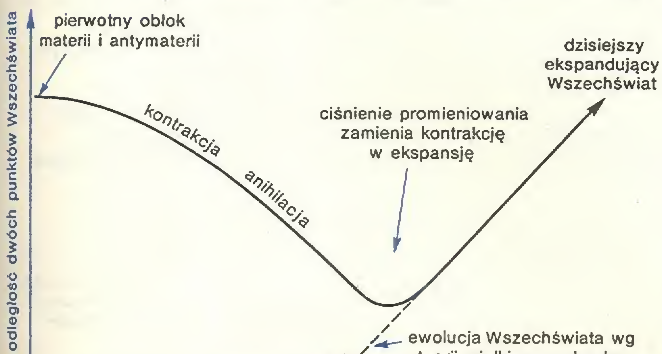
ewolucja Wszechświata z antymaterią

Istniejące dane obserwacyjne świadczą o tym, że kilkanaście miliardów lat temu Wszechświat przechodził przez fazę dużego zgęszczenia materii. W tzw. teorii wielkiego wybuchu (→ Kosmologia) zakłada się nawet, że gęstość materii była wówczas znacznie większa od gęstości materii tworzącej jądra atomowe. Wszechświat o tak dużej gęstości (i odpowiednio wysokiej temperaturze) był niestabilny, wskutek czego zaczął ekspandować. Za takim właśnie obrazem ewolucji Wszechświata przemawia znany od dawna fakt ucieczki galaktyk (opisywanej przez prawo Hubble'a) oraz istnienie wykrytego kilkanaście lat temu promieniowania reliktowego.

W klasycznym modelu wielkiego wybuchu, w którym Wszechświat „pojawił się" w postaci gęstego i gorącego praatomu, trudno jest znaleźć miejsce dla antymaterii: w stanie silnego zgęszczenia materia

i antymateria nie mogłyby ze sobą współistnieć (jeśli pominiemy dość szczególny przypadek, że materia i antymateria pojawiły się od razu rozdzielone). Z punktu widzenia naszych rozważań najwłaściwsze wydaje się założenie, że materia i antymateria istniały początkowo w równych ilościach i były ze sobą dokładnie wymieszane; założenie takie spełnia bowiem w najbardziej naturalny sposób nasz postulat nieodróżnialności materii i antymaterii. Ma ono jednak sens tylko wówczas, jeżeli przyjmiemy równocześnie, że gęstość materii była początkowo bardzo mała (rzędu jednej cząstki w km^3), znacznie mniejsza od gęstości dzisiejszej materii międzygwiazdowej. Wbrew pozorom, założenie takie nie jest sprzeczne z faktami obserwacyjnymi, na których opiera się teoria wielkiego wybuchu. Przyjęty model przesuwa po prostu znany nam początek Wszechświata o wiele miliardów lat wstecz, do czasów, gdy Wszechświat miał postać gigantycznego obłoku rozrzedzonej mieszaniny obu rodzajów materii. Taki stan Wszechświata był również niestabilny. Jeżeli, jak się powszechnie przypuszcza antymateria ma masę grawitacyjną dodatnią, tzn. jeżeli siła grawitacji działająca między cząstkami antymaterii jest siłą przyciągającą, wówczas wzajemne przyciąganie się wszystkich cząstek obłoku powodowało jego kurczenie się. Swobodne kurczenie się obłoku trwało dopóty, dopóki gęstość nie osiągnęła wartości rzędu jednej cząstki w m^3. Przy tej gęstości rozpoczęły się procesy anihilacji prowadzące do pojawienia się w obłoku coraz to silniejszego strumienia promieniowania γ. Promieniowanie to oddziałując z materią powodowało pojawienie się siły pochodzącej od ciśnienia (termicznego ciśnienia gazu oraz ciśnienia promieniowania), skierowanej odwrotnie do kierunku działania sił grawitacji. Postępująca kontrakcja obłoku pociągała za sobą wzrost tempa anihilacji, a tym samym wzrost ciśnienia przeciwstawiającego się kontrakcji. Pozornie pusty i zimny Wszechświat sprzed wielu bilionów lat wkroczył wówczas w dramatyczną fazę ścierania się siły samograwitacji z przeciwnie skierowaną i stale rosnącą siłą ciśnienia promieniowania. W pewnym momencie ciśnienie promieniowania nie tylko zrównoważyło, ale znacznie przewyższyło siłę przyciągania, powodując w rezultacie odwrócenie kierunku ruchu materii. Nasz obłok osiągnął wówczas swoje najmniejsze rozmiary, a następnie zaczął rozszerzać się ze stale wzrastającą prędkością. Można teoretycznie wykazać, że ruch na zewnątrz powinien odbywać się zgodnie z prawem Hubble'a, inaczej mówiąc, dalsza ewolucja naszego

Wszechświat przed wielkim wybuchem



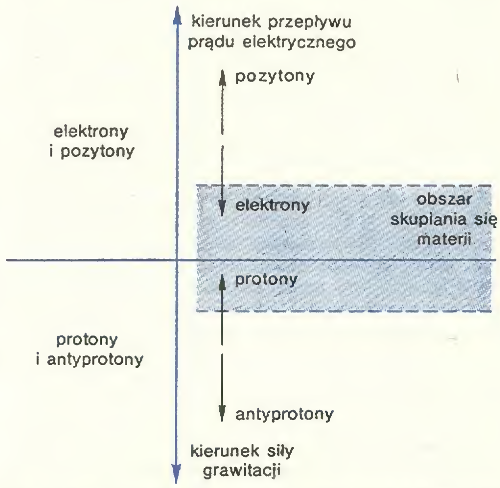
Rys. 3. Schematyczny przebieg historii Wszechświata wg teorii wielkiego wybuchu i w modelu rozrzedzonego obłoku materii i antymaterii. Oba modele przewidują taki sam obecny stan Wszechświata

obłoku powinna wyglądać tak samo, jak ewolucja Wszechświata powstałego w wyniku wielkiego wybuchu (rys. 3).

Obecny Wszechświat nie jest jednak zbudowany z materii rozproszonej, lecz z galaktyk, a więc obiektów o bardzo dużej koncentracji materii. Wprawdzie nie wiemy jeszcze, w którym momencie pojawiły się galaktyki, jednak musimy założyć, że zanim one powstały, nastąpiło rozdzielenie materii od antymaterii, co najmniej w skali odpowiadającej masom gwiazd. Problem wielkoskalowego rozdzielania się materii i antymaterii nie został jeszcze wystarczająco ściśle rozwiązany, jednak mimo to wydaje się celowe przytoczenie w tym miejscu jako ilustracji najprostszego modelu zjawisk, które do takiego rozdzielenia się mogły doprowadzić.

rozdzielanie materii i antymaterii

Rozważmy w tym celu, znajdujący się w polu grawitacyjnym, obłok plazmy składającej się z mieszaniny lekkiego gazu elektronowo-pozytonowego i znacznie cięższego gazu protonowo-antyprotonowego (rys. 4). Pole grawitacyjne doprowadzi do wyraźnej separacji cząstek cięższych od cząstek lżejszych: protony i antyprotony znajdą się przede wszystkim „na dole", podczas gdy obszar zajmowany przez cząstki lżejsze będzie rozciągał się znacznie wyżej — „na



Rys. 4. Jeden z możliwych sposobów rozdzielenia materii i antymaterii

górze" (oczywiście słowa „dół" i „góra" mają tu tylko znaczenie umowne). Nietrudno zrozumieć, że prąd elektryczny płynący ku górze spowoduje ruch elektronów w dół, a protonów — do góry. Na pośredniej wysokości, na której protony spotykają się z elektronami utworzy się obszar zajęty przede wszystkim przez zwykłą materię; obdarzone przeciwnymi ładunkami antycząstki popłyną w kierunkach przeciwnych, tzn. pozytony ku górze, a antyprotony w dół. Oczywiście zmiana kierunku przepływu prądu doprowadziłaby do wydzielenia „czystej" antymaterii. Istnienie pola grawitacyjnego oraz prądu elektrycznego wydaje się założeniem zupełnie naturalnym: pole grawitacyjne może pochodzić z ogólnego pola grawitacyjnego Wszechświata, natomiast przepływ prądu zawsze towarzyszy polom magnetycznym, które we Wszechświecie występują bardzo powszechnie z różnym natężeniem i w różnej skali.

Opisany model jest tylko jakościowym przykładem procesów, które mogły prowadzić do pożądanego rozdzielenia materii i antymaterii, nie stanowi jednak dowodu, że rozdzielenie takie było możliwe w skali rzeczywistych obiektów Wszechświata (galaktyk, gwiazd).

Jak wynika z tego, co zostało powiedziane wyżej, problem istnienia antymaterii we Wszechświecie, choć niezwykle ważny z punktu widzenia naszych ogólnych poglądów na budowę Wszechświata i wypełniającej go materii, nie wyszedł jeszcze poza granice dość ogólnikowych spekulacji. Jest też mało prawdopodobne, by nawet istotne rozszerzenie na-

szych możliwości obserwacyjnych mogło doprowadzić do zmiany sytuacji, która wydaje się szczególnie paradoksalna: pełna symetria własności materii i antymaterii, będąca wyrazem prostoty otaczającego nas świata, jest zarazem okolicznością poważnie utrudniającą zebranie dowodów na to, że Wszechświat jest zbudowany w sposób możliwie najprostszy.

H. Alvén *Kosmologia i antymateria*, Warszawa 1973; J. Nowożyłow *Cząstki elementarne*, Warszawa 1961; *Mezony, grawitacja, antymateria*, Warszawa 1962.

Radioastronomia

Stanisław Zięba

Radioastronomia jest nauką badającą ciała niebieskie przez odbiór wysyłanych przez nie fal radiowych. Należy do grupy tych dziedzin wiedzy, których właściwy rozwój nastąpił dopiero w ciągu ostatnich trzydziestu lat.

początki radioastronomii

Faktycznie jednak narodziny radioastronomii nastąpiły wcześniej, bo w roku 1932, kiedy to amerykański inżynier, Karl Jansky, po raz pierwszy zarejestrował fale radiowe pochodzące z przestrzeni pozaziemskiej. Odkrycie Jansky'ego było przypadkowe i do czasu zakończenia II wojny światowej radioastronomia pozostawała nauką zapomnianą. Pewne pionierskie prace w tej dziedzinie prowadzone były w tym czasie tylko w Stanach Zjednoczonych przez radioamatora Grote Rebera, który z własnych funduszów konstruował i budował specjalne układy do odbioru pozaziemskiego promieniowania radiowego. Wyniki swoich prac opublikował on w latach 1940 i 1942, a więc wtedy, gdy nie było sprzyjającej atmosfery do rozwijania czystej nauki. Mimo to prace te nie pozostały niezauważone i wraz z interesującymi obserwacjami dokonanymi przy okazji badań związanych z radarem stały się podstawą i początkiem współczesnej radioastronomii:

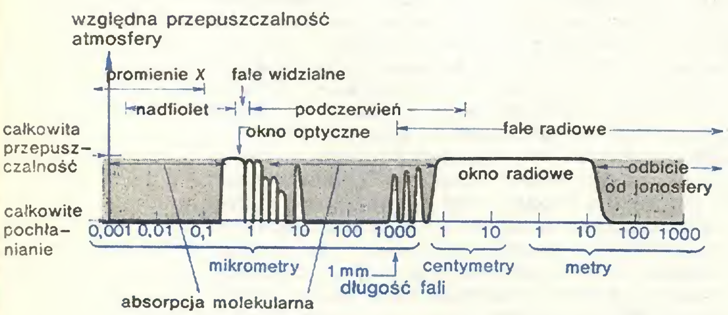
Radioteleskopy

Promieniowanie radiowe dochodzące z kosmosu, podobnie jak promieniowanie widzialne, dociera bez przeszkód przez atmosferę do powierzchni naszego globu. Atmosfera ziemska przepuszcza bowiem z padającego na nią widma fal elektromagnetycznych fale w dwóch zakresach częstości, tworząc jak gdyby dwa okna, przez które możemy spoglądać we Wszechświat (rys. 1). Rozmiary interesującego nas okna radiowego są ograniczone od strony fal milimetrowych pochłanianiem przez tlen atmosferyczny i parę wodną, od strony fal dekametrowych zaś przepuszczalnością jonosfery, która odbija fale dłuższe. Aby zarejestrować dochodzące do nas promieniowanie radiowe, musimy stosować skomplikowane układy odbiorcze, zwane radioteleskopami. Schemat najprostszego radioteleskopu przedstawiony jest na rys. 2. Zasadnicze jego części to antena — zbierająca padające na nią fale elektromagnetyczne, oraz odbiornik — wzmacniający i zapisujący sygnał powstający w oświetlaczu anteny. Sygnał ten zależy od wielkości promieniowania odbieranego przez antenę ze wszystkich „widzianych" przez nią kierunków. Jeżeli antena nastawiona jest na obiekt promieniujący radiowo, to z odebranego sygnału można wyznaczyć gęstość strumienia fal radiowych wysyłanych przez obiekt w określonym przedziale częstości i w określonym kącie bryłowym. Im mniejszy jest kąt bryłowy, w którym następuje uśrednienie promieniowania, tym dokładniej można przebadać strukturę źródła, czyli rozkład kierunków silniej i słabiej promieniujących podobszarów. Parametr decydujący o tej własności radioteleskopu nazwano jego zdolnością rozdzielczą, która najczęściej definiowana jest jako najmniejsza odległość kątowa konieczna do odróżnienia dwóch bliskich źródeł promieniowania.

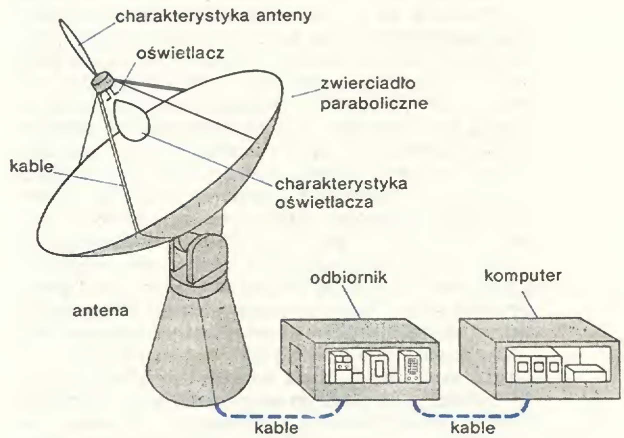
zdolność rozdzielcza radioteleskopu

Drugim, poza zdolnością rozdzielczą, podstawowym parametrem charakteryzującym radioteleskop jest jego czułość, czyli zdolność do odbierania możliwie słabych sygnałów. Duże czułości uzyskuje się budując anteny o dużych aperturach (tj. o dużych powierzchniach zbierających), jak również stosując wzmacniacze o małych szumach własnych, poszerzając pasma odbiorników i wydłużając czas uśredniania

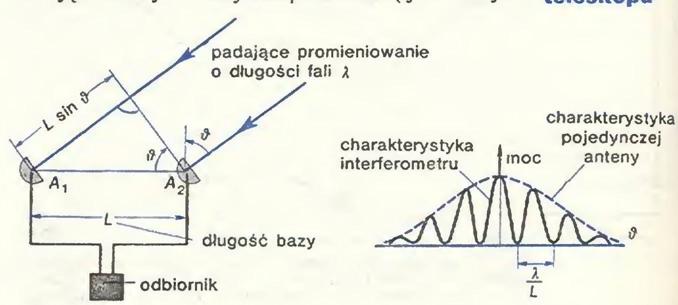
czułość radioteleskopu



Rys. 1. Przechodzenie fal elektromagnetycznych przez ziemską atmosferę (rys. 1, 2, 8 i 13 wg J. D. Kraus *Radio Astronomy*, New York 1966)



Rys. 2. Współczesny radioteleskop

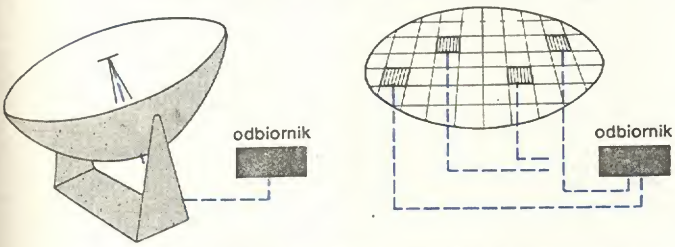


Rys. 3. Układ i charakterystyka prostego interferometru złożonego z dwóch jednakowych anten

(integracji) sygnału. Dzięki złożonym metodom obserwacji i odpowiednim układom radioteleskopów zwanych interferometrami (rys. 3) można znacznie polepszyć zdolność rozdzielczą tworzonych systemów, a więc zmniejszyć rozróżniany już przez układ odstęp kątowy między dwoma źródłami. Zależy on bowiem odwrotnieproporcjonalnie od odległości L (długość bazy) między antenami tworzącymi interferometr. Obecnie w interferometrii o bardzo długich bazach (rzędu tysięcy kilometrów) uzyskuje się zdolności rozdzielcze w granicach 0,1–0,001 sekundy łuku, zależnie od długości fali, na której prowadzone są obserwacje. Najmniejsze wartości otrzymuje się dla fal najkrótszych. Ponieważ przy długich bazach trudno jest połączyć odległe anteny w jeden system odbiorczy, obserwacje wybranego obiektu prowadzi się niezależnie, ale jednocześnie, w obu miejscach, rejestrując odbierane sygnały na taśmach magnetycznych, na których dodatkowo zapisuje się czas z dokładnie zsynchronizowanych zegarów atomowych. Po zakończeniu obserwacji otrzymane taśmy umieszcza się w maszynie cyfrowej, która dokonuje wspólnego opracowania obserwacji obu pozornie niezależnych układów.

metoda syntezy apertury

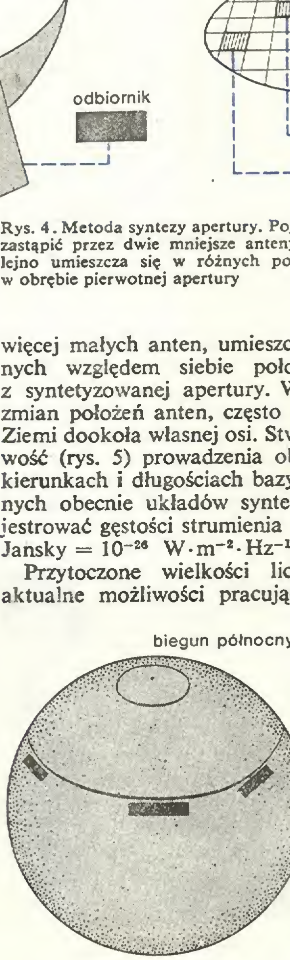
Inną metodą, wykorzystującą współczesną technikę obliczeniową, jest metoda syntezy apertury. Jej zasada została opracowana przez radioastronomów brytyjskich, Sir Martina Ryle'a i Anthona Hewisha, laureatów nagrody Nobla w dziedzinie fizyki z 1974 r. W metodzie tej (rys. 4) radiowy obraz nieba, jaki można byłoby otrzymać za pomocą bardzo dużej, a więc niemożliwej do zbudowania anteny, uzyskuje się zapisując w maszynie cyfrowej obserwacje ustalonego obszaru nieba, wykonane przez układ dwu lub



Rys. 4. Metoda syntezy apertury. Pojedynczą dużą antenę można zastąpić przez dwie mniejsze anteny (interferometr), które kolejno umieszcza się w różnych położeniach względem siebie, w obrębie pierwotnej apertury

więcej małych anten, umieszczanych kolejno w różnych względem siebie położeniach, wybieranych z syntetyzowanej apertury. W metodzie tej, oprócz zmian położeń anten, często wykorzystuje się obrót Ziemi dookoła własnej osi. Stwarza to bowiem możliwość (rys. 5) prowadzenia obserwacji przy różnych kierunkach i długościach bazy. Za pomocą stosowanych obecnie układów syntezy apertury można rejestrować gęstości strumienia rzędu 10^{-3} Jy (1 Jy = 1 Jansky = 10^{-26} $W \cdot m^{-2} \cdot Hz^{-1}$).

Przytoczone wielkości liczbowe charakteryzują aktualne możliwości pracujących układów odbiorczych. Liczby te nie powinny w najbliższej przyszłości ulec zasadniczym zmianom, gdyż głównym celem w konstruowanych obecnie układach jest zwiększenie efektywności ich pracy, poprzez zmniejszenie czasu potrzebnego na wykonanie pełnych obserwacji.



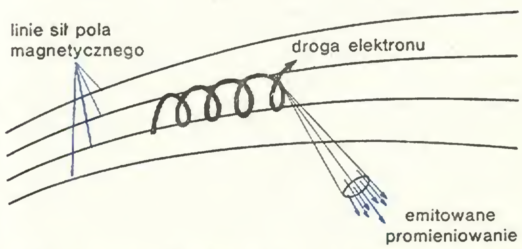
Rys. 5. Zmiany położenia prostego interferometru na powierzchni Ziemi w ciągu dnia oglądane z ustalonego punktu w przestrzeni

Mechanizmy promieniowania

Zanim przejdziemy do omawiania wyników obserwacji radioastronomicznych, zajmiemy się mechanizmami promieniowania prowadzącymi do emisji fal radiowych. Obecnie przypuszcza się, że generacja fal radiowych w kosmosie powodowana jest przez promieniowanie synchrotronowe, promieniowanie termiczne, drgania plazmy oraz przez radiowe promieniowanie monochromatyczne. Jeżeli w prowadzonych badaniach udaje się zidentyfikować mechanizm wywołujący odbierane sygnały radiowe, to istnieje wtedy możliwość poznania szczegółów związanych z naturą fizyczną emitującego obszaru. Różne mechanizmy promieniowania są wytwarzane bowiem w różnych warunkach fizycznych. I tak np. trzy ostatnie z wyżej wymienionych nie wymagają obecności pola magnetycznego, które jest konieczne jeżeli generacja fal ma być wywołana przez promieniowanie synchrotronowe. Właściwie jedyna, jak na razie, droga prowadząca do określenia mechanizmu wytwarzania fal sprowadza się do badania widma radiowego promieniującego obszaru, czyli do wyznaczenia zależności gęstości strumienia od częstości. Badania polaryzacji fal radiowych byłyby tutaj bardzo przydatne, jednakże z powodu dużych trudności, występujących przy tego typu pomiarach, nie są one jeszcze stosowane szerzej w radioastronomii.

promieniowanie synchrotronowe

Źródłem promieniowania synchrotronowego są naładowane cząstki poruszające się w polu magnetycznym. Jeżeli energia tych cząstek jest nierelatywistyczna, promieniowanie to nazywamy promieniowaniem cyklotronowym, pozostawiając nazwę promieniowanie synchrotronowe dla cząstek ultrarelatywistycznych, których energia jest znacznie większa niż m_0c^2. Pojedynczy relatywistyczny elektron krążący w polu magnetycznym B promieniuje fale elektromagnetyczne w małym stożku o osi skierowanej wzdłuż kierunku chwilowej prędkości (rys. 6). Promieniowa-



Rys. 6. Mechanizm synchrotronowy. Elektron poruszając się w polu magnetycznym po linii śrubowej wysyła fale elektromagnetyczne

nie to jest silnie spolaryzowane, a wektor elektryczny fali leży w płaszczyźnie orbity elektronu. Gdy w polu magnetycznym znajdują się elektrony o izotropowym rozkładzie prędkości i rozkładzie energii elektronów opisanym prawem potęgowym $N(E) \sim E^{-\gamma}$, gdzie γ — wskaźnik energetyczny cząstek, to natężenie promieniowania synchrotronowego wyraża się wzorem $I(\nu) \sim \nu^{-\alpha}$. Wskaźnik widmowy α jest powiązany z wskaźnikiem γ zależnością $\alpha = (\gamma - 1)/2$. Dla elektronów kosmicznych $\gamma \approx 2{,}5$. Jeżeli przyjmie się, że elektrony odpowiedzialne za promieniowanie synchrotronowe radioźródeł mają taki sam rozkład energetyczny, to $I(\nu) \sim \nu^{-0,75}$. Wartość $\alpha = 0{,}75$ jest zgodna z obserwowanym średnim wskaźnikiem widmowym radioźródeł (zob. też il. 211, tabl. 57).

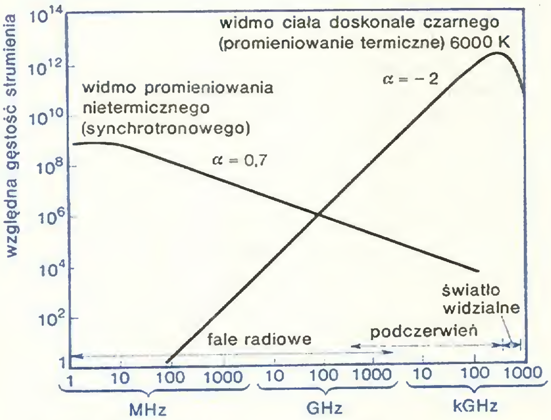
promieniowanie termiczne

Wszystkie ciała, których temperatura jest większa od zera bezwzględnego, emitują fale elektromagnetyczne o różnych długościach. Wykorzystując równanie promieniowania ciała doskonale czarnego można ocenić, ile energii wypromieniuje ciało w poszczególnych przedziałach częstości. Dla fal radiowych uzyskuje się dostatecznie dokładną zależność natężenia od częstości stosując przybliżenie Rayleigha-Jeansa, które prowadzi do wzoru

$$I(\nu) = 2\nu^2 k \cdot T_b / c^2,$$

gdzie k — stała Boltzmanna, a T_b — tzw. temperatura jasnościowa, czyli temperatura jaką posiadałoby ciało doskonale czarne wysyłające w przedziale fal radiowych energię równą energii wypromieniowywanej przez dane ciało.

Jak widzimy na rys. 7 widmo promieniowania synchrotronowego różni się w sposób zasadniczy od widma promieniowania termicznego, którego wskaźnik



Rys. 7. Widmo promieniowania termicznego i synchrotronowego

widmowy α jest równy -2. Na tej podstawie odróżniamy od siebie radioźródła galaktyczne, a więc promieniujące termicznie obszary zjonizowanego wodoru od radioźródeł pozagalaktycznych.

drgania plazmy

Sporadyczne promieniowanie radiowe w stosunkowo wąskim pasmie częstotliwości, ok. $\nu = (e^2 N_e/m)^{1/2}$, może powstawać w wyniku występowania fluktuacji gęstości elektronowej N_e w plazmie. Proces taki jest możliwy przy wzbudzaniu niejednorodnej plazmy strumieniem cząstek poruszających się z prędkościami przewyższającymi prędkość dźwięku. Promieniowanie radiowe tego rodzaju generowane jest podczas niektórych wybuchów radiowych obserwowanych w koronie słonecznej, którą można uważać za plazmę o temperaturze ok. dwóch milionów stopni i gęstości 10^8 elektronów/cm^3. Ponieważ materia w kosmosie najczęściej występuje w stanie plazmy, promieniowanie to może być wytwarzane w wielu obiektach astronomicznych. Teoria tych procesów na razie nie jest jednak dokładnie ilościowo opracowana.

radiowe promieniowanie monochromatyczne

Radiowe promieniowanie monochromatyczne powstaje przy przejściach energetycznych między bliskimi poziomami energetycznymi atomów i cząsteczek. Jest ono scharakteryzowane długością fali oraz jej profilem. Spośród kilkudziesięciu odkrytych już w kosmosie linii radiowych najbardziej rozpowszechniona jest linia neutralnego wodoru o długości fali 21,1 cm, emitowana przy przejściach między podpoziomami struktury nadsubtelnej stanu podstawowego. Na możliwość występowania takiej linii oraz na jej znaczenie w badaniach rozmieszczenia chmur chłodnego wodoru w przestrzeni kosmicznej zwrócił uwagę w 1944 r. holenderski uczony van de Hulst. Przewidywaną przez niego linię po raz pierwszy zaobserwowano w 1951 r. Obecnie obserwuje się już wiele różnego rodzaju linii radiowych. Dużą grupę wśród nich stanowią linie rekombinacyjne, związane z przejściami pomiędzy sąsiednimi poziomami o dużych głównych liczbach kwantowych oraz linie molekularne prostych związków organicznych, których mechanizmy powstawania nie są na razie dostatecznie zbadane. Niektóre z obserwowanych linii molekularnych przedstawia poniższa tabela.

obserwowane linie molekularne

Rok odkrycia	Nazwa atomu lub cząsteczki	Symbol	Długość fali
1951	wodór (neutralny)	HI	21,1 cm
1963	grupa hydroksylowa	OH	18,0 cm
1964	wodór (zjonizowany)	HII	3,4 cm
1966	hel (zjonizowany)	He	18,0 cm
1967	węgiel (zjonizowany)	C	6,0 cm
1968	amoniak	NH_3	1,3 cm
1968	woda	H_2O	1,3 cm
1969	formaldehyd	H_2CO	6,2 cm
1970	tlenek węgla	CO	2,6 mm
1970	cyjanowodór	HCN	3,4 mm
1970	alkohol metylowy	CH_3OH	36,0 cm
1970	kwas mrówkowy	HCOOH	18,0 cm
1971	siarczek węgla	CS	2,0 mm
1971	formamid	$HCONH_2$	6,5 cm
1971	tlenek krzemu	SiO	2,7 mm
1971	cyjanek metylu	CH_3CN	2,7 mm
1971	metyloacetylen	CH_3CCH	3,5 mm
1972	siarkowodór	H_2S	1,8 mm
1973	tlenek siarki	SO	2,9 mm

Źródła promieniowania radiowego

Radiowy obraz nieba, jaki możemy sobie wytworzyć na podstawie wykonywanych radioteleskopami obserwacji, różni się całkowicie od obrazu nieba oglądanego w pogodne noce. Obraz ten zmienia się ponadto dość znacznie wraz ze zmianą długości fali, gdyż w różnych obiektach, zależnie od istniejących w nich warunków fizycznych, dominują, w różnych przedziałach częstości, różne mechanizmy promieniowania (il. 218, tabl. 58 oraz rys. 3 i 4, str. 938).



Rys. 8. Radiowa panorama nieba, pokazująca obraz nieba „widziany oczyma" czułymi na fale radiowe o częstości 250 MHz (otrzymana przez zacieniowanie obszarów między kolejnymi izofotami mapy konturowej)

radiowy obraz nieba

Obiektem najbardziej charakterystycznym na radiowym niebie jest Droga Mleczna (rys. 8). Kontury jej są nieostre, lecz nie dlatego, że składa się z wielu oddzielnych gwiazd. Jest ona po prostu dużą chmurą, w której widać kilkanaście jasnych obiektów, znacznie większych od znanych nam dobrze punktów świetlnych. Na próżno można szukać na radiowym niebie charakterystycznych gwiazd z konstelacji Wielkiego Wozu, Oriona czy Kasjopei. Niebo pokryte jest wieloma nowymi obiektami. Po dokładnym porównaniu z atlasem gwiazd można przekonać się, że niektóre z nich znane są z obserwacji wizualnych. Są to takie obiekty, jak np.: Mgławica Krab, Mgławica Andromedy i inne, z których większość, to położone daleko w przestrzeni mgławice pozagalaktyczne, zwane inaczej galaktykami. Słońce, które wygląda jak dziwny przyćmiony obiekt, dookoła którego pojawiają się nieregularne kontury korony słonecznej, jest w zakresie radiowym tak słabym źródłem, że z trudnością „oświetla" Księżyc, który dostrzegamy tylko dzięki jego własnemu promieniowaniu. Planety są również ledwo „widoczne". Wyjątkiem jest Jowisz, który wysyła impulsowe błyski promieniowania radiowego.

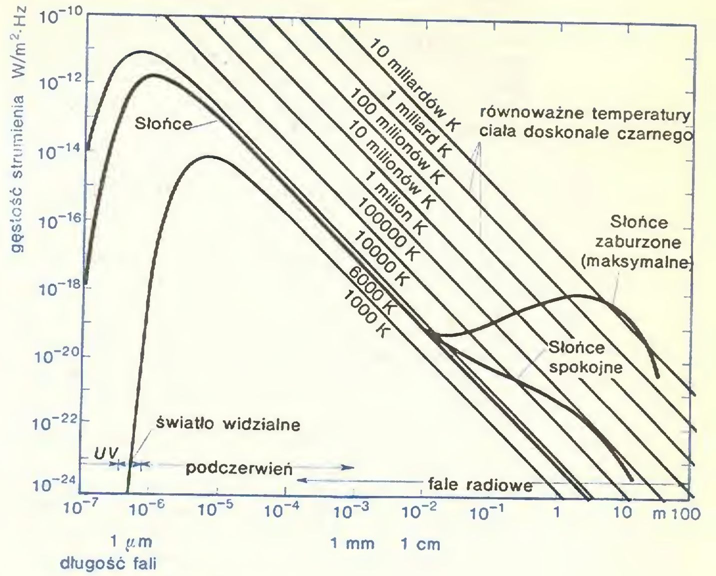
Już z tego krótkiego i uproszczonego opisu nieba radiowego wynika, że w kosmosie występują rozmaite obiekty promieniujące radiowo. Przejdźmy teraz do bardziej szczegółowego opisu promieniowania radiowego tych obiektów oraz do wniosków, jakie można stąd wysunąć co do ich budowy fizycznej.

Promieniowanie radiowe Słońca i gwiazd

Słońce jest najważniejszym spośród ciał niebieskich oświetlających nasze radiowe niebo, chociaż nie dominuje na tym niebie tak wyraźnie, jak Słońce widzialne, a wypromieniowywane przez nie fale radiowe nie wpływają bezpośrednio na bieg naszego codziennego życia. Hipotezy, że Słońce promieniuje również radiowo wysuwane były już w parę lat po odkryciu przez Heinricha Hertza fal radiowych. Wszystkie jednak próby detekcji tych fal kończyły się niepowodzeniami. Dopiero w latach II wojny światowej zarejestrowano za pomocą urządzeń radarowych, w kilku ośrodkach niezależnie od siebie, silne sygnały w postaci szumu, który, jak się okazało, emitowany był z aktywnych obszarów na Słońcu. Wyniki te zostały opublikowane po zakończeniu wojny w 1945 r. Od tego czasu prowadzone są systematyczne badania promieniowania radiowego Słońca. Obecnie uważa się, że na promieniowanie to składa się: stałe promieniowanie termiczne spokojnego Słońca oraz promieniowanie sporadyczne zależne od stanu aktywności Słońca. Promieniowanie sporadyczne można podzielić na tzw. składową wolno zmienną — wykazującą zmiany z dnia na dzień z okresem około dwudziestu siedmiu dni, oraz na wybuchy radiowe — charakteryzujące się dużymi zmianami gęstości strumienia w przedziałach czasu rzędu sekund, minut lub godzin.

promieniowanie termiczne i sporadyczne

Widmo promieniowania Słońca, zarówno w zakresie fal radiowych, jak i optycznych, przedstawia rys. 9. Jak widać, termiczne promieniowanie Słońca spokojnego nie odpowiada jednej określonej temperaturze. Związane to jest z tym, że promieniowanie o różnych długościach fal emitowane jest z różnych warstw atmosfery Słońca. Fale metrowe wysyłane są z gorącej, mającej ponad milion stopni korony, fale centymetrowe zaś z chłodnej, dolnej chromosfery o temperaturze ok. 10 000 stopni. Składowa wolna zmienna jest obserwowana w zakresie fal od 3 do ok. 60 cm, a jej maksimum znajduje się w pobliżu 8 cm. Charakterystyczne zmiany składowej wolno zmiennej zgodne są ze zmianami powierzchni plam na Słońcu. Źródłem tej składowej jest najprawdopodobniej promieniowanie termiczne powstające w jasnych i gęstych kondensacjach koronalnych, o temperaturze ok. dwu milionów stopni, położonych ponad obszarami plam i pochodni. Odmienne mechanizmy powodują rejestrowane wybuchy radiowe, które są najsilniejsze i najbardziej złożone w metrowym za-



Rys. 9. Widmo promieniowania Słońca

kresie długości fal. Są one z reguły stowarzyszone z obserwowanymi w świetle widzialnym rozbłyskami chromosferycznymi. Większość z nich rozpoczyna się w chwili wyrzucenia strumienia naładowanych cząsteczek z aktywnych obszarów związanych z plamami. Mechanizmy powstawania i rozwijania się wybuchów radiowych nie są jeszcze dokładnie poznane. Są to bowiem bardzo złożone procesy, w których powstanie fal radiowych może być spowodowane zarówno promieniowaniem synchrotronowym, jak i różnego rodzaju drganiami plazmy.

Promieniowanie radiowe z innych gwiazd niż Słońce przez długi okres czasu nie było rejestrowane. Przyczyną tego stanu była przede wszystkim zbyt mała czułość pracujących układów. W końcu, po wielu fałszywych alarmach, w 1963 r. udało się zaobserwować krótkotrwały impuls promieniowania radiowego (rozbłysk) dochodzący do nas od czerwonego karła *UV* Ceti. W latach następnych zaobserwowano jeszcze kilkanaście gwałtownych radiowych i optycznych rozbłysków z gwiazd tego typu, które w rezultacie zyskały nazwę gwiazd rozbłyskowych. W miarę upływu czasu wzrastała liczba obiektów gwiazdowych, których promieniowanie radiowe udało się odebrać. I tak w 1970 r. zarejestrowano na falach centymetrowych promieniowanie dochodzące od dwóch gwiazd nowych oraz od trzech znanych układów podwójnych α Scorpii *B*, β Persei (Algol) i β Lyrae. Promieniowanie radiowe tych układów wykazywało w latach 1970–1972 bardzo osobliwe cechy wskazujące na to, że znajdowały się one w tym czasie w okresie silnej aktywności powodującej powstawanie rozbłysków radiowych. Potwierdzeniem takiego przypuszczenia może być fakt, że na 60 badanych w tym czasie układów podwójnych tylko w tych zarejestrowano dostatecznie silną emisję radiową. Wartości wskaźnika widmowego, otrzymywane z obserwacji na różnych długościach fal, wskazywały na termiczną naturę tych rozbłysków. Hipoteza ta wymaga jednak dalszych potwierdzeń, które możemy między innymi uzyskać z obserwacji w pasmie promieni *X* (→ Astronomia promieni *X* i γ). Gdyby bowiem obserwowane rozbłyski faktycznie powstawały w wyniku wzmożonego promieniowania termicznego pewnych obszarów w atmosferach tych gwiazd, to warunki fizyczne panujące w tych obszarach musiałyby również powodo-

gwiazdy rozbłyskowe

rozbłyski radiowe

wać emisję promieniowania *X*. Prawie w tym samym czasie stwierdzono, że niektóre znane źródła promieniowania *X* emitują z kolei silne strumienie fal radiowych w postaci rozbłysków. Szczególnie silne tego typu zjawisko wystąpiło we wrześniu 1972 r. w obiekcie znanym jako Cyg X-3. Reasumując, można stwierdzić, że badania radiowe gwiazd, w których stawiamy powoli pierwsze kroki, przyczynią się do dokładniejszego zrozumienia ewolucji i budowy złożonych układów gwiazd.

Promieniowanie radiowe Księżyca i planet

Planety, podobnie jak Księżyc, „widoczne" są na falach radiowych dzięki własnemu promieniowaniu emitowanemu z nagrzanych powierzchniowych warstw swych globów. Gdyby warstwy te nagrzewały się tylko wskutek promieniowania Słońca, temperatura jasnościowa wyznaczana z obserwacji radiowych nie powinna się zasadniczo różnić od temperatury mierzonej w zakresie fal optycznych i podczerwonych. Taką właśnie zgodność obserwujemy dla Księżyca i większości planet. Dla Wenus otrzymywane temperatury jasnościowe są jednak wyższe, co związane jest z obecnością gęstej atmosfery, która może zatrzymywać wtórne promieniowanie termiczne (efekt szklarniowy). Najbardziej złożone promieniowanie radiowe wysyła Jowisz. Poza promieniowaniem termicznym wyodrębniamy tu składową nietermiczną oraz tzw. błyski dekametrowe składające się z izolowanych grup krótkich impulsów. Błyski te powstają prawdopodobnie wskutek oddziaływania strumieni prędkich naładowanych cząstek z polem magnetycznym Jowisza.

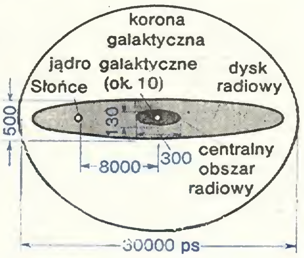
Promieniowanie radiowe Galaktyki

izofoty

Jednym z dużych osiągnięć radioastronomii było zbadanie i zrozumienie struktury naszej Galaktyki. Systematyczne obserwacje Drogi Mlecznej na różnych częstościach doprowadziły do utworzenia dokładnych map linii jednakowych jasności (izofoty). Izofoty dla różnych długości fal różnią się znacznie między sobą zarówno kształtem, jak i wielkością temperatury jasnościowej T_b. W pasmie fal dekametrowych promieniowanie jest izotropowe, a $T_b \approx 100\,000$ K. Przechodząc do fal krótszych T_b maleje (1000–100 K) i wyraźnie spada w miarę oddalania się od równika galaktycznego (rys. 10). Wskaźnik widmowy α dla fal dekametrowych i metrowych wskazuje na promieniowanie synchrotronowe. W przedziale fal krótszych natomiast ogólne promieniowanie Galaktyki ma charakter termiczny. Z przebiegu izofot ogólnego promieniowania radiowego Galaktyki wynika, że składa się ono (rys. 11) z promieniowania sferycznego korony galaktycznej (składowa sferyczna) oraz promieniowania płaszczyzny galaktycznej (składowa płaska — dysk radiowy).

składowa sferyczna i płaska

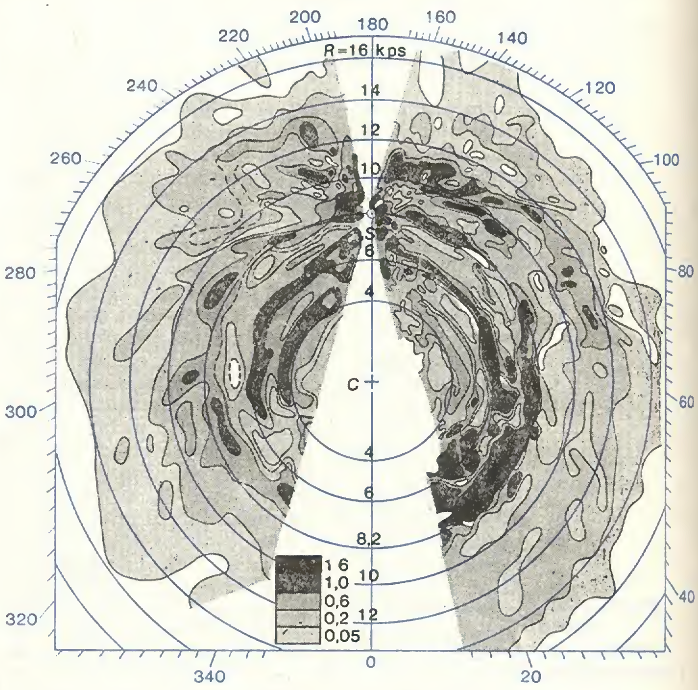
Składowa sferyczna pochodzi z hamowania elektronów relatywistycznych w chaotycznych polach magnetycznych korony galaktycznej. Składowa płaska jest związana z promieniowaniem termicznym cienkiej warstwy obłoków wodoru międzygwiazdowego, zjonizowanego przez gorące gwiazdy koncentrujące się w płaszczyźnie Galaktyki, oraz promieniowaniem elektronów relatywistycznych w uporządkowanych polach magnetycznych ramion spiralnych Galaktyki. Składowa sferyczna oraz synchrotronowa część składowej płaskiej decydują o kształcie izofot dla długofalowej części widma, termiczna część składowej płaskiej zaś dla krótkofalowej części widma.



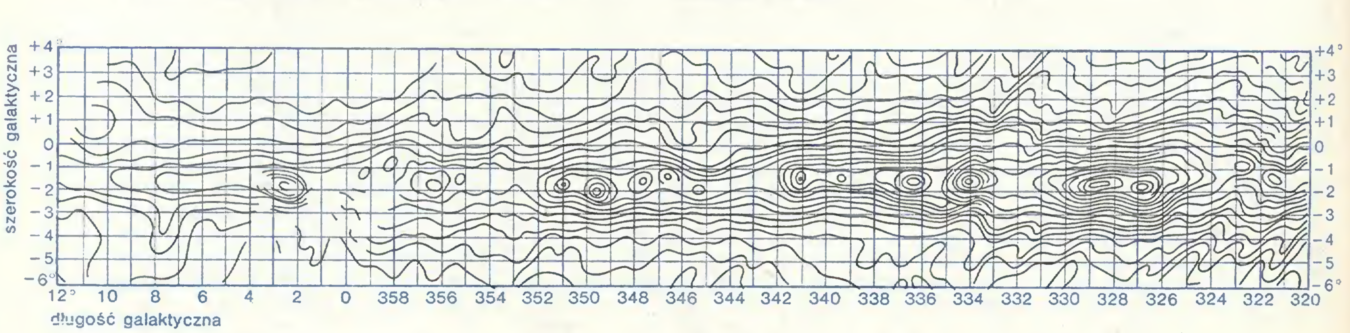
Rys. 11. Budowa Galaktyki według danych radiowych; rozmiary w parsekach

rozkład wodoru w Galaktyce

Badania natężenia promieniowania i profilu linii wodoru neutralnego umożliwiły określenie jego rozkładu w Galaktyce (ramiona spiralne, rys. 12), jego gęstości (ok. 0,1–100 atomów/cm³) i całkowitej masy, temperatury (ok. 100 K) i prędkości radialnych spo-



Rys. 12. Rozkład wodoru neutralnego w Galaktyce (*C* środek Galaktyki, *S* Słońce)



Rys. 10. Izofoty promieniowania radiowego Galaktyki przy $\lambda = 3,5$ m; 1 jednostka wynosi 1000 K

wodowanych obrotem Galaktyki, jak również ruchami własnymi obłoków gazu. Wszystkie przedstawione powyżej wyniki uzyskane zostały w żmudnym procesie opracowywania wyników obserwacji. Wykorzystując teoretyczne zależności wiążące parametry obserwowane z warunkami fizycznymi, starano się określić te ostatnie tak, aby obserwowane natężenia i profile linii były jak najlepiej przez nie odtworzone.

Oprócz termicznych źródeł galaktycznych, jakimi są obłoki zjonizowanego wodoru, obserwuje się w Galaktyce wiele źródeł promieniujących synchrotronowo. Są to między innymi rozszerzające się otoczki będące pozostałościami po wybuchach gwiazd supernowych. Nowy typ galaktycznych źródeł punktowych odkryto w 1967 r. Nazwano je pulsarami. Emisja z pulsarów charakteryzuje się występowaniem silnych krótkotrwałych impulsów o zmiennej amplitudzie pojawiających się regularnie w równych odstępach czasu, których wartości wahają się od 0,01 do kilku sekund w zależności od pulsara. Obecnie najczęściej uważa się, że pulsary są prawdopodobnie szybko rotującymi gwiazdami neutronowymi z silnym polem magnetycznym, na powierzchni których znajdują się obszary będące źródłem impulsów (→ Pulsary).

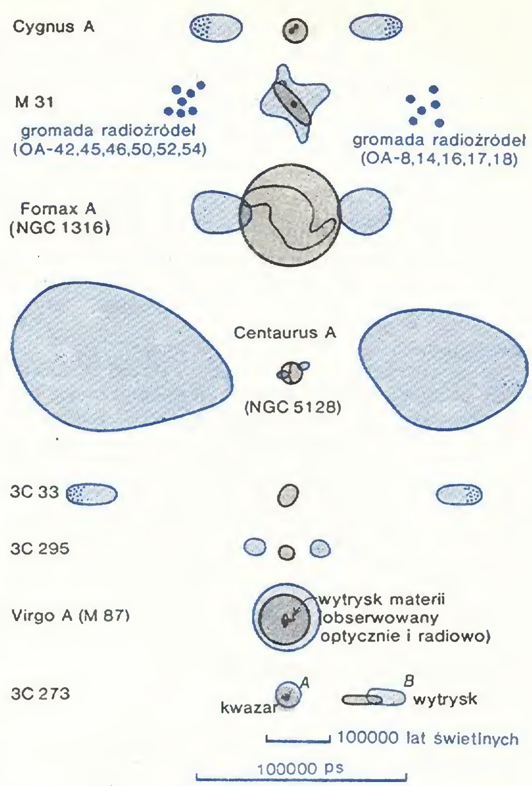
pulsary

Radioźródła pozagalaktyczne

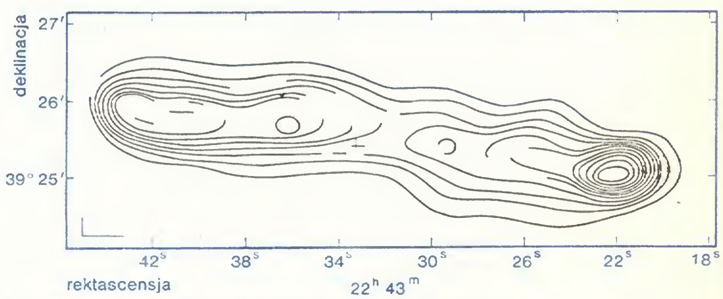
Odbierając promieniowanie radiowe od mniej lub bardziej punktowych źródeł, od początku starano się identyfikować obserwowane obiekty radiowe z obiektami optycznymi. Identyfikację taką można przeprowadzać w zasadzie tylko na podstawie zgodności położeń, co jest zadaniem trudnym a wyniki nie zawsze są całkowicie jednoznaczne, pomimo że zdolność rozdzielcza tworzonych systemów antenowych została w ostatnich latach znacznie polepszona. Pomimo tych trudności już w 1949 r. jedne z silniejszych źródeł, Virgo *A* i Centaurus *A*, zostały zidentyfikowane z bliskimi galaktykami M 87 i NGC 5128. Dopiero jednak identyfikacja bardzo silnego źródła Cygnus *A* z odległą o 600 milionów lat światła galaktyką uzmysłowiła wyraźnie pozagalaktyczną naturę wielu źródeł i postawiła jasno problem wyjaśnienia ich ogromnych źródeł energii. Na początku lat 60-ych niektóre radioźródła zostały zidentyfikowane z punktowymi, gwiazdowymi z wyglądu obiektami optycznymi odznaczającymi się bardzo dużymi przesunięciami swoich linii emisyjnych ku czerwieni. Obiekty te nazwano kwazarami, a ich dalsze badania wykazały, że wiele spośród nich jest silnymi radioźródłami (→ Kwazary).

kwazary

W wyniku ogromnej, wieloletniej pracy różnych ośrodków radioastronomicznych znamy obecnie ponad 10 000 radioźródeł, z których zaledwie kilkaset jest w sposób pewny zidentyfikowanych z galaktykami lub kwazarami. Wprawdzie dokładne pozycje radiowe znane są dla znacznie większej liczby radioźródeł, jednakże nie znajduje się dla nich obiektów optycznych na płytach atlasu Palomarskiego. Dokładne badania radioźródeł za pomocą interferometrów są podstawą analizy ich struktury, która z reguły jest dość złożona. Konfiguracją najczęściej występującą jest układ dwu obszarów położonych symetrycznie względem obiektu optycznego (rys. 13) emitujących fale radiowe. To symetryczne położenie oraz silniej lub słabiej zaznaczone pomosty promieniowania radiowego, łączące obszary wzmożonej emisji radiowej z centralnie położonym obiektem optycznym, wskazują na ich wspólne pochodzenie (rys. 14). W niektórych przypadkach obserwujemy również bardzo zwarte składniki struktur. Często też wśród zwartych źródeł można zauważyć silne zmiany wysyłanych przez nie strumieni mające jakby charakter burz pojawiających się najpierw na falach krótkich. Obserwując radioźródła na różnych częstościach można badać ich widma, które mimo różnic wskazują na



Rys. 13. Radiowa (kolor niebieski) i optyczna (kolor czarny, szary) struktura różnych radioźródeł pozagalaktycznych w przybliżeniu w tej samej skali



Rys. 14. Struktura radioźródła 3C 452 obserwowana na częstości 1407 MHz. Krzyżyk oznacza obiekt optyczny — galaktykę 16^m

synchrotronową emisję fal radiowych. Mechanizm ten jest więc ogólnie przyjęty, pozostaje jednak nadal nie rozwiązany problem źródeł energii dostarczających relatywistycznych elektronów. Wysuwane są różne koncepcje, jak np. zderzenie galaktyk, ich eksplozje czy też ich grawitacyjne zapadanie się. Wymagają one jednak jeszcze wielu opracowań, tak że nie należy się spodziewać szybkiego rozwiązania tego problemu.

J. D. Kraus *Radio Astronomy*, New York 1967; G. L. Verschuur, K. I. Kellermann *Galactic and Extra-Galactic Radio Astronomy*, New York 1974.

Astronomia promieni X i γ

Marcin Kubiak

Przedmiotem zainteresowania tej nowej i szybko rozwijającej się gałęzi astronomii jest pozaziemskie promieniowanie elektromagnetyczne o długości fali mniejszej od ok. 10 nm. Promieniowanie X i γ wygodnie jest charakteryzować za pomocą energii fotonów wyrażonej w kiloelektronowoltach (1 keV = = 1000 eV = $1{,}602 \cdot 10^{-16}$ J). Długość fali λ wyrażona w nanometrach jest związana z energią fotonu wyrażoną w keV za pomocą związku λ(nm) = 124/ /E(keV). Umowną granicą między promieniowaniem X i γ jest energia 0,51 MeV (1 MeV = 1000 keV) odpowiadająca energii spoczynkowej elektronu. Fizyczną podstawą rozróżnienia obu rodzajów fotonów

pochodzenie fotonów X i γ

jest ich pochodzenie: fotony X powstają przede wszystkim w procesach atomowych, natomiast źródłem fotonów γ są oddziaływania jądrowe.

Fotony o energii z omawianego zakresu są całkowicie pochłaniane w atmosferze ziemskiej i mogą być obserwowane w sposób bezpośredni tylko z pokładów wysoko wznoszących się balonów, rakiet i sztucznych satelitów Ziemi. Wprawdzie fotony o energiach większych od ok. 10^{11} eV można obserwować z powierzchni Ziemi za pośrednictwem wielkich pęków atmosferycznych (→ Promieniowanie kosmiczne), jednak zasadniczo możliwości obserwacyjne astronomii X i γ są określone przez stan techniki rakietowej i satelitarnej.

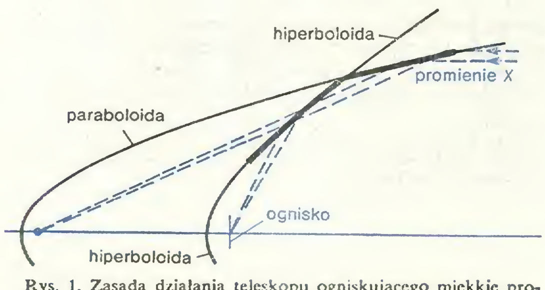
Metody obserwacji

Technika odbioru promieniowania X i γ zależy od zakresu energii, w którym prowadzone są obserwacje. Granice poszczególnych zakresów są określone przez właściwości detektorów.

0,1 keV–10 keV

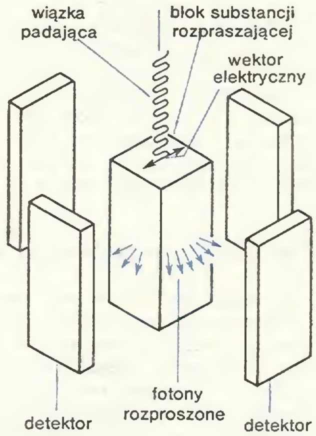
0,1 keV–10 keV (miękkie promieniowanie X). Obserwacje w tym zakresie prowadzi się w zasadzie podobnie jak w dalekim nadfiolecie. Detektorami promieniowania są rentgenowskie emulsje fotograficzne, elektronowe przetworniki obrazów oraz liczniki proporcjonalne. W zakresie tym możliwe jest również ogniskowanie promieniowania i tworzenie obrazów źródeł dzięki wykorzystaniu posiadanych przez niektóre powierzchnie metaliczne (np. beryl, glin, nikiel, złoto) właściwości odbijania promieniowania X padającego pod bardzo małymi kątami. Zasadę działania teleskopu ogniskującego miękkie promieniowanie X wyjaśnia rys. 1. Fotografię takiego teleskopu przedstawia il. 212 (tabl. 57). Obecnie kątowa zdolność rozdzielcza teleskopów tego rodzaju jest rzędu 2″, choć w niedalekiej przyszłości można spodziewać się jej zwiększania aż do 0,3″, tj. do granicy określonej przez geometryczną aberrację układu.

W omawianym zakresie promieniowania możliwe jest również prowadzenie obserwacji spektroskopowych; jako elementy dyspersyjne służą metalizowane siatki dyfrakcyjne lub kryształy. Stosunkowo mała powierzchnia czynna teleskopów rentgenowskich sprawia, że jak dotychczas obserwacje takie zostały wykonane z powodzeniem tylko w przypadku Słońca.



Rys. 1. Zasada działania teleskopu ogniskującego miękkie promieniowanie X. Użycie dwu zwierciadeł pozwala na skrócenie ogniskowej układu

Duże znaczenie dla astrofizyki ma możliwość pomiaru polaryzacji promieniowania X w tym zakresie. Wykorzystuje się do tego celu zjawisko nieizotropowego rozpraszania spolaryzowanej wiązki fotonów X w blokach substancji rozpraszających, np. LiH (rys. 2).



Rys. 2. Zasada działania polarymetru rentgenowskiego. Kąt rozproszenia zależy od kierunku polaryzacji promieniowania padającego

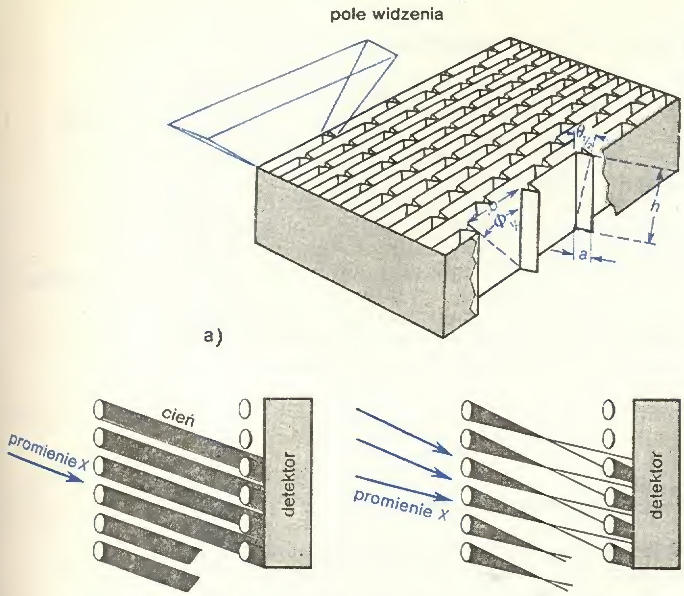
1 keV–20 keV

1 keV–20 keV. Detektorami fotonów X o energii z tego zakresu są liczniki proporcjonalne. Fotony wbiegające do takiego licznika wywołują lawinowe wyładowanie elektryczne w gazie wypełniającym przestrzeń między anodą i katodą. Ponieważ wielkość wyładowania jest proporcjonalna do energii padającego fotonu, połączenie licznika z analizatorem wysokości impulsów prądu umożliwia określenie energii rejestrowanego fotonu (spektroskopia niedyspersyjna). Zakres pomiarowy urządzenia ustala się z jednej strony przez zastosowanie odpowiedniego okna wejściowego, absorbującego wszystkie fotony o energii mniejszej od wybranej energii progowej, a z drugiej strony — przez właściwości gazu wypełniającego obszar między elektrodami: fotony o energiach bardzo dużych mają zbyt mały przekrój czynny na oddziaływanie z materią, by wywołać jonizację w danej ilości gazu. Ograniczenie pola widzenia urządzenia odbiorczego osiąga się przez zastosowanie odpowiednich kolimatorów mechanicznych umieszczonych przed oknem wejściowym (rys. 3).

10 keV–ok. 300 keV

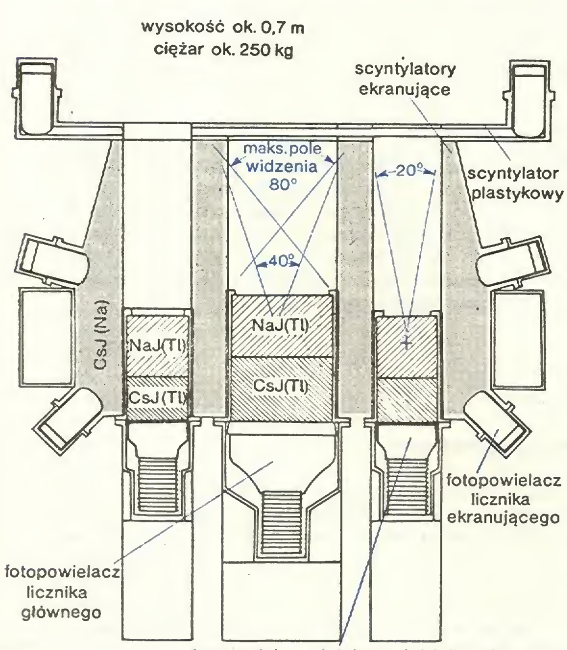
10 keV–ok. 300 keV (twarde promieniowanie X). Detektorami promieniowania w tym zakresie energii są liczniki scyntylacyjne. Zasadniczą częścią licznika jest kryształ (o możliwie dużych rozmiarach), w którym fotony X wywołują błyski luminescencyjne. Kryształ taki musi zawierać centra luminescencji wychwytujące elektrony swobodne uwalniane z atomów zjonizowanych w wyniku absorpcji lub rozpraszania komptonowskiego fotonów X; wychwytowi towarzyszy emisja fotonu światła widzialnego. Najczęściej używa się kryształów NaJ(Tl) — aktywowanych talem kryształów jodku sodu — wysyłających promieniowanie widzialne w zakresie długości fali od 420 do 435 nm. Jedna powierzchnia kryształu pozostaje w kontakcie optycznym z fotopowielaczem rejestrującym poszczególne błyski światła widzialnego, podczas gdy pozostałe powierzchnie są zazwyczaj pokryte substancjami odblaskowymi w celu zwiększenia ilości światła zbieranego przez fotopowielacz. Ilość światła zawarta w błysku jest tylko w przybliżeniu proporcjonalna do energii kwantu padającego, w związku z czym energetyczna zdolność rozdzielcza liczników

scyntylacyjnych jest znacznie mniejsza niż liczników proporcjonalnych i wyraźnie pogarsza się ze wzrostem energii kwantów. Dlatego też liczniki scyntylacyjne są użyteczne tylko do energii kilkuset keV, chociaż w zasadzie reagują one również na przelot kwantów o znacznie większych energiach.



Rys. 3. Przykłady kolimatorów mechanicznych: a) kolimator płytkowy, b) obrotowy kolimator siatkowy

Obserwacje w tym zakresie energii są skomplikowane, ponieważ istnieje niepożądane tło kwantów X i γ powstających w materialnym otoczeniu licznika (obudowa, satelita itp.) pod wpływem szybkich cząstek promieniowania kosmicznego. W celu wyeliminowania tego tła, a jednocześnie ograniczenia pola widzenia układu odbiorczego, licznik scyntylacyjny otacza się dodatkowymi licznikami ekranującymi, tworzącymi z licznikiem głównym układ antykoincydencyjny. Zasadę działania takiego układu wyjaśnia rys. 4.



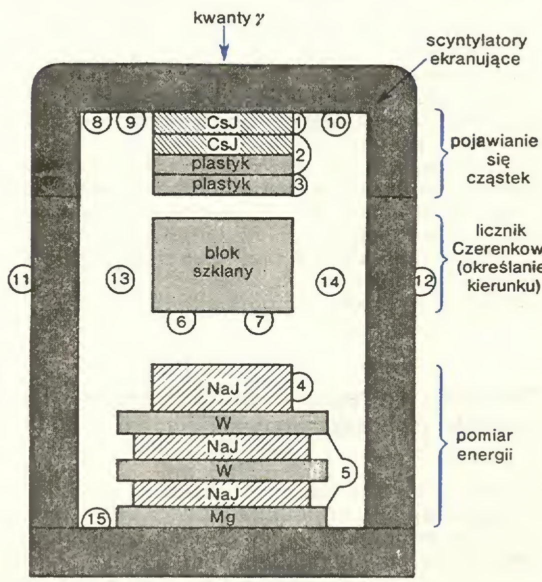
Rys. 4. Schemat ekranującego licznika scyntylacyjnego. Licznikiem zasadniczym jest kryształ NaJ(Tl), scyntylatory ekranujące są zbudowane z kryształów CsJ(Tl). Zarejestrowany zostanie tylko sygnał pochodzący od licznika głównego w nieobecności sygnałów z liczników ekranujących i scyntylatora plastykowego

0,3 MeV–10 MeV

0,3 MeV–10 MeV (miękkie promieniowanie γ). Ponieważ kryształy scyntylacyjne są mało wydajnymi detektorami kwantów o tak dużych energiach, natomiast są dobrymi detektorami cząstek naładowanych elektrycznie, używane obecnie teleskopy γ są zaopatrzone w układ, który zamienia wbiegający kwant γ w mniej lub bardziej liczną lawinę cząstek wtórnych o mniejszych energiach. W najprostszy sposób można to osiągnąć umieszczając na drodze kwantu np. płytę ołowianą o odpowiedniej grubości. Cząstki wtórne wybite z płyty ołowianej są następnie rejestrowane przez licznik lub liczniki scyntylacyjne. Łączna energia zawarta w błyskach dawanych przez cząstki wtórne jest proporcjonalna do energii kwantu pierwotnego.

10 MeV–kilkaset MeV

10 MeV–kilkaset MeV. Kwanty o tak dużych energiach są źródłem cząstek wtórnych poruszających się z prędkościami bliskimi prędkości światła, co umożliwia zwiększenie dokładności określenia kierunku przylotu poszczególnych kwantów γ bez jednoczesnego zawężania całkowitego pola widzenia teleskopu. Jest to o tyle ważne, że strumień kwantów o tych energiach jest już niezwykle mały i teleskop o małym polu widzenia rejestrowałby kwanty tak rzadko, że wykorzystywanie do takich obserwacji kosztownych rakiet i satelitów byłoby pozbawione celu. Zjawiskiem, które w tym wypadku wykorzystujemy, jest promieniowanie Czerenkowa wysyłane przez szybkie cząstki wtórne w bloku przezroczystej substancji o odpowiednio dużym współczynniku załamania (prędkość światła w takim ośrodku jest istotnie mniejsza niż w próżni). Promieniowanie Czerenkowa pojawia się wówczas, gdy przez dany ośrodek przebiega cząstka z prędkością większą od prędkości światła w tym ośrodku i jest skierowane niemal dokładnie wzdłuż drogi cząstki. Oprócz licznika Czerenkowa, pozwalającego określić kierunek biegu cząstek, a tym samym również kierunek wlotu pierwotnego kwantu γ, na



Rys. 5. Licznik Czerenkowa. Liczby w półkolach oznaczają fotopowielacze. Zarejestrowany zostanie tylko taki przypadek, w którym nie będzie sygnału z ekranu (fotopowielacze *8–15*), natomiast nadejdzie sygnał z któregoś z fotopowielaczy *1*, *2*, *3*, sygnał z co najmniej jednego fotopowielacza *7* i *8* oraz sygnał z fotopowielacza *4* lub *4* i *5*

drodze cząstek wtórnych umieszcza się również układ wielu scyntylatorów lub komór iskrowych, śledzących ruch cząstek wtórnych aż do ich całkowitego wyhamowania. Długość drogi hamowania cząstek wtórnych jest miarą ich energii początkowej, a więc rów-

nież miarą energii pierwotnego kwantu γ. Schemat teleskopu opisanego typu jest przedstawiony na rys. 5.

10^{11}–10^{13} eV

10^{11} eV–10^{13} eV. Strumień pozaziemskiego promieniowania γ o energii powyżej ok. 10^{10} eV jest zbyt mały, by można było prowadzić jego obserwacje za pomocą opisanych poprzednio przyrządów wymagających przecięcia się drogi kwantu z detektorem. Uzyskanie znaczącej liczby zliczeń wymaga użycia detektorów o rozmiarach porównywalnych z rozmiarami Ziemi. Dlatego też do obserwacji kwantów o tak dużych energiach wykorzystuje się zjawiska zachodzące w atmosferze ziemskiej, a dokładniej — wielkie pęki atmosferyczne, czyli rozległe kaskady cząstek wtórnych powstające w gazie atmosferycznym pod wpływem promieniowania γ. Wtórne cząstki pęku, poruszające się z prędkościami większymi od prędkości rozchodzenia się światła w atmosferze, są źródłem promieniowania Czerenkowa. Promieniowanie to, przypadające w dziedzinie widzialnej, może być rejestrowane przez zwykłe teleskopy optyczne wyposażone w szybkie kamery fotograficzne i umieszczone na powierzchni Ziemi nawet w znacznych odległościach od wlotu pierwotnego fotonu γ do atmosfery. Błyski promieniowania Czerenkowa mają na zdjęciach wygląd kolistych lub wydłużonych plamek o rozmiarach kątowych rzędu jednego stopnia (il. 214, tabl. 57). Chociaż kształt i orientacja błysków zależą od położenia osi pęku w stosunku do obserwatora, to jednak punkt największej intensywności błysku leży w przybliżeniu w tym kierunku, z którego nadbiegł pierwotny foton γ. Długotrwałe obserwacje mogą doprowadzić do stwierdzenia ewentualnego grupowania się błysków Czerenkowa w pewnych obszarach nieba, a tym samym do wykrycia i zlokalizowania dyskretnych źródeł pozaziemskiego promieniowania γ. Metody tej nie można by jednak użyć do obserwacji promieniowania γ rozłożonego izotropowo na niebie.

10^{14}–10^{16} eV

10^{14} eV–10^{16} eV. Wielkie pęki atmosferyczne wytworzone przez cząstki lub fotony o energiach powyżej ok. 10^{14} eV dobiegają aż do powierzchni Ziemi, gdzie określenie ich składu pozwala jednoznacznie stwierdzić, czy źródłem ich był nukleon czy foton. Dokładniej mówiąc, w pękach pochodzenia fotonowego mniejsza jest zawartość mionów, a więcej jest elektronów i pozytonów. Obserwacje promieniowania γ w tym zakresie prowadzi się więc za pomocą układu detektorów cząstek pokrywającego stosunkowo duży obszar na powierzchni Ziemi. Kierunek osi pęku, pokrywający się w przybliżeniu z kierunkiem przylotu pierwotnego fotonu, można wyznaczyć na podstawie opóźnień czasowych między sygnałami pochodzącymi od różnych liczników. Zazwyczaj, jednocześnie z obserwacjami strumieni cząstek wtórnych prowadzi się również obserwacje optycznych błysków promieniowania Czerenkowa.

Procesy fizyczne prowadzące do emisji promieniowania X i γ

Spośród wszystkich możliwych procesów prowadzących do emisji promieniowania X i γ wyliczymy tylko te, które mogą przebiegać w źródłach kosmicznych dających się bezpośrednio obserwować.

Termiczne promieniowanie ciała doskonale czarnego

Wszystkie ciała materialne znajdujące się w warunkach równowagi termodynamicznej w temperaturze T są źródłami promieniowania termicznego. Zgodnie z prawem Plancka, liczba fotonów dq wysłana w przedziale energii $d\varepsilon$ jest opisana przez wyrażenie

prawo Plancka

$$dq \sim \exp[(\varepsilon/kT)-1]^{-1}d\varepsilon.$$

Wydajnym źródłem fotonów X są ciała o temperaturze 10^6–10^8 K. Ciało będące równie wydajnym źródłem fotonów γ musiałoby mieć temperaturę większą od 10^{10} K, jest jednak mało prawdopodobne, by ciała doskonale czarne o takich temperaturach występowały we Wszechświecie.

Promieniowanie synchrotronowe

Każda cząstka obdarzona ładunkiem elektrycznym poruszająca się w poprzecznym polu magnetycznym doznaje przyspieszenia, a tym samym jest źródłem promieniowania elektromagnetycznego. Promieniowanie to może przypadać w zakresie X i γ, jeżeli prędkości cząstek są bardzo bliskie prędkości światła (cząstki ultra-relatywistyczne). Najwydajniejszym źródłem promieniowania synchrotronowego są elektrony. Jeżeli widmo energetyczne elektronów byłoby opisane przez wyrażenie $dN/dE \sim E^{-\alpha}$, gdzie dN — liczba elektronów o energii zawartej w przedziale dE, wówczas energetyczne widmo fotonów promieniowania synchrotronowego byłoby postaci

$$\frac{dq}{d\varepsilon} \sim \varepsilon^{-\frac{\alpha+1}{2}}.$$

Odwrotny efekt Comptona

W procesie tym kwant promieniowania o dużej energii powstaje w wyniku zderzenia relatywistycznego elektronu z kwantem promieniowania o małej energii. Widmo fotonów promieniowania X i γ powstających w tym procesie jest takie samo jak promieniowania synchrotronowego (przy takim samym widmie energetycznym świecących elektronów).

Promieniowanie hamowania (Bremsstrahlung)

Promieniowanie to jest wysyłane przez elektrony w rzadkiej i gorącej plazmie, doznające przyspieszeń w polach elektrostatycznych ciężkich cząstek obdarzonych ładunkiem dodatnim. Mechanizm ten różni się od omówionych poprzednio pod dwoma ważnymi względami. Po pierwsze, elektrony wysyłające promieniowanie hamowania w zakresie X i γ mogą mieć energię niewiele przewyższającą energię wysyłanych fotonów, innymi słowy mogą być nierelatywistyczne. Po drugie, mechanizm ten jest niezwykle wydajny, tzn. nawet stosunkowo niewielkie ilości plazmy mogą być źródłem intensywnego promieniowania hamowania. W przypadku plazmy znajdującej się w stanie równowagi statystycznej, której elektrony mają energie opisane przez rozkład Maxwella, widmo energetyczne emitowanych fotonów jest postaci

$$\frac{dq}{d\varepsilon} \sim e^{-(\varepsilon/kT)}/\varepsilon.$$

Procesy jądrowe

Emisja promieniowania γ towarzyszy praktycznie wszystkim procesom jądrowym wywoływanym przez oddziaływania silne i elektromagnetyczne. Procesy, które mogą być źródłami kosmicznego promieniowania γ, to przede wszystkim rozpad mezonów π^0, anihilacja par cząstek materii i antymaterii, deekscytacja jąder wzbudzonych oraz rozpad jąder ciężkich.

rozpad mezonów π^0

Mezony π^0 są cząstkami nietrwałymi, rozpadającymi się na dwa fotony γ z połówkowym czasem rozpadu $2\cdot10^{-16}$ s. W układzie odniesienia związanym z mezonem oba fotony mają energię $W_0 = 67{,}5$ MeV. W układzie obserwatora, w którym mezony π^0

mogą mieć bardzo duże prędkości, energie fotonów γ mogą zawierać się w przedziale od $W_0\sqrt{(1-\beta)/(1+\beta)}$ do $W_0\sqrt{(1+\beta)/(1-\beta)}$, gdzie β — stosunek prędkości mezonu do prędkości światła. Ponieważ mezony produkowane w źródłach kosmicznych mają zapewne różne prędkości, widmo fotonów γ pochodzących z ich rozpadu jest rozmyte i ma maksimum w pobliżu energii W_0. Reakcjami, w których mogą się tworzyć mezony π^0 pochodzenia kosmicznego, są przede wszystkim zderzenia bardzo szybkich protonów z protonami, cząstkami α i fotonami o małych energiach.

anihilacja e^-e^+

Zderzenie elektronu z pozytonem prowadzi do anihilacji obu cząstek z jednoczesną emisją dwóch fotonów γ. W przypadku anihilacji w spoczynku oba fotony mają energię 0,51 MeV. W zderzeniach o niezerowej energii jeden z fotonów unosi niemal całą energię kinetyczną cząstek, podczas gdy drugi foton unosi tylko energię spoczynkową 0,51 MeV.

anihalacja $p\bar{p}$

Anihilacja proton-antyproton prowadzi do produkcji mezonów π, w tym również mezonów π^0, które rozpadają się następnie z emisją fotonów γ. Średnio jeden proces anihilacji proton-antyproton jest źródłem 3,5 fotonów γ.

Emisyjne linie X i γ

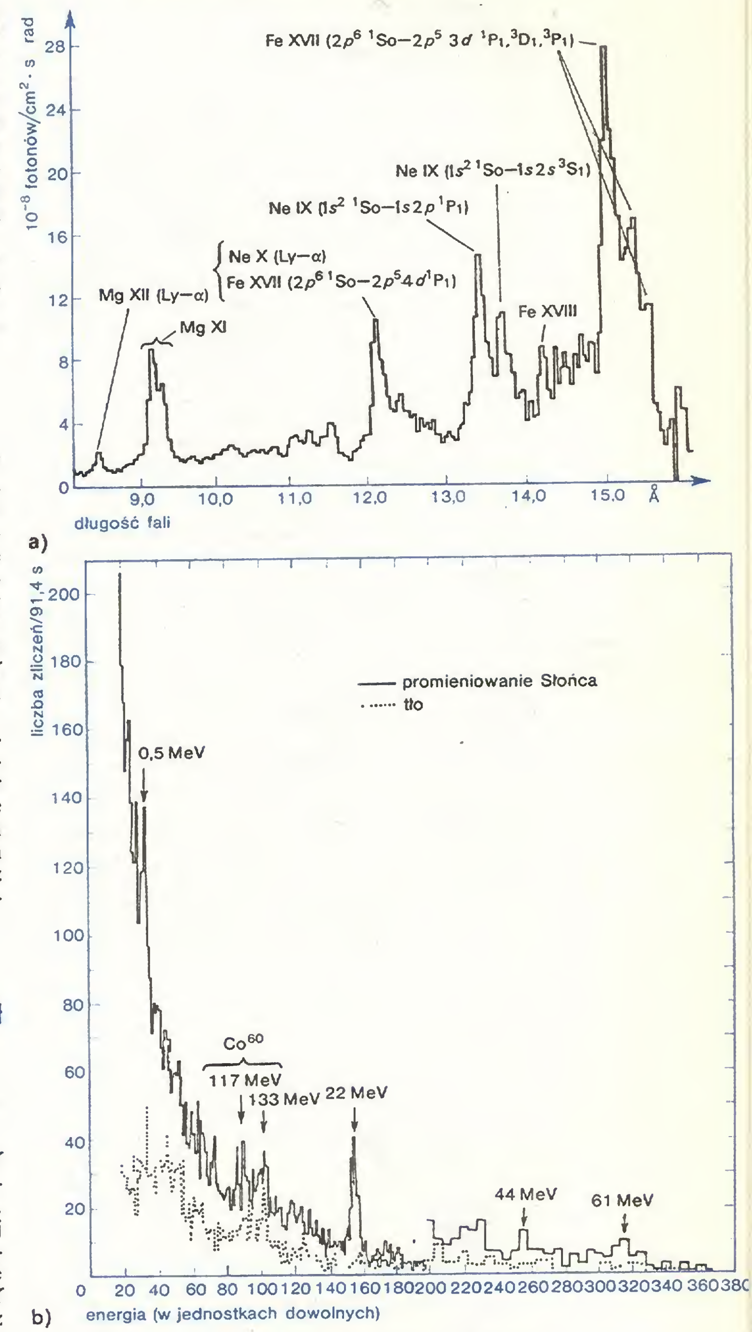
Oprócz wymienionych wyżej mechanizmów produkcji fotonów X i γ o ciągłym rozkładzie widmowym istnieją również procesy prowadzące do powstania fotonów o ściśle określonych energiach. Źródłem linii widmowych przypadających w rentgenowskim zakresie widma są atomy pierwiastków ciężkich, w których — np. wskutek zderzeń — zostały wzbudzone elektrony z najniższych powłok elektronowych, tzw. powłok K. Przejściu elektronu z powłoki wyższej na wolne miejsce w powłoce K towarzyszy wysłanie fotonu X o ściśle określonej energii. Każdy pierwiastek wysyła charakterystyczne dla siebie linie.

Reakcje jądrowe przebiegające w źródłach kosmicznych lub pojedyncze zderzenia szybkich cząstek z jądrami ciężkimi mogą prowadzić do powstania jąder wzbudzonych. Kwantowym przejściom jąder ze stanów wzbudzonych do stanów podstawowych towarzyszy emisja fotonów γ o charakterystycznych dla danego jądra energiach. Źródłem linii γ są również rozpadające się jądra pierwiastków promieniotwórczych.

Obserwacje pozaziemskich źródeł promieniowania X i γ

Słońce

Słońce jest najsilniejszym źródłem promieniowania X i miękkiego promieniowania γ na niebie. Jego obserwacje w tych zakresach energii, rozpoczęte przed ok. dwudziestu laty, dostarczyły już nam bardzo dużej ilości informacji o procesach wielkich energii przebiegających w obszarach koronalnych. Spokojne Słońce jest źródłem miękkiego promieniowania X o widmie ciągłym, wysyłanego przez rozrzedzony gaz korony o temperaturze ok. 1,5 mln stopni (il. 213, tabl. 57). W okresach wzmożonej aktywności na widmo to nakłada się promieniowanie obszarów aktywnych zawierające znacznie twardsze promieniowanie ciągłe oraz liczne emisyjne linie widmowe w dziedzinie X i γ. Widmo liniowe obszarów aktywnych jest najprawdopodobniej promieniowaniem hamowania wysyłanych przez strumienie szybkich elektronów poruszających się w plazmie korony. Liniowe widmo X pochodzi z przejść elektronowych w powłokach wielokrotnie zjonizowanych atomów pierwiastków ciężkich. W liniowym widmie γ zidentyfikowano linię 0,5 MeV pochodzącą z anihilacji par elektron-pozyton, linię 2,2 MeV pochodzącą z reakcji wychwytu neutronu przez jądro wodoru: $n+p \rightarrow d+\gamma$, gdzie d oznacza jądro ciężkiego wodoru — deuteron, oraz linie powstające podczas deekscytacji wzbudzonych jąder C^{12} (4,4 MeV) i O^{16} (6,1 MeV). Przykłady widm rozbłysków słonecznych są przedstawione na rys. 6.



Rys. 6. Widmo promieniowania obszaru aktywnego na Słońcu: a) w zakresie miękkiego promieniowania X, b) w zakresie miękkiego promieniowania γ

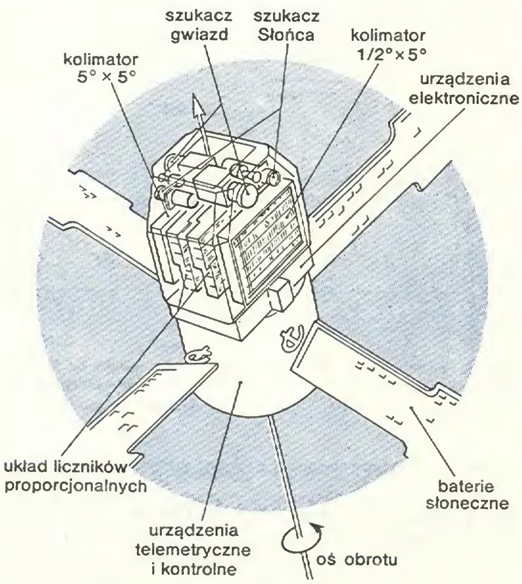
Pierwsze obserwacje źródeł pozasłonecznych

Pierwszej udanej obserwacji źródła promieniowania X znajdującego się poza Układem Słonecznym dokonano 12 VI 1962 r. podczas trwającego 350 s lotu rakietowego. Urządzenie odbiorcze (il. 216, tabl. 58) stanowił licznik Geigera, którego wlot był przesłonięty

wykrycie źródła Sco X-1

scyntylatorem tworzącym z licznikiem układ antykoincydencyjny i mającym za zadanie eliminację zliczeń pochodzących od cząstek promieniowania kosmicznego. Zależność liczby zliczeń od kąta pozycyjnego odbiornika wskazywała na obecność w gwiazdozbiorze Scorpius dość silnego źródła miękkiego promieniowania *X*. Mała kątowa zdolność rozdzielcza nie pozwalała dokładnie umiejscowić źródła, była jednak wystarczająca, by jako źródła odbieranego promieniowania można było wykluczyć ciała układu planetarnego. Na uwagę zasługiwał również fakt, że zliczenia w żadnym kierunku nie malały do zera, co świadczyło o istnieniu ogólnego tła promieniowania *X*. Wkrótce potem na podstawie innych obserwacji można było stwierdzić, że źródłem promieniowania *X* jest również okolica nieba, w której znajduje się pozostałość po wybuchu supernowej — Mgławica Krab. Obserwacje wykonane 7 VII 1964 r. podczas zaćmienia tej okolicy nieba przez Księżyc pozwoliły stwierdzić, że źródłem promieniowania jest zarówno sama mgławica, jak i obszar o średnicy ok. 1′ położony wewnątrz niej. Była to pierwsza optyczna identyfikacja kosmicznego źródła promieniowania *X*. Identyfikacji źródła w gwiazdozbiorze Scorpius dokonano w 1966 r. Obiektem optycznym okazała się niebieska gwiazda 13^m mająca w widmie emisyjne linie wodoru i zjonizowanego helu, wykazująca stosunkowo szybkie i nieregularne zmiany jasności. Źródło to nazwano ScoX-1. (Początkowo nazwa źródła składała się ze skróconej nazwy gwiazdozbioru, w którego kierunku znajduje się źródło, oraz kolejnego numeru odpowiadającego kolejności odkrycia, pokrywającej się zazwyczaj z kolejnością według natężenia promieniowania; litera *X* wskazywała na rentgenowski charakter źródła. W miarę odkrywania coraz to większej liczby źródeł utarł się zwyczaj tworzenia nazwy źródła z symbolu satelity, który źródło to obserwował, oraz liczb podających przybliżone współrzędne równikowe, np. 3U1617-60 A 0620-00 itp.). Optyczne cechy źródła Sco X-1, a zwłaszcza fakt, że w dziedzinie rentgenowskiej wysyła ono w przybliżeniu tysiąc razy więcej energii niż w dziedzinie widzialnej, dowodziły, że odkryto obiekt gwiazdowy, którego istnienia nie można było przewidzieć na podstawie dotychczasowych danych o źródłach optycznych i radiowych. Ten pierwszy wielki sukces astronomii rentgenowskiej spowodował jej dalszy szybki rozwój. Przełomowym momentem było wysłanie 12 XII 1970 r. satelity Uhuru, który miał wyłącznie obserwować promieniowanie *X*. Jego

nazwy źródeł



Rys. 7. Satelita rentgenowski Uhuru (SAS-A)

satelita Uhuru

zadaniem było dokonanie przeglądu całego nieba w zakresie energii od 2 do 20 keV. Schematyczny wygląd takiego satelity jest przedstawiony na rys. 7. Był on wyposażony w dwa niezależne układy odbiorcze, z których każdy składał się z licznika proporcjonalnego przesłoniętego oknem berylowym i kolimatorem mechanicznym. Jedno urządzenie odbiorcze miało pole widzenia $1/2° \times 5°$ i rejestowało fotony do 30 keV, drugie miało pole widzenia $5° \times 5°$ i rejestrowało fotony do 10 keV. Każdy detektor był wyposażony w ośmiokanałowy analizator impulsów, dzięki czemu można było uzyskać ośmiopunktowe widma obserwowanego promieniowania. Czasowa zdolność rozdzielcza zliczeń mogła być w ograniczonym zakresie zmieniana i wynosiła 0,096 s, 0,192 s lub 0,384 s. Detektory mogły wykrywać źródła dające ok. $2 \cdot 10^{-3}$ zliczeń na cm^2 i sekundę, co odpowiada strumieniowi energii rzędu $10^{-15}\ J \cdot cm^{-2} \cdot s^{-1}$. W ciągu ok. 2 lat działania aparatury pomiarowej satelita dokonał przeglądu całego nieba, a uzyskane wyniki posłużyły do sporządzenia katalogu źródeł promieniowania *X* obejmującego 161 obiektów, których położenia na niebie są określone z dokładnością rzędu jednej dziesiątej stopnia kwadratowego.

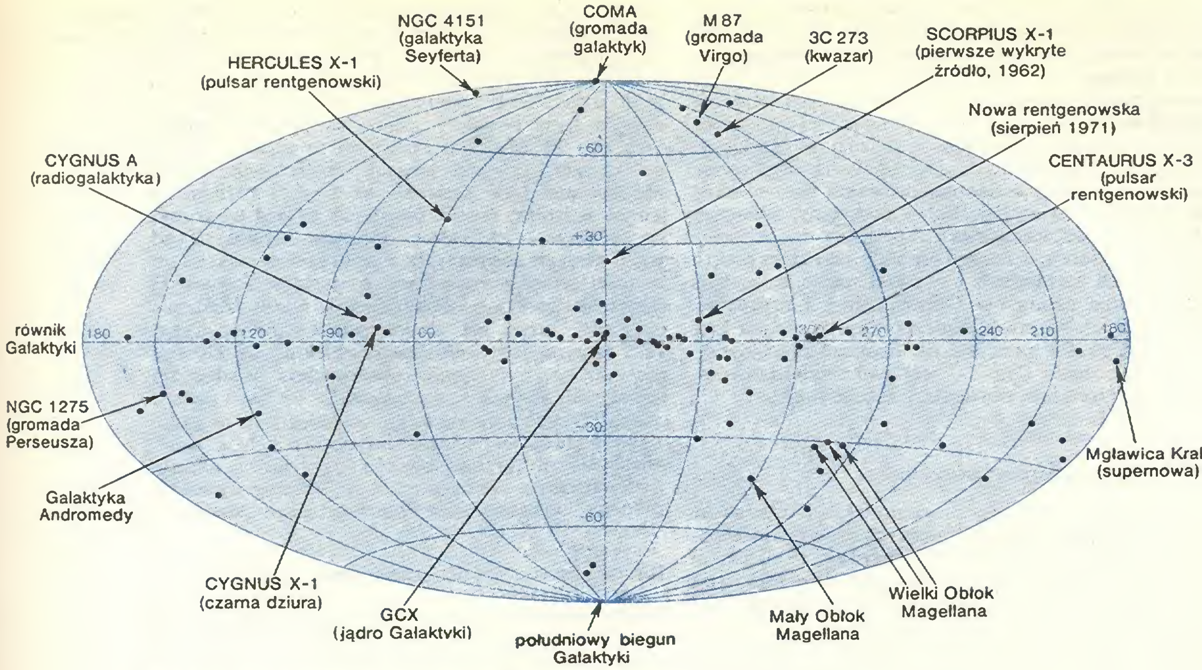
Ostateczny katalog źródeł rentgenowskich obserwowanych przez satelitę Uhuru, tzw. katalog 3U. jest podstawą dalszych prac obserwacyjnych i teoretycznych.

Rozkład źródeł promieniowania *X* na niebie

Rysunek 8 przedstawia rozkład źródeł rentgenowskich we współrzędnych galaktycznych. Już na pierwszy rzut oka sugeruje on istnienie dwu rodzajów źródeł. Część ich wyraznie skupia się wokół płaszczyzny Galaktyki, co pozwala przypuszczać, że są to źródła galaktyczne i że tworzą one podsystem płaski (→ Galaktyki). Pozostałe źródła są rozmieszczone na niebie mniej więcej izotropowo i najprawdopodobniej są pochodzenia pozagalaktycznego. Przypuszczenia te znajdują potwierdzenie zarówno w statystycznych badaniach, jak i w istniejących identyfikacjach optycznych. Źródła skupione wokół płaszczyzny Galaktyki mają średnio większe obserwowane jasności rentgenowskie niż źródła o rozkładzie izotropowym. Wyznaczenie ich rentgenowskiej jasności absolutnej wymaga znajomości ich odległości od nas. Wówczas gdy odległość tę możemy ocenić w sposób pośredni otrzymujemy jasność absolutną rzędu $10^{31}\ J \cdot s^{-1}$, co jest zgodne z wynikami wyznaczeń jasności absolutnej źródeł w Wielkim Obłoku Magellana, którego odległość jest znana. Oceniona na tej podstawie łączna jasność rentgenowska Galaktyki jest rzędu $2 \cdot 10^{32}\ J \cdot s^{-1}$ (dla porównania, energetyczną jasność absolutną Galaktyki we wszystkich zakresach widmowych ocenia się na $10^{37}\ J \cdot s^{-1}$). Podobną absolutną jasność rentgenowską ma również galaktyka M31 w gwiazdozbiorze Andromedy.

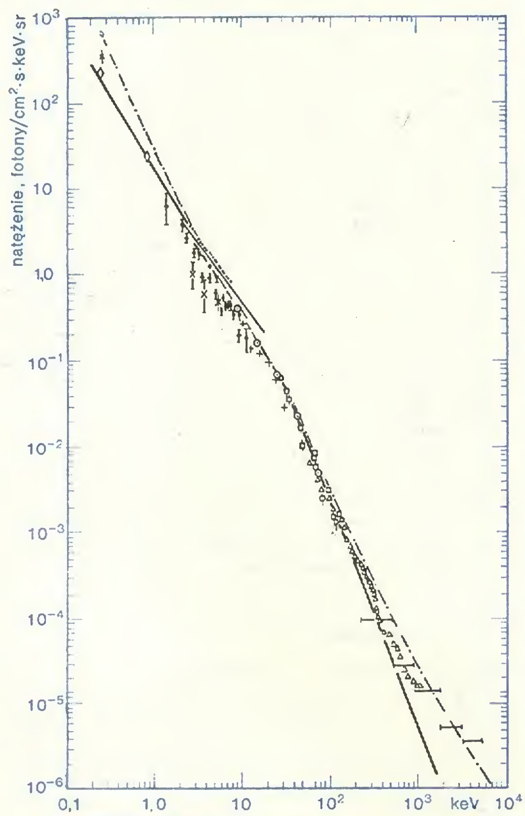
Tło promieniowania *X* i γ

Rozmyte i najprawdopodobniej izotropowe promieniowanie *X* i γ zostało wykryte już w najwcześniejszych eksperymentach i od tej chwili było przedmiotem wielu obserwacji. Na rys. 9 zebrane są wyniki różnych obserwacji dotyczących widma tego promieniowania. Jak dotychczas nie udało się jeszcze wyjaśnić w sposób zadowalający pochodzenia tego promieniowania. Może ono być zarówno pochodzenia lokalnego, tzn. powstawać w naszej Galaktyce, jak i pochodzenia pozagalaktycznego, tzn. powstawać albo w przestrzeni międzygalaktycznej, albo być sumą wielu nierozdzielonych źródeł pozagalaktycznych. Wydaje się jednak, że ta ostatnia możliwość jest najmniej prawdopodobna, choć nie można wykluczyć istnienia pewnej składowej pozagalaktycznej,



Rys. 8. Rozkład źródeł katalogu 3U we współrzędnych galaktycznych

„niebo rentgenowskie"



Rys. 9. Widmo tła rentgenowskiego. Różne symbole odpowiadają różnym eksperymentom

nakładającej się na tło produkowane w naszej Galaktyce. Jednoznaczne rozwiązanie problemu pochodzenia tła rentgenowskiego będzie miało duże znaczenie dla poznania warunków fizycznych panujących w odległych obszarach przestrzeni międzygalaktycznej i międzygwiazdowej.

Źródła podwójne

Kilka spośród znanych źródeł wykazuje periodyczne zmiany jasności rentgenowskiej. Okresy tych zmian, od 4,8 h (Cyg X-3) do 8,95 d (Vela X-1) świadczą o tym, że obiekty te mogą należeć do klasy ciasnych układów podwójnych (→ Ewolucja gwiazd). Przypuszczenie to znalazło pełne potwierdzenie po dokonaniu identyfikacji dwu spośród tych źródeł (Cen X-3 i Her X-1) z układami zaćmieniowymi o okresach zaćmień optycznych równych okresom zaćmień rentgenowskich (2,067 d dla Cen X-3 i 1,7 d dla Her X-1). Oba te źródła są tym bardziej interesujące, że ich składniki rentgenowskie są jednocześnie pulsarami rentgenowskimi (→ Pulsary), tzn. wysyłają promieniowanie *X* w postaci bardzo krótkich impulsów powtarzających się co 4,8 s (Cen X-3) i 1,24 s (Her X-1). Liczne i różnorodne dane obserwacyjne — zarówno rentgenowskie, jak i optyczne — pozwoliły opracować dość prawdopodobny model tych źródeł: Układ podwójny złożony jest z normalnej gwiazdy o cechach charakterystycznych dla gorącego olbrzyma i masie kilku lub kilkunastu mas Słońca, wokół którego po ciasnej orbicie kołowej krąży obracająca się szybko gwiazda neutronowa o masie rzędu 1 masy Słońca. Rozrzedzony strumień materii wypływający z pierwszej gwiazdy (np. w postaci wiatru gwiazdowego) jest częściowo wychwytywany przez gwiazdę neutronową i wzdłuż linii sił pola magnetycznego spływa w jej obszary biegunowe. Ze względu na małe rozmiary gwiazdy neutronowej (rzędu kilku kilometrów) pole grawitacyjne w jej pobliżu ma ogromne natężenie. Cząstki rozrzedzonego strumienia materii opadające na gwiazdę neutronową mogą wyzwalać ogromne ilości energii grawitacyjnej, która ulega przemianie w energię cieplną, a następnie — za pośrednictwem jednego z mechanizmów wspomnianych na wstępie — w energię promieniowania *X*. Szybki obrót gwiazdy neutronowej wokół osi nachylonej względem osi pola magnetycznego sprawia, że promieniowanie wysyłane przez obszary biegunowe dobiega do nas w postaci regularnych impulsów.

źródła Cen X-3 i Her X-1

model: gorący olbrzym i gwiazda neutronowa

Innym źródłem niezmiernie interesującym, a zarazem zagadkowym, jest Cyg X-1. Wykazuje ono bardzo skomplikowane zmiany jasności rentgenowskiej,

źródło Cyg-X1

optycznej i radiowej. Zostało ono zidentyfikowane ze zmienną spektroskopową o okresie 5,6 d. Okresowość ta nie występuje jednak w zmianach jasności optycznej i rentgenowskiej. Promieniowanie *X* dobiega do nas w postaci ciągów impulsów trwających od 1 ms do 100 ms, które jednak nie powtarzają się ściśle periodycznie. Analiza obserwacji w dziedzinie optycznej, wielkość amplitudy prędkości radialnych oraz cechy widmowe gwiazdy widocznej prowadzą do wniosku, że jeżeli jest to zwykły układ podwójny, wówczas składnik niewidoczny musi mieć masę co najmniej kilku mas Słońca. Jedynym trwałym obiektem o bardzo małych rozmiarach i o takiej masie może być tylko czarna dziura (→ Czarne dziury i zapadanie grawitacyjne). Wprawdzie interpretacja ta nie jest konieczna (Cyg X-1 może być np. układem potrójnym), mimo to obecnie Cyg X-1 jest najbardziej prawdopodobnym układem zawierającym taki obiekt.

Pozostałości supernowych

Emisję promieniowania rentgenowskiego obserwuje się z kilkunastu mgławic będących niewątpliwie pozostałościami po wybuchach supernowych w naszej Galaktyce. Najlepiej poznanym obiektem tego typu jest mgławica Krab i znajdujący się wewnątrz niej pulsar (radiowy, optyczny i rentgenowski) NP0531. Źródłem emisji *X* jest obszar mgławicy mający średnicę ok. 100″ (il. 215, tabl. 58). Ok. 15% energii zawarte jest w pulsach o okresie 33 ms, przy czym pulsy rentgenowskie są zgodne w fazie z pulsami optycznymi i radiowymi. Mgławica Krab jest również źródłem promieniowania γ o energii większej od 10^{11} eV (obserwowany strumień $4\cdot10^{-11}$ fotonów $cm^{-2}\cdot s^{-1}$). W przedziale 2–20 keV od mgławicy Krab dobiega do nas strumień równy $2\cdot10^{-15}$ $J\cdot cm^{-2}\cdot s^{-1}$, co przy znanej odległości 2 kps daje absolutną jasność rentgenowską $1{,}5\cdot10^{30}$ $J\cdot s^{-1}$. Stwierdzone w dziedzinie radiowej spowalnianie obrotu pulsara wskazuje na utratę energii rotacyjnej rzędu $3\cdot10^{31}$ $J\cdot s^{-1}$. Wystarcza to do podtrzymywania jasności samego pulsara jak i otaczającej go mgławicy. Nawiasem mówiąc, pulsar NP0531, odkryty najpierw jako pulsar radiowy a następnie optyczny, jest przede wszystkim pulsarem rentgenowskim, wysyłającym w zakresie *X* sto razy więcej energii niż w dziedzinie widzialnej i sto tysięcy razy więcej w dziedzinie radiowej.

mgławica Krab

mgławica Vela X

Równie interesującym obiektem tego typu jest mgławica Vela X, zawierająca pulsar radiowy PSR 0833-45. Jest ona pozostałością supernowej nieco starszą od mgławicy Krab. Źródłem promieniowania *X* jest zarówno rozległa mgławica, jak i pulsar (il. 217, tabl. 58). Część energii promieniowania *X* jest modulowana z okresem radiowym pulsara. Chociaż mgławica Vela X ma wiele cech wspólnych z mgławicą Krab, to jednak ze względu na swą większą rozległość w przestrzeni i starszy wiek może dostarczyć wielu interesujących informacji na temat oddziaływania rozszerzającej się powłoki supernowej z materią międzygwiazdową. Efekty te, nie dające się jeszcze zauważyć u mgławicy Krab, są już dość silnie zaznaczone w promieniowaniu mgławicy Vela X.

Chwilowe źródła promieniowania *X* (nowe rentgenowskie)

Obserwacje prowadzone w latach 1971–75 doprowadziły do wykrycia pięciu interesujących źródeł o wyraźnie aperiodycznym charakterze zmian jasności rentgenowskiej. Wspólną cechą tych źródeł jest szybki wzrost jasności o kilka rzędów wielkości w ciągu dnia lub kilku dni, a następnie stosunkowo powolny spadek w skali czasowej rzędu miesięcy. Wzrostowi jasności towarzyszy na ogół zmiana widma. Najwięcej obserwacji zebrano dla źródła A 0620-00, a ponadto jego położenie zostało określone z tak dużą dokładnością, że udało się dokonać jego identyfikacji optycznej z gwiazdą, która w okresie wzrostu jasności rentgenowskiej zwiększyła swą jasność optyczną z 18^m do 11^m. Jednocześnie, na podstawie archiwalnych zdjęć nieba stwierdzono, że ta sama gwiazda zwiększyła w podobny sposób swoją jasność optyczną ok. 70 lat temu. Pozwoliło to zaliczyć obiekt do kategorii gwiazd nowych powrotnych. Nowa Monocerotis (jak nazwano to źródło) różniła się jednak w istotny sposób od innych gwiazd tej grupy; po pierwsze — jej widmo optyczne różniło się od widm nowych powrotnych; a po drugie, w maksimum jasności promieniowała ona ok. tysiąca razy więcej energii w dziedzinie rentgenowskiej niż w dziedzinie optycznej, czego nie obserwowano u nowej Cygni 1975 (zresztą jedynego innego obiektu, dla którego prowadzone były takie obserwacje). Mechanizm wybuchów nowych rentgenowskich nie jest jeszcze znany. Przypuszcza się tylko, że obiekty te są układami podwójnymi o dość dużych orbitach, a pojawienie się emisji rentgenowskiej jest wynikiem akrecji materii przez jeden ze składników, wzmożonej w wyniku np. okresowej aktywności drugiego składnika.

nowa Monocerotis

Rozbłyskowe źródła promieniowania *X* i γ

Istotny wzrost czasowej zdolności rozdzielczej urządzeń odbiorczych oraz coraz większa liczba satelitów prowadzących obserwacje w zakresie *X* i γ doprowadziły do odkrycia w ciągu ostatnich kilku lat kilkunastu źródeł wysyłających promieniowanie w postaci impulsów trwających od kilku sekund do kilkudziesięciu milisekund. Impulsy te pojawiają się nieperiodycznie i różnią się między sobą całkowitą ilością zawartej w nich energii. W niektórych wypadkach stwierdza się istnienie korelacji między energią zawartą w impulsie i czasem, po którym pojawia się następny impuls: im większa energia impulsu, tym później pojawia się następny impuls. Pierwsze identyfikacje optyczne tych źródeł wskazywały, że obserwowane rozbłyski pochodzą z gromad kulistych. Obecnie znane są jednak również źródła rozbłyskowe nie dające powiązać się z tymi obiektami. Podobne rozbłyski obserwowane były również przez urządzenia odbierające promieniowanie γ. Ponieważ obserwacje w zakresie γ są prowadzone zazwyczaj niezależnie od obserwacji promieniowania *X*, początkowo sądzono, że oba rodzaje rozbłysków mają różne pochodzenie. Obecnie przeważa przypuszczenie, że źródłami obu rodzajów rozbłysków są te same obiekty. Pochodzenie rozbłysków nie zostało jeszcze wyjaśnione. Przypuszcza się tylko, że mogą one być następstwem okresowego opróżniania się „rezerwuarów” plazmy istniejących w pobliżu obiektów masywnych o małych rozmiarach (np. gwiazd neutronowych lub czarnych dziur) i prawdopodobnie wchodzących w skład układów podwójnych.

gromady kuliste

Źródła pozagalaktyczne

Pozagalaktyczne źródła promieniowania *X* mają niewielką obserwowaną jasność rentgenowską i są znacznie trudniejsze do wykrycia. Trudniejsze jest również określenie ich pozycji na niebie z taką dokładnością, by możliwe było dokonanie ich identyfikacji optycznych. Spośród 60 źródeł z katalogu 3U leżących w dużych szerokościach galaktycznych i będących prawdopodobnie źródłami pozagalaktycznymi większość nie została jeszcze zidentyfikowana. Źródłami promieniowania *X* są natomiast z pewnością najbliższe nas galaktyki: Wielki i Mały Obłok Magellana (w których zidentyfikowano również źródła punktowe) oraz M 31. Ich absolutna jasność rentgenowska jest niewielka i zbliżona do jasności rentgenowskiej naszej Galaktyki; jako źródła rentgenowskie obserwujemy je tylko ze względu na małą odległość. Wśród obiektów leżących w znacznie większych odległoś-

ciach, promieniowanie rentgenowskie stwierdzono u jednej galaktyki Seyferta (NGC 4151), u kilku galaktyk aktywnych (będących jednocześnie silnymi radioźródłami lub wykazującymi inne oznaki aktywności), u najjaśniejszego kwazara (3C 273) oraz u kilkunastu gromad galaktyk. Obserwacje gromad galaktyk mogą przynieść szczególnie interesujące wyniki, jeżeli tylko uda się stwierdzić pochodzenie wysyłanego przez nie promieniowania *X*. Jeżeli jego źródłem okaże się np. rozproszona materia międzygalaktyczna w obrębie gromady, fakt ten będzie miał istotny wpływ na rozstrzygnięcie problemu dynamicznej stabilności gromad galaktyk.

obserwacje gromad galaktyk

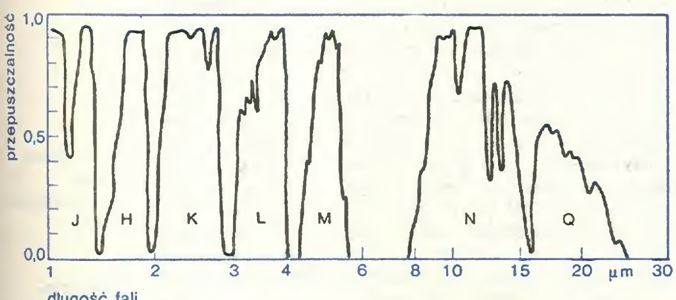
H. Bradt, R. Giacconi (ed.) *X- and gamma-Ray Astronomy*, Dordrecht 1973; R. Giacconi, H. Gursky (ed.) *X-Ray Astronomy*, Dordrecht 1974; S. Kane (ed.) *Solar gamma, X and EUV Radiation*, Dordrecht 1975; F. McDonald, C. Fichtel (ed.) *High Energy Particles and Quanta in Astrophysics*, Cambridge 1974.

Astronomia w podczerwieni

Marcin Kubiak

Nazwą podczerwieni obejmujemy promieniowanie elektromagnetyczne o długościach fal zawartych w przedziale od ok. 1 μm (10 000 Å) do ok. 1000 μm (1 mm). Podczerwień jest przedłużeniem widma widzialnego w kierunku fal dłuższych i podlega tym samym prawom optyki geometrycznej; do skupiania promieniowania podczerwonego używa się zwykłe teleskopy optyczne. Możliwość obserwowania w podczerwieni z powierzchni Ziemi jest ograniczona przede wszystkim przez absorpcyjne właściwości jej atmosfery. Poszczególne składniki atmosfery, a zwłaszcza para wodna (H_2O), dwutlenek węgla (CO_2) oraz ozon (O_3), mają silne pasma absorpcyjne w omawianym przedziale widmowym. Pomiędzy tymi pasmami istnieje jednak kilka zakresów, tzw. okien, w których atmosfera ziemska jest niemal zupełnie przezroczysta. Zależność przepuszczalności atmosfery ziemskiej od długości fali promieniowania podczerwonego, odnosząca się do średniej atmosfery (zawierającej normalną ilość pary wodnej, dwutlenku węgla i ozonu), jest przedstawiona na rys. 1. Chwilowe zmiany składu chemicznego atmosfery, a zwłaszcza jej wilgotności, powodują zmiany szerokości oraz przezroczystości poszczególnych okien. Innym zjawiskiem decydującym o wielkości osłabienia promieniowania podczerwonego w atmosferze ziemskiej jest rozpraszanie na cząstkach materialnych (mgła, smog, pył). Wielkość tego rozpraszania zależy od rozmiarów cząstek i ich koncentracji, natomiast nie zależy prawie od długości fali. Obserwacje w podczerwieni należy więc prowadzić przede wszystkim z miejsc położonych wysoko nad poziomem morza, w okolicach bardzo suchych i wolnych od zanieczyszczeń pyłowych. Z obserwacyjnego punktu widzenia zaletą promieniowania podczerwonego jest fakt, iż nie ulega ono rozproszeniu na atomach i cząsteczkach gazów znajdujących się w atmosferze (tzw. rozpraszanie Rayleigha, którego wynikiem jest barwa nieba w dzień). W zakresie podczerwonym nie ma zatem różnicy między jasnością nieba nocnego i dziennego, co sprawia, że obserwacje w podczerwieni można prowadzić również w ciągu dnia. Przy obserwacjach spoza atmosfery ziemskiej lub spoza jej najgęstszych warstw (z pokładów wysoko lecących samolotów, balonów stratosferycznych, rakiet i sztucznych satelitów) dostępny jest oczywiście cały zakres promieniowania podczerwonego.

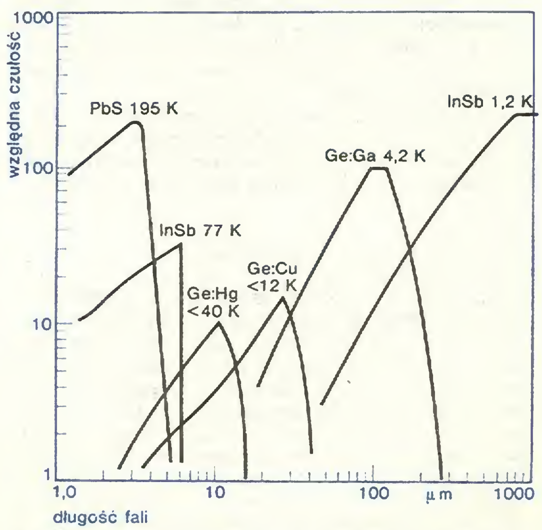
rozpraszanie promieniowania podczerwonego



Rys. 1. Przepuszczalność atmosfery ziemskiej w jednostkach umownych: 1 oznacza zupełną przezroczystość, natomiast 0 całkowitą absorpcję promieniowania. Literami zaznaczone jest przybliżone położenie poszczególnych pasm fotometrii szerokopasmowej opisanych dokładniej w tabeli

Rejestracja promieniowania podczerwonego wymaga specjalnych urządzeń odbiorczych (detektorów) odznaczających się dużą czułością oraz niskim poziomem szumów w interesującym nas zakresie długości fal. Brak takich odbiorników był głównym powodem, dla którego obserwacje w podczerwieni na dużą skalę zostały podjęte stosunkowo niedawno. Chociaż pierwsze pomiary promieniowania podczerwonego Słońca, Księżyca, planet i niektórych najjaśniejszych gwiazd zostały wykonane w latach 30-ych XX w., to jednak za datę narodzin astronomii w podczerwieni należy uznać początek lat 60-ych, gdy dzięki rozwojowi techniki półprzewodnikowej oraz techniki otrzymywania niskich temperatur uzyskano detektory o czułości wystarczającej do rejestracji promieniowania nawet słabych źródeł pozaziemskich. Na rys. 2 przedstawiona jest czułość widmowa niektórych stosowanych obecnie fotokatod. W zakresie tzw. bliskiej podczerwieni (długość fali od 1 μm do ok. 5 μm) energia niesiona przez fotony promieniowania jest wystarczająca do wywołania zjawiska fotoelektrycznego w substancji pokrywającej fotokatodę (Pbs, InSb). Energia niesiona przez fotony o długości fali większej od ok. 5 μm (przeciętnie kilkaset razy mniejsza od energii fotonów promieniowania widzialnego) jest wystarczająca do spowodowania tzw. wewnętrznego zjawiska fotoelektrycznego. (zmiany

detektory promieniowania podczerwonego



Rys. 2. Widmowa czułość kilku substancji będących detektorami promieniowania podczerwonego. Podane są również temperatury, w których detektory te pracują

wewnętrzne zjawisko fotoelektryczne

przewodnictwa kryształu półprzewodnikowego pod wpływem promieniowania podczerwonego) w półprzewodnikach, do których zostały wprowadzone odpowiednie domieszki (Ge:Hg, Ge:Cu, Ge:Ga itp.). Wszystkie substancje czułe na obecność promieniowania podczerwonego są detektorami wydajnymi dopiero w bardzo niskich temperaturach (podanych również na rys. 2). Chłodzenie detektorów ma na celu odpowiednie zmniejszenie poziomu szumów, czyli chaotycznych sygnałów będących następstwem termicznych ruchów cząstek materii samego detektora. Obecnie możliwości detekcji promieniowania podczerwonego sa określone wyłącznie przez poziom szumów mierzonego promieniowania.

eliminacja promieniowania otoczenia

Podstawową trudnością przy obserwacjach w podczerwieni jest wyodrębnienie promieniowania pozaziemskiego spośród promieniowania wysyłanego przez otoczenie detektora. Źródłem promieniowania podczerwonego jest każde ciało o temperaturze wyższej od zera bezwzględnego, a zwłaszcza ciała o temperaturze kilkuset stopni K. Niepożądane promieniowanie podczerwone wysyła więc zarówno atmosfera ziemska (promieniująca jak ciało o temperaturze rzędu 0°C), jak i poszczególne elementy teleskopu znajdującego się w polu widzenia detektora (obiektyw, zwierciadła wtórne itp.). Wpływ promieniowania otoczenia na uzyskane wyniki można wydatnie zmniejszyć lub całkowicie wyeliminować przez staranne i częste wzorcowanie urządzenia obserwacyjnego lub przez zastosowanie sygnału modulowanego w zadany sposób.

Obserwacje pozaatmosferyczne nie zostały jeszcze podjęte na szerszą skalę, pomimo szybkiego rozwoju techniki lotów satelitarnych, i ograniczają się na razie do krótkotrwałych lotów rakietowych, które zresztą, ze względu na krótki efektywny czas obserwacji — rzędu kilku minut — są stosunkowo kosztowne. Główną przeszkodą w podjęciu znacznie ekonomiczniejszych pod tym względem obserwacji z pokładów sztucznych satelitów ziemi jest trudność utrzymania odpowiednio niskiej temperatury w warunkach długotrwałego lotu kosmicznego. Obserwacje promieniowania podczerwonego z przestrzeni kosmicznej wymagałyby chłodzenia nie tylko detektora ale również całego teleskopu, co pociągałoby za sobą konieczność ciągłego dostarczania stosunkowo dużych ilości energii.

Obecnie astronomia w podczerwieni jest jedną z najaktywniej rozwijających się dziedzin astronomii; jej możliwości nie ustępują w zasadzie możliwościom obserwacyjnym w klasycznej dziedzinie widzialnej.

Fotometria szerokopasmowa

Pierwszym, a zarazem stosunkowo najprostszym rodzajem obserwacji podjętych w dziedzinie podczerwonej była fotometria szerokopasmowa. Przepuszczalność atmosfery ziemskiej w powiązaniu z czułością detektorów i właściwościami istniejących filtrów promieniowania podczerwonego określa w sposób naturalny zakres długości fal, w których możemy odbierać pozaziemskie promieniowanie podczerwone.

system fotometryczny 12-pasmowy

Wynikający stąd system fotometryczny został powiązany z tradycyjnym systemem fotometrii trójbarwnej *UBV* (*U* — nadfiolet, *B* — przedział fal niebieskich, *V* — zakres widzialny) w dziedzinie widzialnej w jeden system obejmujący dwanaście pasm. Bliższa charakterystyka poszczególnych pasm tego systemu jest podana w tabeli. Obserwacje w tym systemie są niekiedy uzupełniane przez obserwacje w pasmach pośrednich. W wypadku pasm niedostępnych z powierzchni Ziemi, obserwacje w ich zakresie są wykonywane z wysoko lecących samolotów, z rakiet lub balonów.

Fotometria szerokopasmowa jest najszybszą metodą uzyskiwania informacji o rozkładzie energii w widmie ciągłym w szerokim zakresie długości fal.

Charakterystyka wielobarwnego systemu fotometrii szerokopasmowej

Pasmo	λ_{ef}, μm	$\Delta\lambda$, μm	0 mag, W·cm^{-2}·μm^{-1}
U	0,36	0,066	$4{,}35\cdot10^{-12}$
B	0,44	0,098	$7{,}20\cdot10^{-12}$
V	0,55	0,087	$3{,}92\cdot10^{-12}$
R	0,70	0,21	$1{,}76\cdot10^{-12}$
I	0,90	0,23	$8{,}35\cdot10^{-13}$
J	1,25	0,30	$3{,}40\cdot10^{-13}$
H	1,60	0,30	$1{,}28\cdot10^{-13}$
K	2,20	0,58	$3{,}90\cdot10^{-14}$
L	3,40	0,70	$8{,}10\cdot10^{-15}$
M	5,00	1,13	$2{,}20\cdot10^{-15}$
N	10,20	4,33	$1{,}23\cdot10^{-16}$
Q	22,00	7,5	$7{,}70\cdot10^{-18}$

λ_{ef} — efektywna długość fali, $\Delta\lambda$ — szerokość pasma, 0 mag — bezwględna wartość strumienia energii odpowiadającego zerowej wielkości gwiazdowej w danym pasmie.

Spektrofotometria (fotometria wąskopasmowa)

Dzięki opanowaniu technologii wytwarzania filtrów interferencyjnych o zmiennej charakterystyce widmowej, stały się możliwe również obserwacje fotometryczne w bardzo wąskich zakresach długości fal. Na przykład detektor germanowy (z domieszką miedzi) w połączeniu z filtrem interferencyjnym pozwala uzyskać spektrofotometr mierzący promieniowanie w zakresie 3–14 μm z widmową zdolnością rozdzielczą rzędu 0,1 μm. Dokładność takiego urządzenia jest porównywalna z dokładnością spektrografów, natomiast zakres jego zastosowania jest znacznie szerszy.

Spektroskopia

W krótkofalowym zakresie promieniowania podczerwonego, będącego naturalnym przedłużeniem zakresu widzialnego, technika obserwacji spektroskopowych nie różni się niczym od spektroskopii światła widzialnego. Widma podczerwone w dużej dyspresji można uzyskać za pomocą zwykłych siatek dyfrakcyjnych oraz emulsji fotograficznych czułych na promieniowanie podczerwone. Obserwacje tego rodzaju nie wykraczają jednak poza długość fali ok. 1,2 μm (przy rejestracji fotograficznej) lub ok. 2,5 μm (przy rejestracji fotoelektrycznej, zwłaszcza w obserwacjach rakietowych). W całym zakresie podczerwieni duże zdolności rozdzielcze — rzędu 0,05 μm lub wyjątkowo nawet 0,01 μm — można uzyskać przez zastosowanie interferometrów (np. Fabry'ego–Pèrota lub Michelsona).

interferometry

Technika interferometryczna, chociaż stosunkowo czasochłonna, została z powodzeniem zastosowana w niemal całym zakresie promieniowania podczerwonego dostępnego z powierzchni Ziemi.

Obserwacje bolometryczne

Urządzenie odbierające promieniowanie z bardzo szerokiego przedziału widmowego, a w idealnym wypadku — całe promieniowanie padające — nazywamy bolometrem.

bolometry

Celem obserwacji bolometrycznych jest pomiar całkowitego strumienia energii promienistej dobiegającej do nas od danego ciała niebieskiego. Przy znanej odległości tego ciała, łączna energia dobiegająca do nas w jednostce czasu jest bezpośrednią miarą jego temperatury efektywnej. W praktyce wystarcza, jeżeli używany przez nas bolometr rejestruje promieniowanie w tym zakresie widmowym, w którym przypada większość promieniowania wysyłanego przez obserwowane ciało niebieskie. Urządzeniem czułym w szerokim zakresie pro-

mieniowania podczerwonego — od ok. 2 μm, do ok. 1000 μm — jest np. bolometr germanowy chłodzony ciekłym helem do temperatury poniżej 2 K. Odbiornik taki może spełniać rolę bolometru przy obserwacjach źródeł o temperaturach efektywnych zawartych w przedziale od kilkunastu do ponad tysiąca K. Przy obserwacjach z powierzchni Ziemi bolometr germanowy lub bolometr indowo-antymonowy, którego czułość widmową ilustruje rys. 2, może być wykorzystywany do obserwacji promieniowania w tzw. podmilimetrowym zakresie fal (obszar przejściowy między promieniowaniem podczerwonym i radiowym), w którym atmosfera ziemska staje się ponownie stosunkowo przezroczysta.

bolometr germanowy

Jak już wspominaliśmy, zależnie od temperatury ciała maksimum promieniowanej przez niego energii przypada w różnych obszarach widmowych. W podczerwieni świecą przede wszystkim ciała o temperaturach od ok. 3500 K (maksimum w pobliżu 1 μm) do ok. 3 K (maksimum w pobliżu 1000 μm). Zatem obserwacje w podczerwieni umożliwiają badanie obiektów, które w dziedzinie widzialnej są bardzo słabo albo w ogóle niewidoczne. Ponadto, z obserwacji spektroskopowych uzyskuje się informacje o tych atomach i cząsteczkach, które występują w atmosferach gwiazd i planet (zarówno chłodnych, jak i gorętszych), ale które nie mają linii widmowych w dziedzinie widzialnej. Należy również pamiętać, że istnieją obiekty astronomiczne, których promieniowanie nie ma charakteru wyłącznie termicznego, i które w związku z tym, obok dużych ilości promieniowania widzialnego, mogą wysyłać również silne promieniowanie podczerwone.

Chociaż obserwacje w podczerwieni zostały podjęte stosunkowo niedawno, to jednak przyniosły one już wiele interesujących wyników i znacznie rozszerzyły naszą znajomość zarówno stanu materii we Wszechświecie, jak i procesów fizycznych, które w niej przebiegają. Poniżej podany jest krótki przegląd najważniejszych wyników uzyskanych dzięki obserwacjom w podczerwieni.

Absorpcja międzygwiazdowa

Z obserwacji w dziedzinie widzialnej wiadomo, że materia rozproszona (gaz i pył) w płaszczyźnie Galaktyki wywołuje osłabienie przechodzącego przez nią światła gwiazd. Osłabienie to zależy od długości fali i jest tym mniejsze, im większa jest długość fali (stąd też nazwa — poczerwienienie międzygwiazdowe). Dzięki obserwacjom w podczerwieni zależność absorpcji międzygwiazdowej od długości fali została wyznaczona w szerokim zakresie widmowym. Wyznaczono ją przez porównanie przebiegu widma ciągłego poczerwienionych i niepoczerwienionych gwiazd takiego samego rodzaju. W dziedzinie widzialnej wykorzystuje się w tym celu przede wszystkim jasne gwiazdy gorące, a w dziedzinie podczerwonej — jasne gwiazdy chłodne. Dokładna znajomość krzywej absorpcji międzygwiazdowej jest niezbędna do wielu pośrednich metod wyznaczania odległości ciał niebieskich. Jednocześnie na jej podstawie można uzyskać wiele cennych informacji na temat stanu fizycznego i składu chemicznego materii, która tę absorpcję wywołuje. Obecnie wiadomo, że osłabienie światła gwiazd powodują drobne ziarenka materii w stanie stałym (pył) o charakterystycznych rozmiarach 1 μm. Skład chemiczny pyłu nie jest jeszcze dokładnie znany, jest jednak bardzo prawdopodobne, że zawiera on takie pierwiastki i związki chemiczne, jak tlenek krzemu, żelazo, lód, glin, magnez i wapń. Ocenia się, że łączna masa składowej pyłowej wynosi ok. 1% masy gazu międzygwiazdowego. Na podstawie obserwacji podczerwonego promieniowania gwiazd można przypuszczać, że ziarna pyłu powstają w atmosferach gwiazd chłodnych, skąd przedostają się do przestrzeni międzygwiazdowej.

poczerwienienie międzygwiazdowe

Obserwacje gwiazd

Pierwsze obserwacje gwiazd polegały przede wszystkim na wyznaczeniu przebiegu ich widma ciągłego za pomocą szerokopasmowej fotometrii wielobarwnej, oraz na poszukiwaniu obiektów promieniujących wyłącznie w dziedzinie podczerwonej. Obserwacje obejmowały zarówno gwiazdy gorące (wystarczająco jasne, by można było spodziewać się również mierzalnego strumienia promieniowania podczerwonego), jak i gwiazdy chłodne. Interesujące było stwierdzenie, że stosunkowo dużo gwiazd gorących, zwłaszcza gwiazd z liniami emisyjnymi, świadczącymi o istnieniu wokół nich rozległych otoczek gorącej i prawdopodobnie ekspandującej materii, wykazuje nadwyżki (w stosunku do termicznego promieniowania ciała doskonale czarnego) promieniowania w zakresie fal dłuższych od 3,5 μm. Biorąc pod uwagę, że gorące jasne gwiazdy są na ogół gwiazdami młodymi, nadwyżkę tę można wiązać z promieniowaniem cieplnym pyłu otaczającego gwiazdy młode i będącego pozostałością po pierwotnym obłoku materii gazowo-pyłowej, z którego gwiazdy te powstały. Niezależnym potwierdzeniem tego wniosku jest fakt, że duże nadwyżki promieniowania podczerwonego zostały również stwierdzone przy obserwacji niektórych rodzajów chłodniejszych gwiazd zmiennych (np. tzw. gwiazd typu *T* Tauri i *R* Monocerotis), które według naszych dzisiejszych poglądów są gwiazdami znajdującymi się w początkowych fazach ewolucyjnych. Wyraźne nadwyżki promieniowania podczerwonego zostały stwierdzone również dla wielu normalnych gwiazd chłodnych (np. gwiazd węglowych). Ich widma (w pobliżu 10 μm i 20 μm) wykazują pasma absorpcyjne charakterystyczne dla tlenku krzemu i niektórych związków węgla. Przypuszcza się, że górne warstwy atmosfer tych gwiazd są miejscem tworzenia się ziaren materii w stanie stałym.

gwiazdy gorące

Spektroskopowe obserwacje gwiazd chłodnych w bliskiej i średniej podczerwieni (do ok. 10 μm) dostarczyły interesujących informacji na temat zawartości w ich atmosferach cząsteczek takich związków chemicznych, jak np. CO lub H_2O. Cząsteczki te mają w podczerwieni wiele linii oscylacyjnych i oscylacyjno-rotacyjnych, które w niskich temperaturach powodują nieprzezroczystość atmosfer gwiazdowych. Obserwacje linii rotacyjnych cząsteczek dostarczają ponadto informacji dotyczących względnej obfitości izotopów niektórych pierwiastków (np. węgla) w materii gwiazdowej.

gwiazdy chłodne

Obiekty podczerwone

Podczas wstępnych przeglądów nieba wykryto wiele obiektów, które wysyłają bardzo silne promieniowanie podczerwone, natomiast nie mają odpowiedników w dziedzinie optycznej lub dają się zidentyfikować z obiektami optycznymi o bardzo małej jasności (rys. 3). Podczerwone widma trzech najsłynniejszych obiektów tego rodzaju są pokazane na rys. 4.

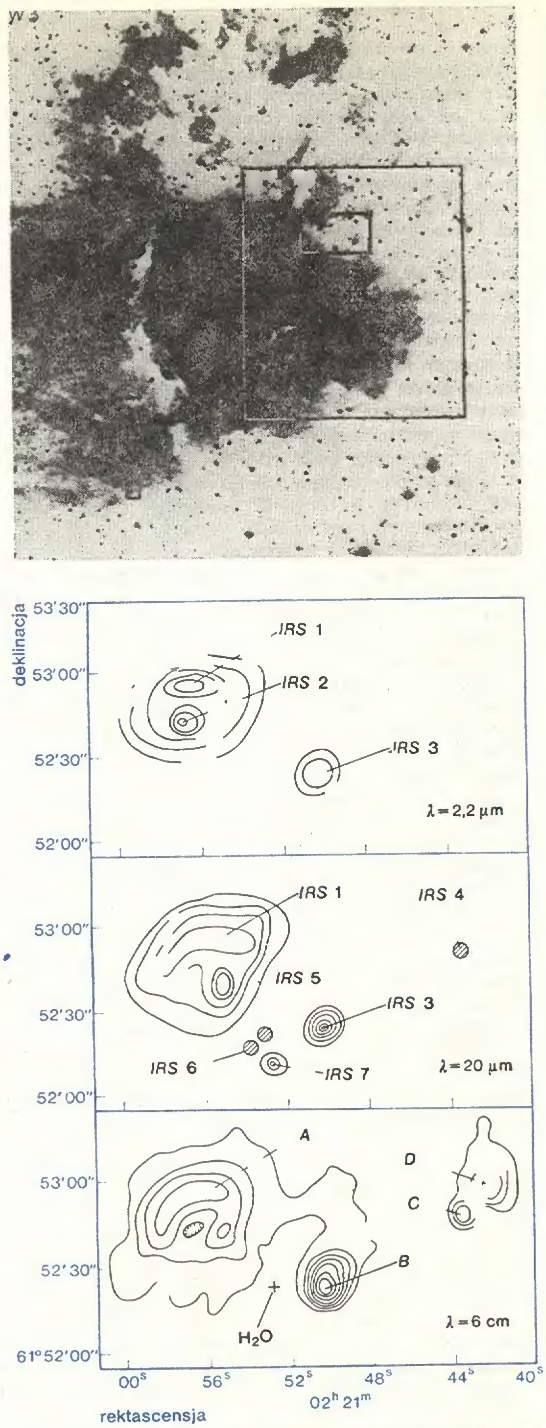
NML Cyg jest jednym z najjaśniejszych obiektów podczerwonych na niebie. Najprawdopodobniej jest to nadolbrzym o cechach promieniowania charakterystycznych dla gwiazdy typu widmowego *M*, wokół którego znajduje się gęsta otoczka gazowo-pyłowa będąca źródłem promieniowania podczerwonego. Optycznym odpowiednikiem źródła NML Cyg jest obiekt 18 wielkości gwiazdowej; niewielka jasność optyczna jest spowodowana silną absorpcją okołogwiazdową. Przypuszcza się, że NML Cyg jest protogwiazdą, w której wnętrzu nie rozpoczęły się jeszcze przemiany jądrowe (czyli nie osiągnęła ona jeszcze ciągu głównego).

NML Cyg

Obiekt Becklina jest to jasny obiekt podczerwony o bardzo małych rozmiarach kątowych, znajdujący się prawdopodobnie wewnątrz rozciągłej mgławicy podczerwonej Kleinmanna-Lowa stanowiącej część

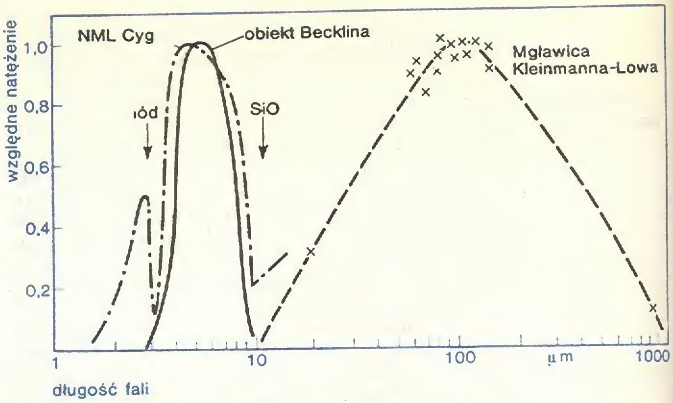
obiekt Becklina

mgławica IC 1795 (W3)



Rys. 3. U góry — zdjęcie mgławicy IC 1795 (W3) wykonane w świetle widzialnym (o długości fali odpowiadającej linii wodoru H_α). Obszar wewnątrz mniejszego prostokąta zawiera liczne źródła promieniowania podczerwonego i radiowego. U dołu — konturowe mapy tego obszaru wykonane w promieniowaniu podczerwonym (2,2 μm i 20 μm) i radiowym (6 cm). Zwróćmy uwagę, że obiekty podczerwone nie mają widocznych odpowiedników optycznych. Źródła promieniowania podczerwonego są oznaczone symbolem *IRS*; duże litery (*A–D*) oznaczają obszary intensywnego promieniowania radiowego (wg L.N. Mavridis (ed.), vol. *2 Stars and the Milky Way System*, Berlin 1974)

gazowo-pyłowej Mgławicy Oriona. Przypuszcza się, że obiekt Becklina jest zwykłym nadolbrzymem typu widmowego *F* (temperatura ok. 7000 K), którego światło widzialne ulega ogromnej absorpcji ok. 80 wielkości gwiazdowych. Pasma absorpcyjne w widmie



Rys. 4. Względny rozkład natężeń w podczerwonych widmach NML Cyg, obiektu Becklina i podczerwonej Mgławicy Kleinmanna-Lowa (krzyżykami zaznaczone są punkty obserwacyjne)

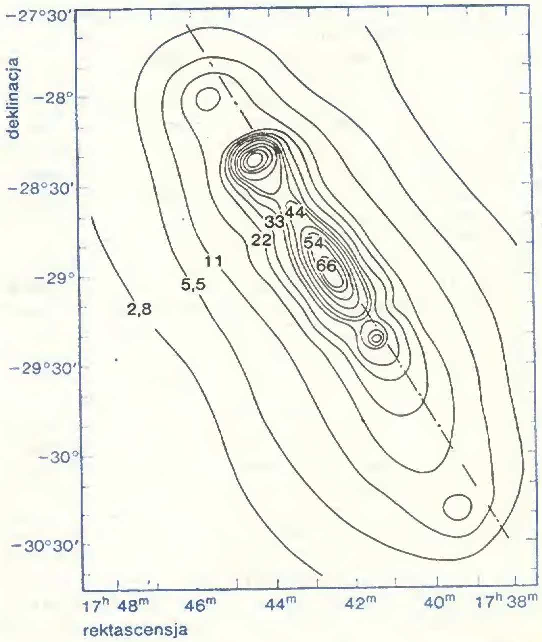
obiektu Becklina są produkowane przez pył znajdujący się w mgławicy Kleinmanna-Lowa.

Mgławica Kleinmanna-Lowa

Poczerwona Mgławica Kleinmanna-Lowa jest rozciągłym źródłem długofalowego promieniowania podczerwonego o rozmiarach kątowych rzędu 30–40 sekund łuku. Mgławica ta nie wysyła dającego się zmierzyć promieniowania w zakresie fal krótszych od ok. 20 um, jest natomiast jeszcze stosunkowo jasna w zakresie podmilimetrowym (900 μm). Obserwowany rozkład natężenia w widmie jest w przybliżeniu taki, jak oczekiwany rozkład promieniowania ziaren pyłu o temperaturze 50–70 K. Mgławica ta jest prawdopodobnie stosunkowo dużym i gęstym obłokiem pyłu międzygwiazdowego, wewnątrz którego przebiegają procesy powstawania gwiazd.

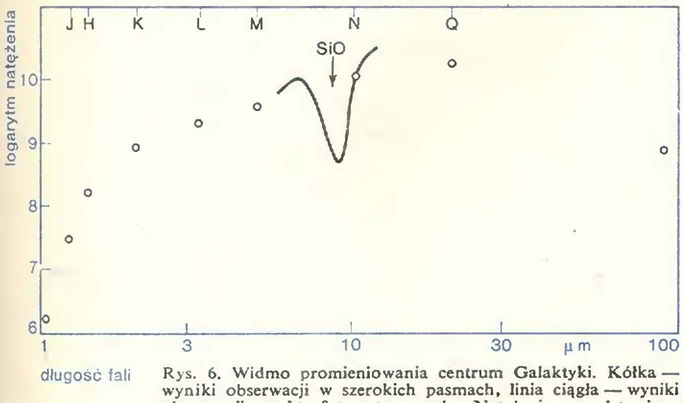
Centrum Galaktyki

Dzięki temu, że absorpcja międzygwiazdowa jest odwrotnie proporcjonalna do długości fali, jądro Galaktyki, całkowicie przesłonięte w dziedzinie



Rys. 5. Konturowa mapa okolic centrum Galaktyki wykonana w promieniowaniu podczerwonym o długości fali 100 μm. Liczby oznaczają natężenie w jednostkach względnych. Przerywaną linią (prostą) zaznaczony jest przebieg płaszczyzny symetrii Galaktyki. (wg V. Manno and J. Ring (ed.) *Infrared Techniques for Space Research*, Dordrecht 1972)

widzialnej, staje się widoczne w promieniowaniu podczerwonym. Fotometryczne obserwacje okolic centrum Galaktyki (rys. 5) ujawniły, że jądro Galaktyki w podczerwieni jest obiektem rozciągłym (o rozmiarach rzędu 1° dla fal z przedziału 75–125 μm i rzędu 30" dla fal 10 i 20 μm), wysyłającym stosunkowo silne promieniowanie. Na tle promieniowania ciągłego widoczne jest pasmo absorpcyjne SiO (9,7 μm) (rys. 6). Pochodzenie tego promieniowania nie jest jeszcze wyjaśnione.



Rys. 6. Widmo promieniowania centrum Galaktyki. Kółka — wyniki obserwacji w szerokich pasmach, linia ciągła — wyniki obserwacji spektrofotometrycznych. Natężenie przedstawione jest w skali logarytmicznej, w jednostkach umownych

Obiekty pozagalaktyczne

Źródłami promieniowania podczerwonego są niektóre galaktyki oraz kwazary. Zwykłe galaktyki oraz galaktyki Seyferta mają widmo promieniowania podczerwonego przypominające typowe widma podczerwone gorących i gęstych obłoków zjonizowanej materii międzygwiazdowej występujących w naszej Galaktyce.

Nasuwa się przypuszczenie, że mechanizm powstawania emisji jest w obu wypadkach taki sam (termiczne promieniowanie pyłu, linie emisyjne niektórych pierwiastków zjonizowanych lub promieniowanie nietermiczne). Wyraźną a ponadto zmienną w czasie nadwyżkę promieniowania podczerwonego stwierdziły również obserwacje najjaśniejszego kwazara 3C 273.

Słońce i planety

Dla badania Słońca i Księżyca (dwu najsilniejszych źródeł promieniowania podczerwonego na niebie) obserwacje w podczerwieni odgrywają przede wszystkim rolę uzupełniającą w stosunku do obserwacji prowadzonych w innych dziedzinach widma. Są one jednak bardzo ważne dla badania planet, które ze względu na stosunkowo niską temperaturę, główną część swej energii wysyłają w podczerwieni. Istotną rolę dla określenia bilansu energetycznego planet mają obserwacje bolometryczne; szczególnie interesujące jest stwierdzenie, że Jowisz wysyła nieco więcej energii niż jej otrzymuje od Słońca: dowodzi to, że ma on własne źródło energii. Celem spektroskopowych obserwacji planet jest przede wszystkim wyznaczenie składu chemicznego ich atmosfer, a zwłaszcza obfitości cząsteczek (np. CO_2, NH_3, CH_4) mających silne pasma absorpcyjne w podczerwieni i występujących w atmosferach planetarnych.

bilans energetyczny Jowisza

P. J. Brancazio, A. G. W. Cameron *Infrared Astronomy*, New York 1968; H. L. Hackforth *Infrared Radiation*, New York 1960; V. Manno, J. Ring (ed.) *Infrared Detection Techniques for Space Research*, Dordrecht 1972; L.N. Mavridis (ed.), vol. 2, *Stars and the Milky Way System*, Berlin 1974.

Ewolucja gwiazd

Józef Smak

Termin ewolucja oznacza w astronomii rozwój pojedynczego obiektu lub grupy obiektów, a nie odnosi się do zmian określonych cech danego typu obiektów zachodzących z pokolenia na pokolenie. Ewolucja gwiazdy obejmuje zatem proces powstania gwiazdy z materii międzygwiazdowej, następnie zachodzące w jej wnętrzu procesy termojądrowe i związane z nimi zmiany parametrów fizycznych i składu chemicznego materii gwiezdnej, kończące się wyczerpaniem źródeł energii i ewentualnym zwrotem części materii do ośrodka międzygwiazdowego, oraz osiągnięcie przez gwiazdę końcowego stadium ewolucji.

Ewolucję gwiazdy można rozważać w aspekcie źródeł jej energii. Słońce, które może być uważane za przeciętną gwiazdę, wypromieniowuje w ciągu sekundy energię równą w przybliżeniu $3{,}86\cdot10^{26}$J. Masa Słońca równa jest $1{,}98\cdot10^{30}$ kg. Wynika stąd, że tempo produkcji energii słonecznej wynosi średnio ok. $2\cdot10^{-4}$ J/(kg·s). Równocześnie zaś z danych geologicznych, odnoszących się do wieku najstarszych zbadanych śladów życia organicznego na Ziemi, wynika, że Słońce dostarczało Ziemi światła i ciepła w tych samych — w przybliżeniu — jak obecnie ilościach przez co najmniej ostatnie kilkaset milionów lat.

energia grawitacyjna gwiazdy

Pierwszej próby zidentyfikowania źródeł energii gwiazd dokonali w połowie XIX w. H. Helmholtz i W. Thomson (Kelvin). Zwrócili oni uwagę na zasoby energii grawitacyjnej gwiazdy: w wypadku stałego, powolnego kurczenia się gwiazdy, jej energia potencjalna przechodzi w energię cieplną, pokrywając przy tym straty powstające przez wypromieniowanie. Jeżeli gwiazda znajduje się w stanie równowagi mechanicznej, wtedy jej energia wewnętrzna (cieplna) U, energia potencjalna Ω oraz energia całkowita $E = U+\Omega$ spełniają związki:

$$E = C_1\Omega, \quad \Omega = -C_2\frac{GM^2}{R},$$

gdzie G — stała grawitacji, M — masa gwiazdy, R — promień gwiazdy, zaś C_1 i C_2 — stałe rzędu 1, zależne — odpowiednio — od stanu materii i rozkładu gęstości we wnętrzu gwiazdy. Korzystając z powyższych zależności można oszacować, jaka była całkowita ilość energii wypromieniowanej przez kurczące się pra-Słońce przy przejściu od konfiguracji początkowej, o nieskończenie dużych rozmiarach ($\Omega = 0$, $E = 0$), do stanu obecnego. Otrzymujemy w przybliżeniu $\Delta E = 5\cdot10^{41}$ J. Jeżeli przyjmie się, że w trakcie procesu kurczenia się Słońce promieniowało równie intensywnie jak obecnie, to otrzymuje się, że oszacowane powyżej zasoby energii grawitacyjnej mogły wystarczyć zaledwie na ok. 40 mln lat. Zatem mechanizm Helmholtza–Kelvina, jakkolwiek odgrywa istotną rolę na pewnych szczególnych i krótkotrwałych etapach ewolucji gwiazd, nie może być głównym źródłem ich energii.

Już w latach dwudziestych obecnego stulecia podejrzewano (m.in. J. H. Jeans), że głównym źródłem energii gwiezdnej są reakcje jądrowe, połączone z wydzielaniem znacznych ilości energii, zgodnie z einsteinowską zależnością $E = mc^2$. Datą przełomową był jednak dopiero rok 1938, kiedy to H. A. Bethe wykazał możliwość zamiany wodoru w hel

reakcje jądrowe

we wnętrzach gwiazd w tzw. cyklu węglowo-azotowym. Wreszcie w 1952 r. E. E. Salpeter opisał inny cykl zamiany wodoru w hel (tzw. cykl proton-proton) oraz zwrócił uwagę na tzw. reakcję 3α, polegającą na zamianie helu w węgiel. Wkrótce potem rozwinęły się badania reakcji jądrowych we wnętrzach gwiazd na różnych etapach ich ewolucji, a równocześnie powstały teoria ewolucji gwiazd i teoria powstawania pierwiastków chemicznych (→ Powstawanie pierwiastków chemicznych).

reakcja 3α

niedobór masy

Wydajność energetyczną reakcji jądrowej określa tzw. niedobór masy, tj. różnica mas jąder atomowych stanowiących elementy wejściowe i produkty końcowe reakcji. Na przykład wynikiem cyklu węglowo-azotowego lub cyklu proton-proton jest zamiana 4 protonów w jądro helu; niedobór masy wynosi tu: $4\cdot 1{,}685\cdot 10^{-27} - 6{,}694\cdot 10^{-27} = 4{,}6\cdot 10^{-29}$ kg, zaś odpowiadająca mu ilość energii jest równa w przybliżeniu $4\cdot 10^{-12}$ J. Zamieniając na hel 1 gram materii wodorowej otrzymujemy ok. $6{,}3\cdot 10^{11}$ J. Proces zamiany wodoru w hel ma zatem wydajność energetyczną mogącą zapewnić świecenie Słońca przez czas rzędu $3\cdot 10^{18}$ s = 10^{11} lat (przy założeniu, że zamianie wodoru w hel ulegnie cała masa Słońca). Wydajność reakcji z udziałem jąder helu i jąder cięższych jest znacznie niższa. Na przykład, w procesie 3α wydajność wynosi ok. $6\cdot 10^{10}$ J/g.

Tempo zachodzenia określonej reakcji jądrowej we wnętrzu gwiazdy zależy od lokalnych wartości parametrów fizycznych, głównie temperatury i gęstości, oraz składu chemicznego. Reakcje zachodzą najszybciej, a energia wydzielana jest najobficiej w centralnych częściach gwiazdy, tj. tam, gdzie panuje najwyższa temperatura.

Sytuacja przedstawia się jednak odmiennie, gdy wewnętrzne obszary gwiazdy pozbawione są już określonego paliwa jądrowego: wtedy maksimum zachodzenia reakcji przypada na warstwę pośrednią, stanowiącą przejście do obszarów zewnętrznych gwiazdy, bogatych w paliwo.

Reakcje jądrowe zachodzące we wnętrzu gwiazdy pokrywają wydatek energetyczny gwiazdy na promieniowanie, dzięki czemu zachowana jest równowaga termiczna. Spadek tempa produkcji energii, wywołany np. wyczerpywaniem się zapasów paliwa, wywołuje zachwianie równowagi termicznej, kurczenie się gwiazdy prowadzące do podwyższenia temperatury wewnętrznej i wzrostu tempa produkcji energii, aż do odzyskania równowagi. Po całkowitym wyczerpaniu się zapasów paliwa proces kurczenia się gwiazdy może się stać na krótko jedynym znaczącym źródłem energii oraz może — a nawet musi — prowadzić do wzrostu temperatury i gęstości materii aż do wartości, w których mogą zachodzić reakcje jądrowe z udziałem następnego, obfitego paliwa. Taki autoregulujący mechanizm umożliwia gwieździe kontrolowanie jej zapasów energetycznych oraz przestawianie się z jednego typu paliwa jądrowego na następne.

utrata materii przez gwiazdy

Ważnym zjawiskiem związanym z zaawansowanymi stadiami ewolucji jest proces utraty materii przez gwiazdy, zachodzący bądź przez powolny wypływ, bądź też jednorazowe wyrzucenie w przestrzeń międzygwiazdową znacznej części masy całej gwiazdy, jak np. w przypadku gwiazd supernowych. Utrata materii przez gwiazdę stanowi zwrot tej materii do ośrodka międzygwiazdowego. Ponieważ przynajmniej część wyrzuconej przez gwiazdę materii brała uprzednio udział w procesach jądrowych, zatem skład chemiczny zwracanej do ośrodka międzygwiazdowego materii różni się od pierwotnego składu materii, z której powstała gwiazda. Oznacza to, że ubocznym wynikiem ewolucji gwiazd jest stałe wzbogacanie materii międzygwiazdowej w Galaktyce w produkty wewnątrzgwiezdnych reakcji jądrowych. Jest to jeden z najważniejszych mechanizmów powstawania pierwiastków chemicznych we Wszechświecie.

Powstawanie gwiazd

Proces powstawania gwiazd nie jest zjawiskiem jednorazowym, ale zachodził nieprzerwanie w różnych epokach istnienia → Galaktyki i zachodzi obecnie. Dowodem tego jest obecność w Galaktyce młodych gwiazd, których wiek wynosi ok. 10^6 lat, podczas gdy wiek Galaktyki wynosi ok. $(1\text{–}1{,}5)\cdot 10^{10}$ lat. Przykładem młodych gwiazd są tzw. gwiazdy typu *O*, wyróżniające się bardzo dużymi jasnościami. Gwiazda typu *O*, której masa jest dwudziestokrotnie większa od masy Słońca, wypromieniowywuje 10^5 razy więcej energii niż Słońce; innymi słowy, tempo zużywania zapasów energii (w przeliczeniu średnio na 1 gram materii) jest tu ok. 5000 razy większe niż w przypadku Słońca. Zatem — nawet gdyby cała masa gwiazdy mogła zostać wykorzystana jako paliwo jądrowe — zapasów tych mogłoby wystarczyć zaledwie na kilkadziesiąt milionów lat; z dokładniejszych rozważań wynika, że wiek tych gwiazd nie może być większy od podanego wyżej, tzn. rzędu 10^6 lat. Niezależnym potwierdzeniem tej oceny jest fakt, że gwiazdy typu *O* tworzą w przestrzeni luźne zgrupowania — tzw. asocjacje — które są niestabilne dynamicznie: po upływie czasu rzędu 10^6–10^7 lat muszą ulec rozpadowi, zaś wchodzące w ich skład gwiazdy — ulec wymieszaniu z innymi, starszymi gwiazdami w Galaktyce. Młodymi gwiazdami są także tzw. gwiazdy typu *T* Tauri, występujące również w asocjacjach, często z gwiazdami typu *O*, a z reguły — w obszarach bogatych w materię międzygwiazdową.

asocjacje gwiazd

Gwiazdy powstają z materii międzygwiazdowej. Prawdziwość tego stwierdzenia wynika z jednej strony z braku jakichkolwiek dowodów na istnienie jakiegoś innego „tworzywa przedgwiazdowego", z drugiej zaś — z wielu faktów obserwacyjnych. Oto najważniejsze z nich. Młode gwiazdy występują wyłącznie w obszarach bogatych w materię międzygwiazdową (il. 221, tabl. 59). Obserwuje się je zarówno w dostępnych do obserwacji częściach naszej Galaktyki, jak też i w innych galaktykach; bogate w materię międzygwiazdową galaktyki nieregularne zawierają dużo młodych gwiazd, a ubogie w materię międzygwiazdową galaktyki eliptyczne wyróżniają się równocześnie brakiem gwiazd młodych.

materia międzygwiazdowa

Materia międzygwiazdowa, występująca jako gaz i pył, wypełnia przestrzeń międzygwiazdową w sposób niejednorodny, z wyraźną tendencją do koncentracji w obłokach. Masy obłoków zawarte są w bardzo szerokich granicach, od ok. 1 masy Słońca ($= 1{,}98\cdot 10^{30}$ kg) w przypadku najmniejszych, możliwych do zaobserwowania kondensacji materii, do 10^3–10^4 mas Słońca, w przypadku największych tego typu obiektów w Galaktyce. Jeszcze bardziej zróżnicowane są parametry opisujące fizyczny stan materii w obłokach. O ile średnia gęstość materii międzygwiazdowej w pobliżu płaszczyzny Galaktyki (wliczając w to obłoki i rozrzedzony ośrodek między obłokami) wynosi ok. 10^{-23} g/cm³ (co w przybliżeniu odpowiada kilku atomom wodoru na cm³), to gęstość w typowych obłokach równa jest ok. 10^{-21} g/cm³, zaś w najgęstszych obłokach, z których obserwuje się m.in. promieniowanie radiowe rodników OH, osiągać może wartość rzędu 10^{-14} g/cm³ (co odpowiada w przybliżeniu gęstości 10^{10} atomów wodoru na cm³). Średnia gęstość materii we wnętrzu przeciętnej gwiazdy w początkowym etapie ewolucji jest rzędu 1 g/cm³, co oznacza, że przejściu od średniej gęstości typowego obłoku materii międzygwiazdowej do gęstości wewnątrzgwiezdnej musi odpowiadać zmiana rozmiarów liniowych o czynnik rzędu 10^{-7}.

Obłok materii międzygwiazdowej może kurczyć się pod wpływem własnego przyciągania grawitacyjnego tylko wtedy, gdy jego parametry fizyczne spełniają tzw. kryterium Jeansa. Z kryterium tego wynika, że masa obłoku nie może być mniejsza od:

kryterium Jeansa

$$m_0 = \left(V_s/\sqrt{4\pi G}\right)^3 \varrho^{-1/2},$$

gdzie G — stała grawitacji, V_s — prędkość dźwięku w obłoku, ϱ — gęstość. Przy typowych wartościach parametrów m_0 jest rzędu setek lub nawet tysięcy mas Słońca; oznacza to, że tylko najmasywniejsze obłoki spełniają kryterium Jeansa i tylko one mogą brać udział w procesie tworzenia się gwiazd oraz że zjawiska prowadzące do powstawania masywnych obłoków są pierwszym ogniwem procesu powstawania gwiazd.

Kurczenie się obłoku materii międzygwiazdowej jest procesem złożonym, nie dającym się opisać za pomocą prostego modelu. Wzrost gęstości i temperatury sprzyjają tworzeniu się złożonych rodników i cząsteczek; z tą fazą procesu identyfikuje się supergęste obłoki, będące źródłami emisji radiowej rodników OH i in. Wzrost gęstości powoduje, że lokalne zagęszczenia w obłoku zaczynają spełniać kryterium Jeansa, co prowadzi do rozpadu pierwotnego obłoku na wiele kurczących się zgęszczeń, których masy mogą już być porównywalne z masami gwiazd; zagęszczenia takie można uważać za protogwiazdy. W tym stadium stan fizyczny materii różni się znacznie od stanu początkowego, tj. takiego, jaki panuje w przeciętnym obłoku materii międzygwiazdowej; w wyniku zgęszczania się i ogrzewania powstają nie tylko cząsteczki, ale także i pył. Gdy na dalszym etapie rozwoju jedna (lub więcej) z protogwiazd staje się gorącą gwiazdą typu widmowego *O* lub *B*, ciśnienie jej promieniowania spowoduje „wydmuchanie" pyłu w przestrzeń międzygwiazdową. Naszkicowany tu mechanizm może tłumaczyć pochodzenie pyłu w materii międzygwiazdowej.

moment obrotowy gwiazdy

Ważnym czynnikiem w procesie powstawania gwiazd jest występowanie momentu obrotowego. Z prawa zachowania momentu pędu wynika, że w procesie kurczenia się pierwotnego obłoku, gdy jego rozmiary zmieniają się o czynnik rzędu 10^{-7} (por. wyżej), nawet chaotyczne ruchy materii wewnątrz obłoku z prędkościami rzędu kilku cm/s mogłyby doprowadzić do powstania szybkorotujących gwiazd; istnienie zaś bardziej systematycznego (choć nawet bardzo powolnego) ruchu obrotowego obłoku wykluczałoby powstanie jakiejkolwiek stabilnej konfiguracji gwiazdopodobnej, ze względu na nadmiar momentu obrotowego. Kurcząca się protogwiazda może pozbyć się nadmiaru momentu obrotowego albo przekazując go otaczającej ją materii, albo pozbywając się części własnej materii. Pierwszy z tych procesów może zachodzić, gdy protogwiazda ma pole magnetyczne, którego linie sił — „wmrożone" w otaczający ośrodek — są elementem sprzęgającym hamowaną protogwiazdę i otaczającą ją materię. W drugim wypadku, wyrzucana jest ta część jej materii, dla której siła odśrodkowa nie może już być równoważona siłami grawitacji; jeżeli protogwiazda już ma wyraźną symetrię osiową, to wyrzucane są najszybciej obracające się, równikowe części konfiguracji. W związku z tym uważa się, iż „produktem ubocznym" powstawania gwiazd są układy planetarne.

Charakterystyczną cechą materii protogwiazdy jest duża przezroczystość dla promieniowania podczerwonego (→ Astronomia w podczerwieni). Wydzielona we wnętrzu konfiguracji energia jest więc niemal w całości wypromieniowywana na zewnątrz, co sprawia, że kurczenie się konfiguracji nie powoduje znaczniejszego wzrostu temperatury ani ciśnienia gazu. Kurczenie się ma więc charakter swobodnego spadku, który może zostać zahamowany dopiero przez wzrost ciśnienia. Następuje to wtedy, gdy w ostatniej fazie swobodnego spadku temperatura we wnętrzu konfiguracji osiąga od kilku do kilkunastu tys. stopni, rośnie gwałtownie stopień jonizacji materii i jej nieprzezroczystość, wzrasta też ciśnienie osiągając wartości od kilkudziesięciu do kilkuset barów. Od tego momentu dalsze kurczenie się będzie zachodziło przy zachowaniu równowagi mechanicznej: ciśnienie gazu w danej warstwie musi równoważyć ciężar warstw położonych wyżej. Można przyjąć, że w tym momencie dotychczasowa protogwiazda staje się gwiazdą. Rozpoczynający się w tym momencie wstępny etap ewolucji gwiazdy nazywa się etapem kontrakcji (kurczenia się) lub fazą przed ciągiem głównym. Na początku tego etapu w całej gwieździe występuje konwekcja, czyli mieszanie się materii; jest ona głównym mechanizmem transportującym energię z coraz gorętszego wnętrza gwiazdy na zewnątrz. Później konwekcja we wnętrzu gwiazdy zanika i w gwieździe można rozróżnić: wnętrze (w równowadze promienistej), w którym energia przenoszona jest wyłącznie przez promieniowanie, oraz zewnętrzną otoczkę, w której nadal panuje konwekcja.

kontrakcja gwiazdy

Kurczenie się gwiazdy ulega chwilowemu zahamowaniu, gdy w jej wnętrzu osiągnięta zostaje temperatura wystarczająco wysoka, aby mogły zajść reakcje jądrowe z udziałem tzw. lekkich pierwiastków: litu, berylu i boru, a także ciężkiego izotopu wodoru — deuteru. Na przykład deuter ulega spalaniu w reakcji:

$$D^2+H^1 \rightarrow He^3+\gamma+5{,}49\ \text{MeV},$$

co następuje już w temperaturach powyżej 0,5 mln stopni. Ponieważ obfitości tych pierwiastków są znikomo małe, ich wypalanie trwa krótko i nie wpływa na dalszą ewolucję gwiazdy.

Etap kontrakcji trwa — w zależności od masy gwiazdy — od ok. 100 tys. lat (dla gwiazd najmasywniejszych) do kilkudziesięciu milionów lat (dla gwiazd takich jak Słońce) i tylko dla najmniej masywnych gwiazd (rzędu 0,1 masy Słońca) może być porównywalny z wiekiem Galaktyki (ok. $(1\text{–}1{,}5)\cdot10^{10}$ lat).

Stadium spalania wodoru i helu

Etap kontrakcji kończy się, gdy w centralnych częściach gwiazdy temperatura osiąga ok. 10 mln stopni. W tym momencie rozpoczynają się reakcje jądrowe z udziałem najobfitszego i najwydatniejszego paliwa, jakim jest wodór. W gwiazdach o masach rzędu 1 masy Słońca spalanie wodoru zachodzi w tzw. cyklu protonowym:

cykl protonowy

$$H^1+H^1 \rightarrow D^2+e^++\nu+1{,}44\ \text{MeV},$$
$$D^2+H^1 \rightarrow He^3+\gamma+5{,}49\ \text{MeV},$$
$$He^3+He^3 \rightarrow He^4+H^1+H^1+12{,}85\ \text{MeV}.$$

(W temperaturach powyżej 16 mln stopni cykl protonowy kończy się innymi reakcjami niż podana powyżej, ale i one prowadzą do wyprodukowania He^4 i wyzwolenia energii). W gwiazdach masywnych, o masach rzędu kilku, lub więcej, mas Słońca, kontrakcja ustaje dopiero w chwili osiągnięcia temperatur rzędu 20 mln stopni, w których wodór spala się w tzw. cyklu węglowo-azotowym:

cykl węglowo-azotowy

$$C^{12}+H^1 \rightarrow N^{13}+\gamma+1{,}95\ \text{MeV},$$
$$N^{13} \rightarrow C^{13}+e^++\nu+2{,}22\ \text{MeV},$$
$$C^{13}+H^1 \rightarrow N^{14}+\gamma+7{,}54\ \text{MeV},$$
$$N^{14}+H^1 \rightarrow O^{15}+\gamma+7{,}35\ \text{MeV},$$
$$O^{15} \rightarrow N^{15}+e^++\nu+2{,}71\ \text{MeV},$$
$$N^{15}+H^1 \rightarrow C^{12}+He^4+4{,}96\ \text{MeV}.$$

Ten cykl polega więc także na zamianie 4 protonów na jądro helu oraz wydzielaniu energii.

wiek zero gwiazdy

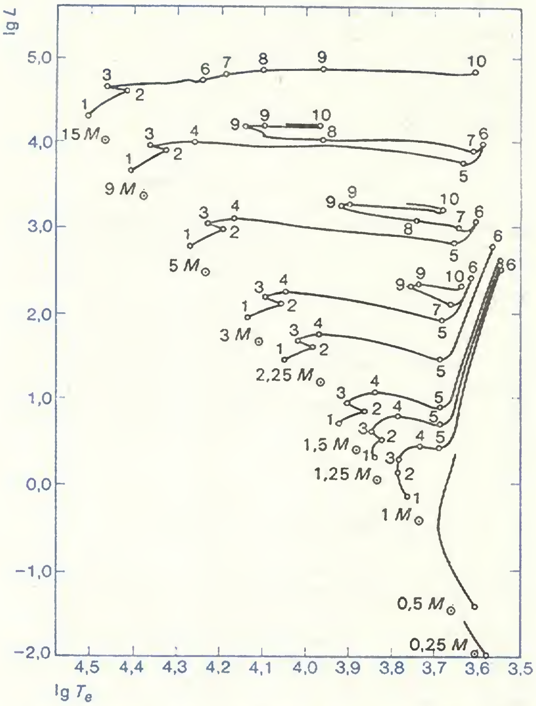
Moment zakończenia się procesu kurczenia i rozpoczęcia we wnętrzu gwiazdy reakcji spalania wodoru określany jest umownie jako wiek zero. Gwiazda wieku zero jest jednorodna chemicznie. Ma ona jasność i rozmiary określone jednoznacznie jej masą i składem chemicznym. Wyniki obliczeń teoretycznych dotyczących ewolucji gwiazd przedstawia się na

wykres H–R wykresie jasność absolutna–temperatura efektywna, którego obserwacyjnym odpowiednikiem jest wykres jasność absolutna–barwa (lub typ widmowy), czyli tzw. wykres Hertzsprunga–Russella (H–R). Rysunek 1 pokazuje wyniki obliczeń ewolucyjnych dla gwiazd o różnych masach, od 0,25 do 15 mas Słońca. Punkty oznaczone numerem 1 odpowiadają gwiazdom o wieku zero; punkty te tworzą tzw. ciąg główny wieku zero na wykresie H–R.

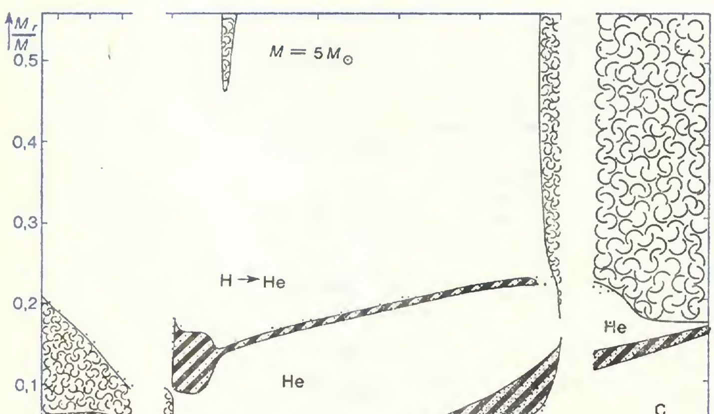
etapy początkowe ewolucji

Początkowe etapy ewolucji prześledzimy na przykładzie gwiazdy o masie 5 mas Słońca (rys. 1 i 2). Pierwszy, najdłuższy etap ewolucji polega na spalaniu się wodoru w centralnych częściach gwiazdy. Z wyjątkiem gwiazd o małych masach, spalanie odbywa się w cyklu węglowo-azotowym, przy czym ilość wyzwalanej energii jest tak duża, że jedynym dostatecznie wydajnym sposobem przekazywania tej energii na zewnątrz jest gwałtowne mieszanie się (konwekcja) materii w centralnych częściach gwiazdy, tworzących tzw. jądro konwektywne. Spalanie się wodoru oraz mieszanie się materii sprawia, że w całym obszarze jądra obfitość wodoru maleje, obfitość zaś helu rośnie. Gdy wyczerpują się zapasy paliwa w jądrze, gwiazda gwałtownie się kurczy (przejście od punktu 2 do 3 na rys. 1), dopalają się resztki wodoru w jądrze i zaczyna się spalać wodór w warstwie pośredniej, na granicy pomiędzy pozbawionym już wodoru jądrem a bogatymi w wodór warstwami zewnętrznymi. Warstwa ta jest początkowo dość gruba, ale wkrótce — wskutek wypalania się wodoru — jej wewnętrzne części dołączają się do bezwodorowego jądra. Wyprodukowanie dostatecznych ilości energii w cienkiej warstwie wymaga odpowiednio wysokiej temperatury; wzrost temperatury pochodzi stąd, że wewnętrzne części gwiazdy kurczą się, podczas gdy równocześnie warstwy zewnętrzne — rozszerzają się.

jądro konwektywne



Rys. 1. Zmiany ewolucyjne gwiazd o różnych masach przedstawione na tzw. wykresie Hertzsprunga–Russella. L jest jasnością absolutną gwiazdy w jednostkach Słońca, T_e temperaturą powierzchni gwiazdy. Numery 1–10 oznaczają kolejne stadia ewolucji. Słońce znajduje się obecnie w punkcie 2 na torze ewolucyjnym gwiazdy o 1 masie Słońca



Rys. 2. Zmiany w czasie struktury gwiazdy o masie 5 mas Słońca. Na osi pionowej — masa w jednostkach całej gwiazdy; $M_r/M = 1$ na powierzchni gwiazdy. Grube, poprzeczne kreski oznaczają obszary, w których zachodzą reakcje jądrowe; cienkie, zawirowane kreski — obszary konwekcji; punkty — obszary, w których skład chemiczny jest wynikiem zachodzenia reakcji jądrowych

czerwone olbrzymy

Przejściu od punktu 4 do punktu 5 (na rys. 1) odpowiada przemiana gwiazdy ciągu głównego w czerwonego olbrzyma — gwiazdę o ok. dziesięciokrotnie większych rozmiarach i 3–4-krotnie niższej temperaturze efektywnej. Na tym etapie w zewnętrznych obszarach gwiazdy rozpoczyna się mieszanie materii — tworzy się rozległa tzw. zewnętrzna warstwa konwektywna. Jest ona charakterystyczną cechą wszystkich chłodnych gwiazd, a zwłaszcza olbrzymów i nadolbrzymów. Ilekroć gwiazda, w trakcie ewolucji, staje się olbrzymem, w jej zewnętrznych warstwach pojawia się natychmiast konwekcja. W niektórych wypadkach konwekcja przenika na tyle głęboko, że może spowodować wydostanie się na powierzchnię gwiazdy produktów reakcji jądrowych. W stadium czerwonego olbrzyma temperatura centralnych części gwiazdy osiąga wartość rzędu 100 mln stopni. Wtedy rozpoczyna się proces spalania helu w węgiel w reakcji 3α:

zewnętrzna warstwa konwektywna

$$3He^4 \rightarrow C^{12} + 7{,}3\ MeV.$$

(Na rys. 1 odpowiada temu punkt 6). Od tej chwili w gwieździe istnieją dwa źródła energii: centralne — helowe, i warstwowe — wodorowe. Spalanie helu odbywa się podobnie jak spalanie wodoru. Początkowo zachodzi ono w centralnym jądrze konwektywnym, w którym maleje obfitość helu, a wzrasta obfitość węgla. Wkrótce (rys. 2) tworzy się jądro węglowe (pozbawione zarówno wodoru, jak i helu), na granicy którego hel spala się w źródle warstwowym. Tej fazie odpowiada punkt 10 na rys. 1. Zmianom we wnętrzu gwiazdy od momentu zapalenia się helu do tworzenia jądra węglowego towarzyszą zmiany struktury warstw zewnętrznych. Rozmiary gwiazdy ulegają najpierw zmniejszeniu, a potem ponownemu zwiększeniu, tak że na wykresie H–R gwiazda zatacza charakterystyczne pętle (rys. 1). W tym stadium gwiazda przechodzi kilkakrotnie przez pas niestabilności pulsacyjnej (→ Gwiazdy zmienne pulsujące) i staje się wtedy gwiazdą zmienną — cefeidą.

cefeidy

tempo ewolucji gwiazd

Dla gwiazd o różnych masach różne jest przede wszystkim tempo procesów ewolucyjnych. Wynika to z prostych rozważań dotyczących bilansu energetycznego gwiazdy. Tempo wypromieniowywania energii, czyli jasność absolutna L, jest silnie zależna od masy m gwiazdy; w przybliżeniu $L = \text{const}\cdot m^n$, gdzie — przykładowo — dla mas rzędu 5 mas Słońca $n = 3$. Zapasy paliwa są oczywiście proporcjonalne do masy gwiazdy. A zatem tempo zużywania paliwa, czyli tempo ewolucji jest w przybliżeniu dane wyrażeniem o postaci: $\text{const}\cdot m^{n-1}$, tzn. zależy silnie od masy gwiazdy. Poniższa tabela podaje czasy ewolucji pomiędzy poszczególnymi punktami na wykresie H–R dla kilku wybranych mas.

Druga ważna różnica między przebiegiem ewolucji gwiazd o różnych masach wiąże się z występowaniem

Tempo ewolucji gwiazd o różnych masach (masy gwiazd podane są w jednostkach masy Słońca, czas — w jednostkach 10^7 lat)

Punkty	Masa			
	9	5	2,25	1,25
1–2	2,144	6,547	46,02	280,3
2–3	0,060	0,217	1,647	18,24
3–4	0,009	0,137	3,696	104,5
4–5	0,015	0,075	1,310	14,63
5–6	0,006	0,049	3,829	40,0
6–7	0,049	0,605		
7–8	0,010	0,102		
8–9	0,328	0,900		
9–10	0,016	0,093		

różnic gęstości materii i wynikającą stąd odmiennością równania stanu. We wnętrzu typowej gwiazdy ciągu głównego, takiej jak np. Słońce, materia spełnia równanie stanu gazów doskonałych, które przestaje jednak obowiązywać w warunkach dużej gęstości materii (przykładowo: powyżej 100 g/cm^2 przy temperaturze ok. 10^6 K, powyżej 10^6 g/cm^3 przy temperaturze ok. 10^9 K itd). Powodem tego jest tzw. zwyrodnienie elektronów wchodzących w skład materii gwiezdnej. Zgodnie z zakazem Pauliego, zwyrodniałe elektrony mają (średnio) większe prędkości niżby to wynikało z panującej w materii temperatury. Ciśnienie elektronów jest znacznie większe od ciśnienia jonów i ono określa ciśnienie gazu. Równanie stanu takiej materii wiąże ze sobą tylko gęstość i ciśnienie, nie zawiera zaś temperatury; np. gdy gęstość wynosi poniżej 10^7 g/cm^3 mamy $p \sim \varrho^{5/3}$, gdy zaś powyżej — $p \sim \varrho^{4/3}$. W stadium ciągu głównego efekty zwyrodnienia są istotne tylko w wypadku gwiazd o skrajnie małych masach, poniżej 0,5 masy Słońca. W trakcie ewolucji, gdy wskutek kurczenia się gwiazdy rośnie w jej wnętrzu zarówno gęstość, jak i temperatura, może w pewnym momencie dojść do zwyrodnienia (elektronów). W gwiazdach o masach mniejszych od ok. 2 mas Słońca następuje to już wkrótce po wypaleniu się wodoru w centralnych częściach gwiazdy. Dalsze kurczenie się bezwodorowego jądra, konieczne dla uzyskania odpowiednio wysokiej temperatury w warstwie spalania wodoru, powoduje dalszy wzrost gęstości i „pogłębianie" się zwyrodnienia. Rośnie jednak także temperatura tego zwyrodniałego, bezwodorowego jądra. W pewnym momencie zostaje osiągnięta temperatura wystarczająca do zapalenia się helu. I tu zjawisko przebiega zupełnie inaczej niż w gwiazdach masywnych, w których wnętrzach elektrony nie są zwyrodniałe. Gdy materia spełnia równanie stanu gazów doskonałych, jej ogrzanie wskutek pojawienia się nowych źródeł energii powoduje wzrost ciśnienia i natychmiastowe rozkurczenie się konfiguracji, a w konsekwencji spadek gęstości i temperatury. Ustala się szybko stan równowagi z uwzględnieniem nowych źródeł energii. Gdy materia składa się m.in. ze zwyrodniałych elektronów, wzrost temperatury nie wywołuje rozkurczenia się konfiguracji, rosnąca temperatura powoduje natomiast gwałtowny wzrost wydajności reakcji jądrowych i trwa to dopóty, dopóki temperatura nie osiągnie wartości, przy której (przy danej gęstości) elektrony przestają być zwyrodniałe; wtedy dopiero włącza się samoregulujący mechanizm opisany powyżej i dalszy przebieg reakcji jądrowych ma charakter kontrolowany. Zapalenie się helu w gwiazdach o masach poniżej ok. 2 mas Słońca ma więc przebieg gwałtowny; nosi ono nazwę błysku helowego.

zwyrodnienie elektronów

błysk helowy

Późne stadia ewolucji

Po wypaleniu się helu w centralnych częściach gwiazdy, jej strukturę można opisać następująco: We wnętrzu gwiazdy istnieje jądro węglowe, o stale rosnącej masie; w jądrze tym, wskutek wcześniejszego zachodzenia reakcji jądrowych, występuje także znaczna obfitość tlenu, stąd nazwa — jądro węglowo-tlenowe. Energia produkowana jest w dwu warstwach, w których pali się wodór i hel, a które w miarę zużywania paliwa przemieszczają się na zewnątrz. W tym stadium jasność gwiazdy zależy niemal wyłącznie od masy zawartej w węglowo-tlenowym jądrze:

jądro węglowo-tlenowe

$$L/L_{\odot} = 59250\,(M_{\text{jądra}}/M_{\odot} - 0{,}522),$$

gdzie $L/L_{\odot}$ — jasność w jednostkach jasności Słońca, $M_{\text{jądra}}/M_{\odot}$ — masa jądra w jednostkach słonecznych. W gwiazdach o masach mniejszych od 8 mas Słońca z biegiem czasu (tj. wskutek wzrostu gęstości) w węglowo-tlenowym jądrze następuje zwyrodnienie elektronów, tak jak w przypadku helowych jąder gwiazd o małych masach.

Na tym etapie ewolucji gwiazda, niezależnie od jej masy, jest chłodnym olbrzymem lub nadolbrzymem. Zarówno dane obserwacyjne, jak i rozważania teoretyczne dotyczące stabilności takich gwiazd wskazują, że na tym etapie ważnym czynnikiem jest utrata materii przez gwiazdę.

chłodne olbrzymy

Obserwacje dostarczają wielu przykładów utraty materii przez gwiazdy znajdujące się na zaawansowanych etapach ewolucji. Mgławice planetarne (tabl. 59, il. 222) są szczególnym przykładem tego zjawiska, ponieważ wyrzucona przez gwiazdę otoczka gazowa świeci na ogół bardzo silnie, co pozwala na szczegółowe badanie jej struktury. Gwiazdy zmienne długookresowe typu Mira tracą materię w tempie 10^{-6} mas Słońca na rok. Tak szybka utrata materii jest charakterystyczna dla wszystkich olbrzymów i nadolbrzymów późnych typów widmowych.

Dalszy przebieg ewolucji gwiazdy zależy zatem od rozmiarów i parametrów fizycznych jądra węglowo-tlenowego oraz od tempa utraty materii. Jakkolwiek szczegóły opisu pozostają jeszcze niepewne, to wyodrębnić można trzy typy ewolucji, w zależności od początkowej masy gwiazdy.

Gwiazdy o masach małych, tj. mniejszych od ok. 1,4 masy Słońca, nigdy nie osiągają stanu, w którym temperatura jądra węglowo-tlenowego byłaby wystarczająco duża do zachodzenia reakcji spalania węgla. Końcowe etapy ewolucji takich obiektów polegają więc na przesuwaniu się ku powierzchni i stopniowym wygasaniu warstw spalania wodoru i helu. Równocześnie wskutek niestabilności dynamicznej warstw zewnętrznych ulegają one wyrzuceniu, najprawdopodobniej w postaci mgławicy planetarnej. Gwiazda kończy ewolucję jako stygnący biały karzeł. Podobny przebieg ma ewolucja gwiazd o początkowych masach zawartych w przedziale od 1,4 do ok. 3 mas Słońca, w przypadku których następuje utrata znacznej części masy początkowej, tak że temperatura jądra węglowo-tlenowego nie osiąga wartości krytycznej „zapłonu" węgla oraz końcowa masa gwiazdy jest mniejsza od 1,4 masy Słońca.

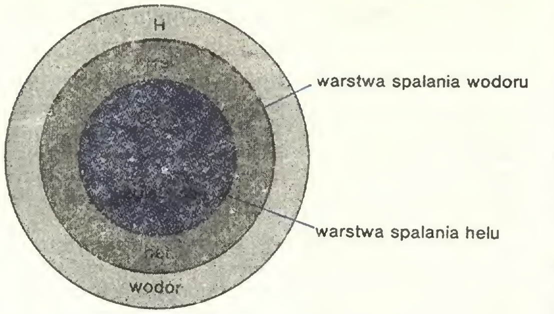
białe karły

Białe karły, będące jednym z możliwych produktów końcowych ewolucji gwiazd, mają gęstości materii w granicach 10^4–10^8 g/cm^3. W tych gęstościach następuje zwyrodnienie elektronów i obowiązuje równanie stanu zwyrodniałego gazu elektronowego (por. wyżej). Gwiazda zbudowana z takiej materii może tworzyć konfigurację stabilną jedynie wtedy, gdy jej masa jest mniejsza od pewnej wartości krytycznej, będącej funkcją składu chemicznego. Jeśli materia jest pozbawiona wodoru, to wartość krytyczna nosi nazwę granicy Chandrasekhara i wynosi ok. 1,4 masy Słońca. Istnienie tej granicy oznacza, że gwiazda mająca zakończyć swą ewolucję jako biały karzeł musi — przynajmniej w końcowym stadium ewolucji — mieć masę mniejszą od granicy Chandrasekhara.

granica Chandrasekhara

Gwiazdy o masach pośrednich, tj. zawartych w granicach od ok. 1,4 do ok. 8 mas Słońca, ewoluują wg podobnego schematu. Początkowo energia produkowana jest w dwu warstwach spalających wodór

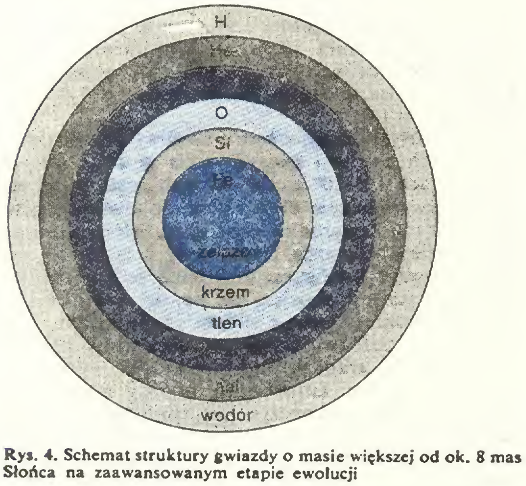
i hel, przy równoczesnym powiększaniu się masy jądra węglowo-tlenowego (rys. 3). W jądrze tym zaczynają się pojawiać neutrina powstające w procesach zachodzących dzięki istnieniu uniwersalnych oddziaływań Fermiego (np. tzw. fotoneutrina pow-



Rys. 3. Schemat struktury gwiazdy o masie mniejszej od ok. 8 mas Słońca, przed zapaleniem się węgla w jądrze

stające w procesie: $\gamma+e^- \rightarrow e^-+\nu+\bar{\nu}$). Neutrina te wynoszą na zewnątrz energię i powodują chłodzenie się jądra. Wskutek kurczenia się jądra, w warstwie, w której spala się hel, utrzymuje się wysoka temperatura i gęstość, co prowadzi do zwyrodnienia elektronów. Gwiazdy z przedziału mas od ok. 3 do ok. 8 mas Słońca ewoluują podobnie, zmierzając do konfiguracji, w której jądro węglowo-tlenowe ma masę równą ok. 1,4 masy Słońca, zaś gęstość i temperatura centralna są równe — odpowiednio — ok. $2 \cdot 10^9$ g/cm³ i ok. $3 \cdot 10^8$ K. W tych warunkach następuje zapalenie się węgla, przy czym — ze względu na zwyrodnienie — zjawisko to ma charakter gwałtowny, jeszcze gwałtowniejszy niż w przypadku błysku helowego. Proces zapalenia się węgla prowadziłby do rozerwania całej gwiazdy, gdyby nie działały jakieś dodatkowe procesy chłodzenia. Ilość wyzwalanej energii oraz częstość zachodzenia takich zjawisk w Galaktyce są zgodne z danymi obserwacyjnymi odnoszącymi się do gwiazd supernowych. Z drugiej strony wiadomo jednak, że pozostałościami po wybuchach supernowych są gwiazdy neutronowe, co nie pozwala na przyjęcie modelu, w którym rozerwaniu ulega cała konfiguracja. Równocześnie wiadomo, że w cyklu reakcji zapoczątkowywanych zapaleniem się węgla produkowane są znaczne ilości żelaza; gdyby wszystkie gwiazdy o masach od 3 do 8 mas Słońca ulegały — na tym etapie ewolucji — całkowitemu rozpyleniu

błysk węglowy



Rys. 4. Schemat struktury gwiazdy o masie większej od ok. 8 mas Słońca na zaawansowanym etapie ewolucji

w przestrzeń międzygwiazdową, to obfitość żelaza w materii międzygwiazdowej musiałaby być wielokrotnie wyższa od obserwowanej. Wyjście z tego dylematu musi polegać na poszukiwaniu odpowiedniego mechanizmu chłodzenia, łagodzącego przebieg błysku węglowego. Sądzi się, że chłodzenie jest konsekwencją tzw. procesów Urca, polegających na tym, że w gazie zawierającym zwyrodniałe elektrony, przy określonej temperaturze może ustalić się równowaga między dwoma procesami: rozpadem β, polegającym na przemianie jądra o ładunku Z w jądro o ładunku $Z+1$ przy równoczesnej emisji elektronu i antyneutrina, oraz odwrotnym rozpadem β, polegającym na wychwycie elektronu przez jądro o ładunku $Z+1$ i jego przemianie w jądro o ładunku Z przy równoczesnej emisji neutrina. Zatem więc produkowane w takich procesach neutrina i antyneutrina, unosząc energię na zewnątrz, mogą skutecznie chłodzić jądro węglowo-tlenowe.

procesy Urca

Gdy występuje chłodzenie, błysk węglowy niekoniecznie powoduje rozerwanie całej konfiguracji. Jeżeli chłodzenie jest dostatecznie wydajne, to gęstość centralna może osiągnąć wartość rzędu 10^{10} g/cm³, przy której wewnętrzne obszary jądra stają się dynamicznie niestabilne. Wybuch supernowej może zatem polegać na wyzwoleniu znacznych ilości energii i wyrzuceniu w przestrzeń warstw zewnętrznych gwiazdy, w tym zewnętrznych obszarów jądra, ale przy równoczesnym „zapadnięciu" się wewnętrznych obszarów jądra. Produktem wybuchu jest więc rozszerzająca się otoczka gazowa oraz gwiazda neutronowa (→ Pulsary).

gwiazdy neutronowe

Gwiazdy masywne, tj. o masach większych od ok. 8 mas Słońca, odznaczają się na tyle wysokimi temperaturami centralnymi, a równocześnie względnie małymi gęstościami centralnymi, że zapalenie się węgla — przy braku zwyrodnienia — ma przebieg łagodny. Po zapaleniu się i wypaleniu węgla następuje zapalanie się i wypalanie następnych pierwiastków. Ewolucja prowadzi do konfiguracji, której jądro jest złożone z żelaza, zaś kolejne koncentryczne warstwy — z coraz lżejszych pierwiastków, aż do bogatej w wodór warstwy powierzchniowej (rys. 4). Przebieg ewolucji zależy od procesów mieszania materii we wnętrzu gwiazdy oraz procesów utraty materii z warstw zewnętrznych. Jeżeli chodzi o ewolucję żelaznego jądra, to oprócz wzrostu masy jądra zachodzi także wzrost jego temperatury centralnej i gęstości. Ten etap ewolucji kończy się w momencie osiągnięcia temperatury krytycznej, w której jądra żelaza ulegają rozpadowi na jądra helu i neutrony. W tym momencie jądro gwiazdy się zapada i przekształca bądź w gwiazdę neutronową, bądź w czarną dziurę (→ Czarne dziury i zapadanie grawitacyjne). Otoczka gwiazdy ulega wyrzuceniu na zewnątrz. Ponieważ jednak większa część materii z warstw zewnętrznych uległa wyrzuceniu już wcześniej, przeto opisany tu proces końcowy nie może być utożsamiany z wybuchem supernowej.

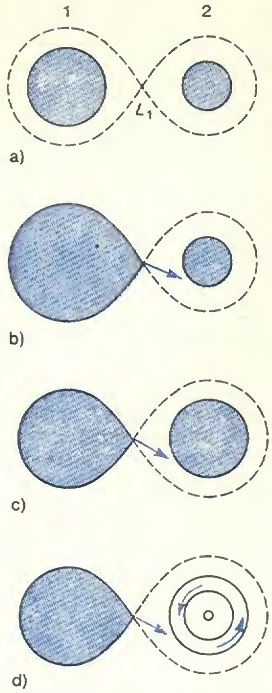
jądro żelazne

czarne dziury

Ewolucja gwiazd — składników układów podwójnych przebiega identycznie jak gwiazd pojedynczych tylko do momentu, w którym rozpoczyna się wymiana materii między składnikami. Przyczyna tego zjawiska jest następująca. Kształt gwiazdy spełniającej warunek równowagi hydrostatycznej opisują powierzchnie jednakowego potencjału, które dla gwiazdy pojedynczej są koncentrycznymi sferami; jedna z nich określa powierzchnię gwiazdy. Gdy gwiazda wchodzi w skład układu podwójnego, powierzchnie jednakowego potencjału określane są siłami przyciągania grawitacyjnego obydwu składników oraz siłą odśrodkową. Szczególne znaczenie ma tzw. granica Roche'a, która jest największą, nie wspólną dla obydwu gwiazd, powierzchnią jednakowego potencjału (rys. 5a). Gwiazda, której rozmiary są mniejsze od granicy Roche'a, różni się nieznacznie — pod względem budowy i ewolucji — od gwiazdy pojedynczej. Gdy gwiazda wypełni granicę Roche'a (rys. 5b) następuje wypływ materii przez tzw. wewnętrzny punkt Lagrange'a L_1 (rys. 5a) i jej przepływ do drugiego składnika. Proces ten trwa od 10^4 do 10^6 lat. W tej fazie gwiazda tracąca materię

układy podwójne

granica Roche'a



Rys. 5. Schemat wczesnych stadiów ewolucji ciasnego układu podwójnego (a-c). Linią przerywaną zaznaczona jest granica Roche'a. Strzałki na rys. b) i c) oznaczają przepływ materii z gwiazdy 1 do gwiazdy 2. Rys. d — ciasny układ podwójny typu nowej. Składnik 2 jest białym karłem i wokół niego wiruje gazowy dysk. W przypadku układów będących źródłami promieniowania rentgenowskiego składnik 2 jest gwiazdą neutronową

przekazuje drugiej gwieździe znaczną część swojej pierwotnej masy. Po zakończeniu tego procesu (rys. 5c) gwiazda wypełniająca powierzchnię Roche'a jest mniej masywna od drugiej gwiazdy układu oraz różni się od jakiejkolwiek gwiazdy pojedynczej na dowolnym etapie ewolucji zarówno budową wewnętrzną (występowanie obszarów o różnym składzie chemicznym) jak i parametrami zewnętrznymi (jasność, temperatura efektywna). Ścisły, modelowy opis przebiegu naszkicowanego tu procesu pozwala na wyjaśnienie obserwowanych własności układów podwójnych typu Algola oraz pochodzenia ewolucyjnego tych układów.

Dalsze etapy ewolucji układów podwójnych nie zostały dotąd dokładnie zbadane. Rozważania jakościowe dotyczące przebiegu ewolucji obydwu składników oraz związanych z tym procesów wymiany i utraty materii oraz momentu pędu przez układ pozwalają na wytłumaczenie pochodzenia układów podwójnych, w których jeden ze składników osiągnął końcowy stan ewolucji i jest białym karłem lub gwiazdą neutronową. Drugi składnik znajduje się nadal na wcześniejszym etapie ewolucji, wypełnia granicę Roche'a i przekazuje materię składnikowi pierwszemu. Rozmiary białego karła lub gwiazdy neutronowej są bardzo małe w porównaniu z rozmiarami układu, wskutek czego struga materii, mająca znaczny moment pędu względem gwiazdy docelowej, nie spada bezpośrednio na jej powierzchnię, ale musi utworzyć wokół niej wirujący dysk (rys. 5d). Tego typu gazowe dyski są charakterystyczną cechą układów podwójnych typu nowych oraz układów będących źródłami promieniowania rentgenowskiego (→ Astronomia promieni X i γ).

gazowy dysk

I. Iben *Stellar evolution within and off the main sequence*, Ann. Rev. of Astr. and Astrophys. 5, 571 (1967); I. Iben *Post main sequence evolution of single stars*, Ann. Rev. of Astr. and Astrophys. 12, 125 (1974); B. Paczyński *Późne stadia ewolucji gwiazd*, Post. Astr. 21, 9 (1973); R. J. Tayler (ed.) artykuły w tomie: *Late Stages of Stellar; Evolution*, I. A.U. Symposium, No. 66, 1974.

Galaktyki

Michał Różyczka

Wyspy Wszechświata

Nazwa „galaktyka", wywodząca się od gr. wyrazu *gala* (mleko), wprowadzona została w II połowie XVIII w. przez Wiliama Herschela. Z obserwacji Herschela wynikało, że wszystkie gwiazdy (wśród nich i Słońce) rozmieszczone są w skończonej objętości i tworzą w otaczającej pustce „wyspę" materii. Oglądane z Ziemi, czyli z wnętrza wyspy, skupiają się na sferze niebieskiej w świecącym pasie Drogi Mlecznej.

model galaktyki Herschela

W końcu XVIII w. obserwacje astronomiczne sprowadzały się na ogół do pomiarów pozycji ciał należących do Układu Słonecznego. Szczególnie chętnie obserwowano komety, które początkujący obserwatorzy często mylili z mgławicami, obiektami — jak komety — rozmytymi, ale nie zmieniającymi położenia wśród gwiazd. By zmniejszyć prawdopodobieństwo popełniania takich pomyłek, sporządzono w 1784 r. pierwszy katalog mgławic, który ma obecnie znaczenie wyłącznie historyczne. Wprowadzone w nim oznaczenia są jednak powszechnie używane i tak np. widoczna gołym okiem w gwiazdozbiorze Herkulesa gromada kulista znana jest jako M 13, zaś pozostałość po wybuchu supernowej w gwiazdozbiorze Byka — jako M 1.

mgławice

Nawet po ukazaniu się katalogu mgławice nie zyskały popularności wśród astronomów, zajęli się nimi natomiast filozofowie. Według Kanta wszystkie mgławice były Wyspami Wszechświata — tj. niezwykle odległymi zbiorowiskami gwiazd rozdzielonymi pustą przestrzenią. Wyprzedzając Herschela Kant twierdził, iż Słońce jest jedną z gwiazd Drogi Mlecznej. Każda mgławica miała według niego składać się z tylu gwiazd, ile jest skupionych w Drodze Mlecznej i każda powinna dać się rozdzielić na gwiazdy za pomocą odpowiednio dużego teleskopu. Hipoteza Kanta czekała na pełną weryfikację obserwacyjną prawie 150 lat. Tymczasem W. Herschel, który dysponował najpotężniejszym wówczas teleskopem, stwierdził, iż niektóre z mgławic na pewno nie są zbiorowiskami gwiazd. W innych, które udawało mu się rozdzielać na gwiazdy, znajdował tyle zaledwie świecących punktów, ile zawiera drobna cząstka Drogi Mlecznej. Przez cały XIX w. galaktyka pozostała jedyną Wyspą Wszechświata, której istnienie uznawano za dowiedzione.

model Wszechświata Kanta

Nowe jakościowo dane o mgławicach zebrano dopiero kilkadziesiąt lat po śmierci Herschela. W 1845 r. Wiliam Parsons (lord Rosse) zaobserwował w paru mgławicach strukturę spiralną. Czternaście lat później Wiliam Huggins rozpoczął pierwsze obserwacje spektroskopowe ciał niebieskich. Najciekawsze wyniki otrzymano z obserwacji mgławic nie dających się rozdzielić na gwiazdy. Znaleziono wśród nich obiekty o widmach niemal identycznych z widmami zwykłych gwiazd. Zgodnie z jedną z postawionych hipotez — widmo takiej mgławicy miało powstawać w wyniku nakładania się dużej liczby widm bardzo dalekich gwiazd. Zdobyto pierwszy argument przemawiający za tym, iż niektóre z wgławic mogą być Wyspami Wszechświata. Szczególną uwagę zwrócono na odkryte przez lorda Rosse wgławice spiralne, które mając widma podobne do widm gwiazd nie dawały się rozdzielać na gwiazdy, mimo budowania coraz większych teleskopów.

W końcu lat siedemdziesiątych ubiegłego wieku rozpowszechniły się obserwacje fotograficzne. Wykonano na ich podstawie obszerne katalogi mgławic

(NGC, IC), a także kontynuowano badania spektroskopowe. Porównując zdjęcia wykonane w dużych odstępach czasu opisano ruchy gwiazd w galaktyce. Na początku XX w. rozpoczęto badania rozmieszczenia przestrzennego gwiazd, możliwe do przeprowadzenia dzięki nowym metodom wyznaczania odległości. W miarę napływu informacji zmieniały się wyobrażenia o kształcie i rozmiarach galaktyki. Stwierdzono, że nazywane niegdyś mgławicami gromady kuliste i otwarte, które rozdzielił na gwiazdy już Herschel, znajdują się w odległościach nie większych niż odległości pojedynczych gwiazd. W 1918 r. Harlow Shapley przedstawił kolejny model galaktyki, sporządzony na podstawie badań rozmieszczenia przestrzennego gromad kulistych. Wyznaczona przez Shapleya wartość średnicy galaktyki była dziesięciokrotnie większa od przyjmowanych poprzednio, a Słońce, które we wszystkich dotychczasowych modelach umieszczano w centrum galaktyki, znalazło się prawie na jej skraju. Próby przeniesienia wyobrażeń o galaktyce na mgławice skończyły się niepowodzeniem. Kilka lat przed ukazaniem się pracy Shapleya wykonano w jednym z obserwatoriów serię zdjęć paru mgławic spiralnych. Popełniając wykryty znacznie później błąd zmierzono na nich zerowe w rzeczywistości przesunięcia kątowe i wyznaczono prędkości kątowe obrotu mgławic. Po przemnożeniu ich przez promień galaktyki Shapley otrzymał prędkości liniowe obrotu wielokrotnie większe od prędkości światła. Założenie, że mgławice spiralne są Wyspami Wszechświata podobnymi do galaktyki, prowadziło zatem do absurdu. Shapley musiał je odrzucić i uznać galaktykę za twór unikalny.

model galaktyki Shapleya

Problem mgławic doczekał się ostatecznego rozwiązania dopiero kilka lat później, gdy Edwin Hubble, za pomocą 2,5-metrowego teleskopu na Mount Wilson, zdołał rozdzielić mgławicę spiralną M 31 na gwiazdy i wyznaczyć ich odległości. Okazało się, że M 31 (Mgławica Andromedy) leży daleko poza granicami naszej Galaktyki. Z następnych prac Hubble'a wynikało, że obiektami pozagalaktycznymi są wszystkie mgławice, które dają się ułożyć w pewien ciąg typów morfologicznych i których widma podobne są do widm zwykłych gwiazd. Ciąg ów znany jest obecnie jako ciąg Hubble'a, natomiast zaproponowana przez niego nazwa mgławice pozagalaktyczne nie przyjęła się. Obiekty te nazywa się obecnie po prostu galaktykami. Naszą Galaktykę wyróżnia się pisząc ją wielką literą bądź nazywając Układem Drogi Mlecznej. Termin „mgławice" zachowany został do oznaczania gazowych i pyłowych obłoków materii międzygwiazdowej znajdujących się w Galaktyce.

ciąg Hubble'a

Układ Drogi Mlecznej

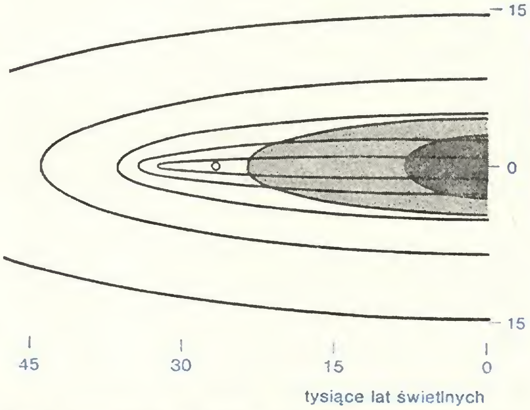
Najsilniej świecąca część Układu Drogi Mlecznej przypomina kształtem silnie spłaszczony dysk. Nie można jednak określić granicy układu: Galaktyka przechodzi w sposób ciągły w przestrzeń międzygalaktyczną. Za umowną granicę Galaktyki, obejmującą wszystkie podsystemy, można przyjąć powierzchnię kuli o średnicy ok. 200 000 lat świetlnych (l.św.). Średnica świecącego dysku nie przekracza 90 000 l.św., a największa grubość — 10 000 l.św. Znajduje się w nim ok. 10^{12} gwiazd, a całkowitą jego masę ocenia się na 10^{11}–10^{12} mas Słońca. Przeważająca część masy emitującej promieniowanie skupiona jest w zagęszczeniu centralnym — kulistym tworze (il. 223 i 224, tabl. 60) o średnicy zaledwie 5000 l.św. Tworzą je gwiazdy populacji II, których łączna masa stanowi $^1/_5$ masy Galaktyki. Gęstość materii w zagęszczeniu centralnym jest kilkaset razy większa niż w bezpośrednim otoczeniu Słońca i wynosi ok. 20–30 mas Słońca/l.św.^{3}

Rozkład przestrzenny gwiazd w Galaktyce ma dwie symetrie: zwierciadlaną — względem płaszczyzny Galaktyki, i osiową — względem osi Galaktyki ustawionej do płaszczyzny Galaktyki prostopadle i przechodzącej przez obszar największej gęstości. Obszar ów zwano dawniej jądrem, obecnie — coraz powszechniej używa się nazwy zagęszczenie centralne. Płaszczyzna Galaktyki (także: płaszczyzna równikowa lub równik) przecina sferę niebieską wzdłuż pasa Drogi Mlecznej. Zagęszczenie centralne leży w kierunku wyznaczonym przez gwiazdozbiór Strzelca. Z Ziemi jest niewidoczne — zasłaniają je obłoki materii międzygwiazdowej nieprzezroczyste dla promieniowania widzialnego.

zagęszczenie centralne

Rozmieszczenie i ruchy gwiazd

Powierzchnie jednakowych gęstości przestrzennych gwiazd (przez gęstość przestrzenną gwiazd należy rozumieć liczbę gwiazd w jednostce objętości) są współosiowymi elipsoidami obrotowymi o różnych stopniach spłaszczenia i różnych rozmiarach (rys. 1). W latach czterdziestych XX w. W. Baade zauważył, że powierzchnie jednakowej gęstości gwiazd o danym typie widmowym tworzą rodzinę elipsoid podobnych,



Rys. 1. Przenikanie się podsystemów: powierzchnie stałych gęstości podsystemów Układu Drogi Mlecznej w modelu Galaktyki Oorta (przekrój płaszczyzną prostopadłą do równika Galaktyki i przechodzącą przez środek Galaktyki oraz Słońce). Kółeczko oznacza położenie Słońca. Model Oorta składa się z trzech podsystemów: płaskiego (dwie wewnętrzne silnie spłaszczone elipsy), pośredniego (trzy zewnętrzne słabiej spłaszczone elipsy) i sferycznego (dwie wewnętrzne najsłabiej spłaszczone, zacieniowane elipsy)

tzn. mających jednakowe stopnie spłaszczenia, i podzielił Galaktykę na podsystemy (populacje). Wyróżniamy obecnie pięć podsystemów:

podsystemy gwiazd

1) Skrajny podsystem płaski — 3% masy Galaktyki, ok. 2 mld gwiazd (skrajna populacja I): błękitne nadolbrzymy, gwiazdy typu *T* Tauri, cefeidy, materia międzygwiazdowa.

2) Podsystem płaski. — 8% masy Galaktyki, ok. 5 mld gwiazd (starsza populacja I): gwiazdy ciągu głównego od karłów do typu widmowego *A* włącznie.

3) Dysk (populacja II): mgławice planetarne, gwiazdy nowe, zmienne typu *RR* Lyrae z okresami zmian blasku mniejszymi od dziesięciu godzin, zmienne długookresowe z okresami większymi od 140 dni.

4) Podsystem pośredni — wraz z dyskiem zawiera ok. 50 mld gwiazd, czyli 67% masy Galaktyki (pośrednia populacja II): zmienne długookresowe z okresami mniejszymi od 140 dni.

5) Halo — 22% masy Galaktyki, ok. 16 mld gwiazd (skrajna populacja II): gwiazdy typu *RR* Lyrae z okresami większymi od dziesięciu godzin, gromady kuliste, podkarły zawierające minimalne ilości pierwiastków cięższych od helu.

Im silniej spłaszczony jest podsystem, tym wyraźniej jest w nim zaznaczona płaszczyzna Galaktyki.

W podsystemach słabo spłaszczonych (tzw. podsystemach sferycznych) płaszczyzna Galaktyki nie jest wyróżniona, wyraźne maksimum gęstości pojawia się natomiast w centrum.

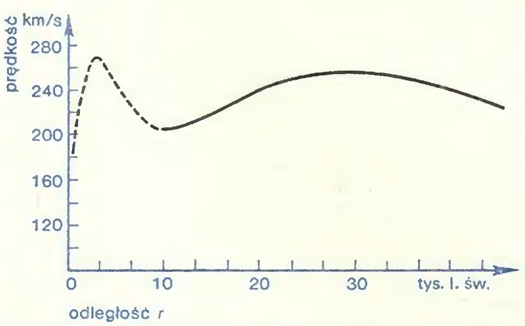
W podsystemach płaskich przeważają gwiazdy zawierające dużo pierwiastków cięższych od helu (nazywanych powszechnie metalami), zaś w podsystemach sferycznych — gwiazdy zawierające bardzo mało metali. Metale powstają w reakcjach syntezy jądrowej zachodzących we wnętrzach gwiazd. Zaobserwowano, iż gwiazdy w trakcie ewolucji tracą część swej masy. Wynika stąd, że gwiazdy podsystemu pośredniego i halo (populacja II) musiały powstać we wcześniejszych etapach ewolucji Wszechświata, tj. wtedy, gdy pierwiastków ciężkich było mało. Te z nich, które ewoluowały szybciej, zwróciły przetworzoną i wzbogaconą w metale materię do ośrodka międzygwiazdowego. Materia międzygwiazdowa skupiona jest obecnie w płaszczyźnie równikowej i tylko sporadycznie występuje w dużych odległościach od równika. W okolicach równika powstają gwiazdy młodszej generacji (populacja I) z bogatych w metale obłoków. Młoda gwiazda porusza się początkowo w płaszczyźnie równikowej, w której działają słabe siły zakłócające. W efekcie ich działania po dostatecznie długim czasie zaawansowana już ewolucyjnie gwiazda przechodzi do podsystemu mniej spłaszczonego, rozpraszając część swej masy w płaszczyźnie Galaktyki.

ramiona spiralne Galaktyki

Podsystem płaski ma strukturę spiralną zaznaczoną najwyraźniej w płaszczyźnie równikowej. Tworzą ją trzy ramiona spiralne: Perseusza, Oriona i Strzelca (wg ostatnich, nie potwierdzonych jeszcze obserwacji, Galaktyka ma cztery ramiona spiralne). W ramionach spiralnych skupione są niemal wszystkie gwiazdy młode i jasne oraz świecące obłoki gazowe (tzw. mgławice emisyjne). Płynące z obserwacji innych galaktyk wrażenie dominowania ramion nad pozostałymi częściami podsystemu płaskiego jest w znacznej mierze wrażeniem pozornym. W rzeczywistości gęstość materii w ramionach jest zaledwie trzykrotnie większa od średniej gęstości podsystemu płaskiego. W zewnętrznej części ramienia Oriona, w odległości około 30 000 lat świetlnych od środka Galaktyki i około 30 lat świetlnych nad jej płaszczyzną równikową, znajduje się Układ Słoneczny. Słońce należy do żółtych karłów często spotykanych w Galaktyce i niczym się nie wyróżnia spośród gwiazd leżących w jego otoczeniu.

ruchy gwiazd

Wszystkie gwiazdy obiegają centrum Galaktyki po mniej lub bardziej spłaszczonych orbitach. Na początku naszego wieku stwierdzono, iż średni ruch gwiazd należących do pewnych grup kinematycznych można zinterpretować jako rotację owych grup. W końcu lat czterdziestych grupy kinematyczne utożsamiono z podsystemami. Najszybciej rotują



Rys. 2. Prędkość liniowa obrotu Galaktyki (v) w funkcji odległości od centrum (r). Podobnie, tzn. w sposób opisany III prawem Keplera, obracają się pierścienie Saturna. Dla $r < 10\,000$ l.św. (linia przerywana) dane obserwacyjne nie są zbyt pewne

podsystemy płaskie, w których orbity gwiazd są prawie kołowe, zaś ruchy własne gwiazd rozumiane jako odchylenia od średniego ruchu całego podsystemu — niewielkie (prędkości ich nie przekraczają 20–30 km/s). W podsystemach sferycznych orbity gwiazd są silnie wydłużonymi elipsami, dowolnie nachylonymi do równika Galaktyki. Ruch obrotowy całości jest powolny, natomiast ruchy własne — szybkie (300–500 km/s) i nieuporządkowane. Podsystemy rotują różniczkowo, tzn. z prędkością kątową zależną od odległości od osi obrotu (w podobny sposób obracają się pierścienie Saturna). Nawet w płaszczyźnie Galaktyki, w której opisywane ruchy są najprostsze, zależność owa jest dość skomplikowana (rys. 2). Kształt funkcji $v(r)$ zależy od rozkładu przestrzennego masy. W małych odległościach od centrum v jest proporcjonalne do r. Oznacza to, iż zagęszczenie centralne jest na tyle silnie powiązane siłami grawitacji, że obraca się jak ciało sztywne. Kierunek obrotu Galaktyki można określić podając, że ramiona „nawijają się" na oś. Wydaje się to być cechą wspólną wszystkich galaktyk spiralnych.

teoria fal gęstościowych

Natura ramion spiralnych, a zwłaszcza ich stabilność (teoretycznie powinny one wskutek rotacji różniczkowej i istnienia ruchów własnych, rozpłynąć się już po 1–2 obrotach Galaktyki) nie są w pełni wyjaśnione. Teoria fal gęstościowych sformułowana przez B. Lindblada, a rozwinięta przez C. Lina i F. Shu, zyskała obecnie największe uznanie. Zgodnie z nią ramię spiralne jest wędrującym przez dysk zaburzeniem (falą gęstości). Gwiazdy, które obecnie należą do ramion, po upływie pewnego czasu znajdą się w przestrzeni między ramionami, by jeszcze później trafić znowu do ramion, lecz już w innych miejscach. Stabilność takich fal gęstościowych została wykazana w wielu eksperymentach numerycznych. Sposób powstawania fal i czynniki determinujące ich liczbę są nadal nieznane.

gromady gwiazd

Gwiazdy należące do dysku tworzą często małe skupienia, zwane asocjacjami, gromadami ruchomymi, bądź gromadami otwartymi. Składniki takich skupień są ze sobą powiązane genetycznie (powstają z jednego dużego obłoku materii międzygwiazdowej), a wektory ich prędkości są do siebie równoległe. Najbogatsze z wymienionych skupień zawierają po kilka tysięcy gwiazd należących z reguły do populacji I. O wiele bogatsze (do 10^5 gwiazd) gromady kuliste utworzone są z gwiazd populacji II. Obszarem maksymalnej koncentracji gromad kulistych jest centrum Galaktyki.

Materia rozproszona

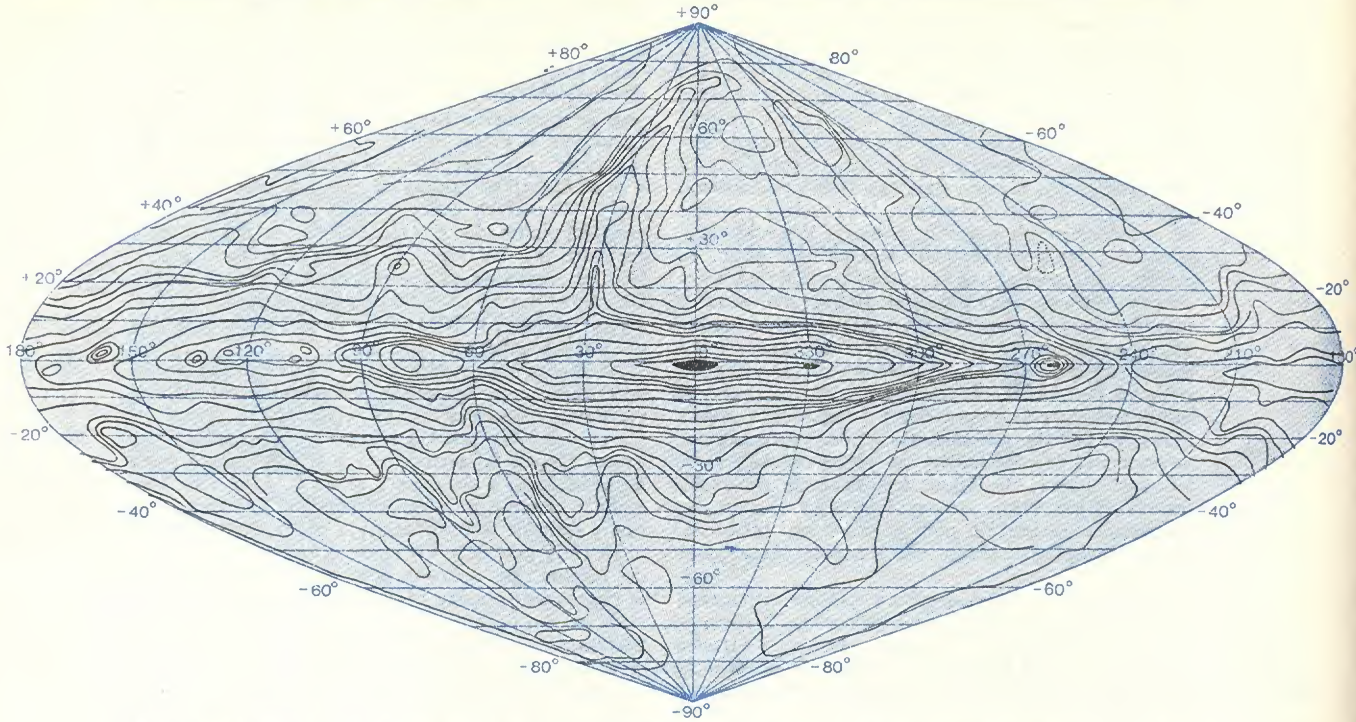
Najważniejszym źródłem danych o materii rozproszonej są obserwacje radiowe. Promieniowanie radiowe Galaktyki zostało wykryte w 1932 r. przez Karla Jansky'ego. Początki radioastronomii przypadają jednak dopiero na drugą połowę lat czterdziestych naszego wieku (→ Radioastronomia).

Wygląd „nieba radiowego" widzianego z Ziemi przedstawia rys. 3. Większość promieniowania radiowego pochodzi z obszarów maksymalnej koncentracji materii rozproszonej, czyli z okolic równika Galaktyki. Maksimum główne natężenia leży w centrum Galaktyki. Szereg maksimów pobocznych obserwuje się w miejscach, w których promień widzenia obserwatora styczny jest do jednego z ramion spiralnych. Wyrastające z płaszczyzny równikowej tzw. ostrogi radiowe (największa z nich ciągnie się wzdłuż południka galaktycznego 30°) najprawdopodobniej są tworami lokalnymi. Wiąże się je obecnie z niejednorodnościami ramienia Oriona, albo z dawnymi wybuchami bliskich supernowych. Tło radiowe Galaktyki jest mieszaniną promieniowania synchrotronowego i termicznego. Wykazano niedawno, że składowa termiczna tła powstaje w wyniku nałożenia się słabych obrazów licznych źródeł dyskretnych. Pas emisji termicznej jest bardzo wąski, grubość jego na ogół nie przekracza 2–3° (na częstości 2,7 GHz). Nieco szerszy pas emisji synchrotronowej ma miejsca-

„niebo radiowe"

mi grubość 5°. Cechą charakterystyczną promieniowania tła jest szybkie zwężanie się pasa emisji ze wzrostem częstości rejestrowanych fal.

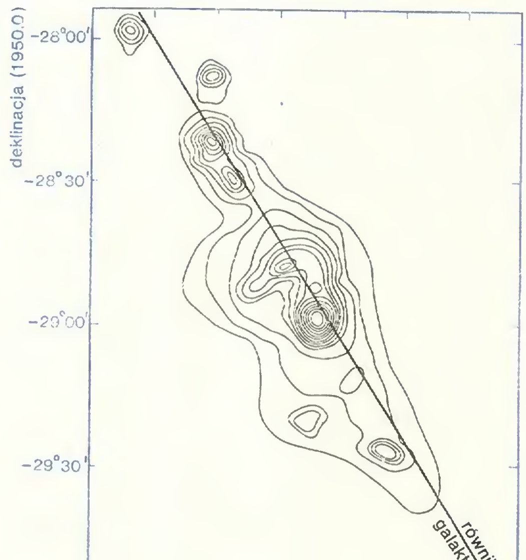
Według niektórych obserwatorów podsystemy płaskie otoczone są rozciągłym halo radiowym wysyłającym promieniowanie synchrotronowe. Struktury po-



Rys. 3. Mapa radiowa nieba na częstości 150 MHz. „Poziomice" (izofoty) odpowiadają ustalonym wartościom natężenia promieniowania radiowego. Punkt (0°) jest środkiem Galaktyki. Wzdłuż południka 30° biegnie największa z ostróg radiowych. Na przecięciu się południków 90° i 270° z równikiem widać piki emisji pochodzące z ramion spiralnych

źródła dyskretne

Nazwą źródła dyskretne obejmuje się obłoki materii międzygwiazdowej, pulsary i radiogwiazdy. Najbardziej interesująca grupa silnych źródeł dyskretnych leży w centrum Galaktyki (rys. 4). Termiczno-synchrotronowe źródło Sagittarius A (SGR A) uważane jest za właściwe jądro Galaktyki. Pozostałe z widocznych na rys. 4 radioźródeł są obłokami pyłowo-gazowymi i mają widmo czysto termiczne.



Rys. 4. Mapa radiowa centrum Galaktyki na częstości 8 GHz

dobne do halo wykryto także w dwóch sąsiednich galaktykach.

obszar H I

Najdokładniejsze informacje dotyczące rozmieszczenia i składu materii rozproszonej pochodzą z obserwacji radiowego widma liniowego. Prócz znanej od dawna linii 21 cm wodoru neutralnego H I, w ostatnich latach wykryto wiele linii prostych związków chemicznych. Ostre maksimum gęstości wodoru neutralnego (10 atomów/cm^3) leży w centrum Galaktyki. Już w odległości 2000 l.św. od centrum gęstość H I wynosi zaledwie 0,2 atomów/cm^3. Rośnie następnie powoli, by w odległości 15 000 l.św. osiągnąć 0,6 atomów/cm^3. Na tym poziomie utrzymuje się przez dalsze 2000 l.św., po czym ponownie opada. Części centralne Galaktyki są zatem otoczone wyraźnym pierścieniem wodoru neutralnego. Całkowita masa pierścienia stanowi ok. 10% masy Galaktyki. Jego grubość waha się od 450 do 600 l.św. Począwszy od odległości 30 000 l.św., płaszczyzna pierścienia zaczyna się odchylać od płaszczyzny równikowej Galaktyki w stronę bieguna północnego Galaktyki w przedziale długości galaktycznych 20–140° i w stronę południowego w przedziale 200–240°. Na zewnętrznej krawędzi pierścienia odchylenie to osiąga wartość 2000 l.św. Zniekształcenie pierścienia H I wywołane jest prawdopodobnie przez siły grawitacyjne pochodzące od najbliższych galaktyk (Obłoków Magellana).

globule

Ośrodek międzygwiazdowy, który składa się głównie z wodoru, jest niejednorodny: są to obłoki o różnych rozmiarach. Średnica przeciętnego obłoku wynosi 30 l.św., gęstość — ok. 25 atomów /cm^3, temperatura — ok. 100 K. Średnia gęstość ośrodka między obłokami jest stukrotnie mniejsza. Najgęstsze ze znanych obłoków, tzw. globule (gęstość do 10 000 atomów/cm^3) uważane są przez niektórych badaczy za gwiazdy znajdujące się w fazie kondensacji przed zapaleniem wodoru.

obszar H II

Zaledwie 3% masy wodoru międzygwiazdowego wysyła promieniowanie widzialne. Są to tzw. obszary

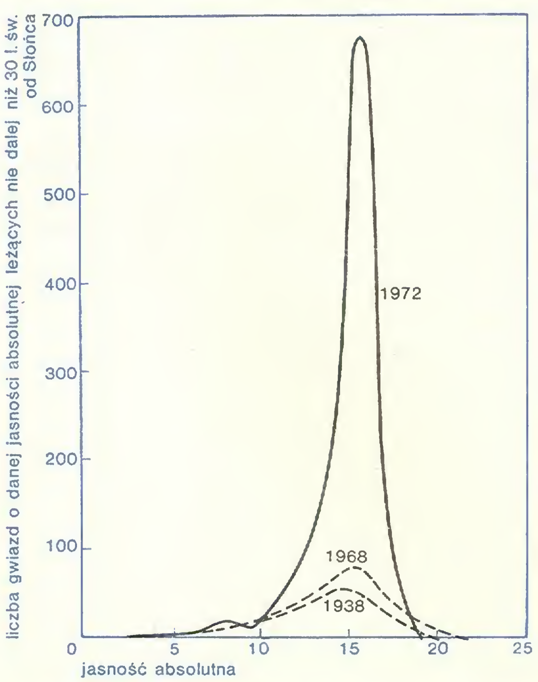
H II wodoru zjonizowanego. Z obłokami H I i H II sąsiadują często obłoki pyłowe złożone prawdopodobnie z ziaren lodowo-grafitowych, widoczne niekiedy jako ciemne plamy na tle jasnych mgławic emisyjnych. Pył kosmiczny jest katalizatorem umożliwiającym powstawanie różnych cząsteczek chemicznych. Liczba zidentyfikowanych cząsteczek i rodników sięga obecnie 30; do występujących najczęściej należą OH, H_2O, CO i H_2CO. Gęstą chmurę pyłowo-gazową zawierającą rodniki OH i cząsteczki CHOH wykryto niedawno w centrum Galaktyki, w pobliżu źródła SGR A. Niektóre obłoki dostatecznie gęste i zawierające odpowiednio dużo OH lub H_2O emitują niezwykle duże ilości energii w postaci promieniowania odpowiadającego liniom widmowym tych cząsteczek powstającym na Ziemi w zjawisku maserowym. Próby wytłumaczenia mechanizmu pompowania cząsteczek obłoku na wyższe poziomy energetyczne nie dały dotychczas rezultatu i zasada działania „masera międzygwiazdowego" pozostaje niewyjaśniona.

ruchy obłoków

Materia rozproszona bierze udział w ruchu obrotowym Galaktyki. W odległościach od centrum większych niż 12 tys. l.św. obłoki poruszają się po dobrze określonych, prawie kołowych orbitach. Bliżej centrum (5–6 tys. l.św.) prędkości ruchów radialnych są porównywalne z prędkościami orbitalnymi i pojęcie orbity obłoku traci sens. W obszarze tym kilkanaście obłoków wodorowych, o łącznej masie ok. 10^7 mas Słońca, oddala się z dużymi prędkościami od środka Galaktyki. Wektory ich prędkości leżą w płaszczyźnie równikowej lub są do niej nachylone pod niewielkimi tylko kątami. Największy z obłoków nazywany bywa ramieniem ekspandującym. W bezpośrednich okolicach jądra ruchy radialne zanikają. Materia międzygwiazdowa tworzy tu wraz z gwiazdami populacji II dwie struktury: zewnętrzny pierścień o promieniu 2000 l.św. i szerokości 1500 l.św. oraz wewnętrzny dysk o promieniu 350 l.św. Masa pierścienia i dysku oceniana jest na 10^{10} mas Słońca, z czego na materię międzygwiazdową przypada zaledwie 0,1%, a więc o wiele mniej niż w okolicach Słońca.

ramię ekspandujące

Analiza zaburzeń ruchu obrotowego materii międzygwiazdowej prowadzi do przypuszczenia, iż przed mniej więcej trzynastoma milionami lat w jądrze Galaktyki nastąpił wybuch. Uwolniła się przy tym energia równa mocy promieniowania miliardów Słońc. Materia z okolic jądra została wyrzucona pod kątem ok. 30° do płaszczyzny równikowej i częściowo ściągnięta do niej na powrót przez siły grawitacyjne Galaktyki i pole magnetyczne.



Rys. 5. Ilustracja niepewności wiedzy o Galaktyce. Takim zmianom ulegały w ciągu trzydziestu lat wyobrażenia o kształcie funkcji świecenia gwiazd w otoczeniu Słońca — czyli w najlepiej zbadanym obszarze Galaktyki

pole magnetyczne Galaktyki

Dokonane różnymi metodami oceny średniego natężenia pola magnetycznego Galaktyki dały zgodne wartości $3\text{–}5 \cdot 10^{-10}$ T (lokalne pole może być nawet dziesięciokrotnie słabsze lub silniejsze od średniego). Dzięki badaniom polaryzacji światła gwiazd wiadomo, że linie sił pola układają się wzdłuż ramion spiralnych. Energia zawarta w polu elektromagnetycznym jest nikłym ułamkiem energii grawitacyjnej Galaktyki, lecz nawet tak słabe pole wystarcza w zupełności do utrzymania w Galaktyce → Promieniowania kosmicznego — strumieni cząstek elementarnych o dużych energiach. Źródłem promieniowania kosmicznego są najprawdopodobniej wybuchy supernowych i — być może — jakieś nie znane jeszcze procesy zachodzące w jądrze Galaktyki.

funkcja świecenia

Nasza wiedza o Galaktyce jest jeszcze bardzo niepełna i obraz Galaktyki ulega ciągłym zmianom. Najlepszym dowodem tego jest rys. 5, przedstawiający tzw. funkcję świecenia gwiazd tak jak się przedstawiała w latach 1938, 1968 i 1972. Mimo że do wyznaczania funkcji świecenia używano gwiazd leżących w bezpośrednim otoczeniu Słońca, a więc w najlepiej poznanym rejonie Galaktyki, różnice są olbrzymie.

Ciąg Hubble'a

Klasyfikacja morfologiczna galaktyk, wprowadzona przez Edwina Hubble'a, jest jeszcze powszechnie stosowana, choć całkowicie straciła przypisywane jej niegdyś znaczenie teoretyczne. Ciąg Hubble'a jest uporządkowany wg rosnącej symetrii (il. 225, tabl. 61). Poniższe zestawienie obejmuje nazwy i skrótowe oznaczenia typów hubble'owskich oraz ich charakterystyki morfologiczne.

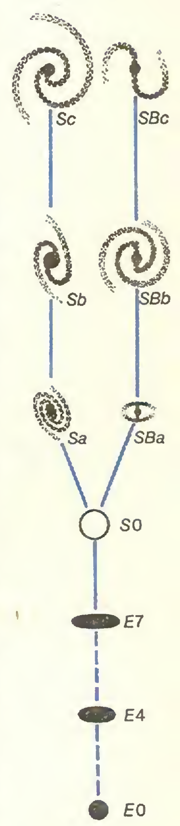
1) Galaktyki nieregularne (*I* 1). Brak jakiejkolwiek symetrii. Wyglądem przypominają jasne obłoki materii międzygwiazdowej. Typową galaktyką nieregularną jest Mały Obłok Magellana (il. 225a, tabl. 61).

2) Galaktyki spiralne (*S*). Niepełna symetria osiowa. Materia świecąca tworzy cienki dysk i eliptyczne zagęszczenie centralne, z którego wybiegają wtopione w dysk ramiona. W galaktykach oglądanych z boku widoczny jest ciemny pas materii pyłowej skupionej w płaszczyźnie równikowej. Zależnie od wielkości zagęszczenia centralnego i stopnia rozwoju ramion spiralnych, galaktyki *S* zalicza się do jednego z trzech podtypów: *Sc* (Mgławica Trójkąta, il. 225b, tabl. 61), *Sb* (Mgławica Andromedy, il. 225c, tabl. 61) oraz *Sa* (M 104, il. 225d, tabl. 61). Galaktyką spiralną typu *Sb* jest także Układ Drogi Mlecznej.

3) Galaktyki spiralne z poprzeczką (*SB*). Niepełna symetria osiowa. Ramiona spiralne zaczynają się w pewnej odległości od zagęszczenia centralnego. Początki ich połączone są przechodzącym przez środek galaktyki pomostem świecącej materii — tzw. poprzeczką. Przyczyny formowania się poprzeczek nie są jeszcze znane. Typ *SB* dzieli się na trzy podtypy, analogiczne do podtypów *S*: *SBc*, *SBb* i *SBa* (il. 225 e–g, tabl. 61).

4) Galaktyki soczewkowate (*S0*). Pełna symetria osiowa. Zagęszczenie centralne otoczone jest niezbyt rozległym, lecz grubym dyskiem. Galaktyki *SB* nie mają ramion spiralnych. W żadnej z nich nie stwierdzono obecności materii pyłowej (il. 225h, i, tabl. 61).

5) Galaktyki eliptyczne (*E*). Symetria osiowa przechodząca w sferyczną. Brak dysku. Na sferze niebieskiej widoczne są jako elipsy o różnych stopniach



spłaszczenia. Miarą obserwowanego spłaszczenia jest stosunek $10 \cdot (a-b)/a$, gdzie a i b — wielka i mała półoś elipsy. Obserwowane stopnie spłaszczenia zawierają się w przedziale (0–7) — stąd podział typu *E* na osiem podtypów: od *E*7 do *E*0. Rzeczywiste stopnie spłaszczenia można wyznaczać bądź grupowo (metodami statystycznymi), bądź indywidualnie (metodami fotometrycznymi). Ostatnia metoda pozwala także na odróżnienie galaktyk *E* od galaktyk *S*0, widzianych od strony bieguna (por. il. 225l, h, tabl. 61) Z badań częstości występowania obserwowanych stopni spłaszczenia wynika, iż istnieją galaktyki o stopniu spłaszczenia równym 0, czyli o pełnej symetrii sferycznej. W galaktykach *E* bardzo rzadko obserwuje się materię pyłową. Do przedstawicielek typu *E* należą towarzyszki Mgławicy Andromedy galaktyki M 32 (*E5*, il. 225j, tabl. 61) i NGC 185 (*E2*, il. 225k, tabl. 61) oraz ogromna galaktyka sferyczna M 87 (*E*0, il. 225l, tabl. 61). Pozostałe 0,5% stanowią galaktyki osobliwe (pekuliarne), nie należące do żadnego z wymienionych typów. Jest o nich mowa w rozdziale Jądra galaktyk.

częstość występowania różnych typów

Typy galaktyk

Typ	Częstość występowania, %
*E*0 + *E*1 + ... + *E*7 + *S*0	17
Sa + *SBa*	19
Sb + *SEb*	25
Sc + *SBc*	36
*I*1	2,5

Kilkanaście lat po ukazaniu się prac Hubble'a jego klasyfikację uzupełniono karłowatymi galaktykami eliptycznymi *I*2, które mają cechy strukturalne galaktyk eliptycznych, różnią się od nich nikłymi rozmiarami i bardzo małą gęstością gwiazd. Wyjątkowo małe jasności powierzchniowe utrudniają, a przy nieznacznym pogorszeniu warunków atmosferycznych wręcz uniewożliwiają, obserwacje galaktyk *I*2. Odkryto ich zaledwie kilka, istnieją jednak podstawy by przypuszczać, że typ *I*2 jest w Kosmosie reprezentowany bardzo licznie. Typowi przedstawiciele galaktyk *I*2 to tzw. systemy Sekstans i Fornax (il. 225 m, n, tabl. 61).

Rozmieszczenie przestrzenne galaktyk

Niejednorodności w rozkładzie galaktyk na sferze niebieskiej znane były już Herschelowi, jednak dokładnie zostały zbadane dopiero przez Hubble'a. Związane z nimi problemy są dziś uważane za jeden z najważniejszych w kosmologii i teorii ewolucji galaktyk.

pas unikania

Nie obserwuje się zupełnie galaktyk leżących w obrębie tzw. pasa unikania, ciągnącego się wzdłuż równika Galaktyki: ich światło jest całkowicie pochłaniane przez materię międzygwiazdową. Jest to przykład pozornej niejednorodności rozkładu, wywołanej niedoskonałością obserwacji. Prawdziwy rozkład galaktyk na sferze niebieskiej, otrzymany po uwzględnieniu zafałszowań spowodowanych absorpcją, jest nadal niejednorodny. Galaktyki grupują się na sferze w układach zwanych (zależnie od liczebności) grupami bądź gromadami. Z badań statystycznych wynika, że widoczne na niebie grupy i gromady nie powstają w wyniku przypadkowego zlewania się bliższych galaktyk z dalszymi, lecz są odbiciem rzeczywistych niejednorodności rozkładu przestrzennego. Na realność fizyczną bogatych gromad wskazuje ich regularna budowa (il. 229, tabl. 63). Ponadto w niektórych mało liczebnych grupach zaobserwowano mosty świecącej materii, łączące poszczególne galaktyki. Większość obserwatorów zdaje się obecnie skłaniać ku stwierdzeniu, iż w ogóle nie istnieją galaktyki pojedyncze, nie należące do żadnej gromady. Nie zaobserwowano natomiast, aby gromady galaktyk tworzyły ugrupowania wyższego rzędu, tj. gromady gromad. Oznacza to, że dostępna obserwacjom część Kosmosu jest jednorodna w skali gromad.

grupy i gromady galaktyk

Niewiele dotychczas wiadomo o przeciętnej gromadzie. Najlepiej zbadane są gromady liczące po kilkaset i więcej galaktyk (il. 229, tabl. 63). Są to zwykle układy niezbyt zwarte. Gęstość przestrzenna galaktyk w gromadzie rośnie jednak znacząco w miarę zbliżania się do centrum, w którym często można znaleźć galaktykę przewyższającą znacznie jasnością i rozmiarami wszystkie pozostałe składniki gromady. Galaktyka centralna jest zazwyczaj galaktyką aktywną (patrz rozdział Jądra galaktyk). Z centralnych okolic paru gromad odebrano promieniowanie rentgenowskie interpretowane jako promieniowanie termiczne gazu międzygalaktycznego. Jest to jedyny, jak dotąd, dowód istnienia materii międzygalaktycznej.

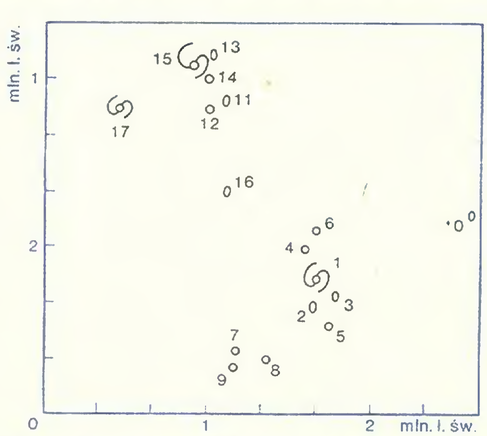
Zakładając, że gromada jest tworem stabilnym, który istnieje w niezmienionej postaci od paru miliardów lat, i mierząc prędkości radialne jej składników można (z twierdzenia o wiriale) wyznaczyć jej masę. Otrzymane wartości przewyższają dziesięciokrotnie masy ocenione na podstawie zsumowania przybliżonych mas poszczególnych galaktyk. Zatem — albo gromady nie są stabilne (czemu jednak przeczy sam fakt ich obserwowania), albo błędne są wyobrażenia na temat mas pojedynczych galaktyk. Według najbardziej prawdopodobnej hipotezy „brakująca" w galaktykach masa skupiona jest w czerwonych karłach ciągu głównego, należących zarówno do pierwszej, jak i do drugiej populacji gwiazd. Teoria ewolucji gwiazd przewiduje ich istnienie, lecz z powodu znikomych jasności są one niemożliwe do zaobserwowania.

gromada Virgo

Największą w naszym otoczeniu Wszechświata gromadą jest gromada Virgo licząca kilkaset jasnych i kilka tysięcy słabszych galaktyk. Środek jej leży w odległości ok. 35 000 000 l.św. od Słońca. W jego pobliżu znajduje się dominująca w gromadzie galaktyka M 87 typu *E*0, niezwykle masywna (ok. $10^{13}\ M_\odot$) i otoczona wyjątkowo liczną chmurą gromad kulistych. Niektórzy obserwatorzy zaliczają do gromady Virgo Układ Lokalny Galaktyk, zawierający Układ Drogi Mlecznej.

Układ Lokalny Galaktyk

Układ Lokalny (jak i większość dokładniej zbadanych małych grup galaktyk) jest tworem niestabilnym: jego składniki rozbiegają się. Do Układu Lokalnego zalicza się wszystkie galaktyki położone w odległości mniejszej od 2 000 000 l.św.; znamy ich dwa-



Rys. 6. Schemat rozmieszczenia galaktyk w Układzie Lokalnym. Numery odpowiadają numerom porządkowym w umieszczonej w tekście tabeli. Kółeczka oznaczają galaktyki eliptyczne, owale — galaktyki nieregularne, spirale — galaktyki spiralne

dzieścia parę. Nasza Galaktyka, której towarzyszą dwie galaktyki satelitarne (Obłoki Magellana), jest największa w Układzie. Dorównuje jej prawie rozmiarami Mgławica Andromedy, także mająca dwie galaktyki satelitarne. Reszta składników Układu to galaktyki pojedyncze, wielokrotnie mniejsze od naszej. Badania Układu Lokalnego Galaktyk prowadzą zatem do wniosku, iż najliczniej reprezentowane są we Wszechświecie galaktyki o małych masach, a wśród nich — karłowate galaktyki *I2*. Krótkie charakterystyki składników Układu Lokalnego zawarte są w tabeli, a ich rozmieszczenie przedstawia rys. 6.

Galaktyki nie objęte zestawieniem (w 1975 r. było ich 4) należą do klasy *I2*. Zostały one odkryte w ostatnich latach i brak o nich na razie dokładniejszych danych. Jedna z nich, odkryta w połowie 1975 r., leży w odległości zaledwie 60 000 l.św., a więc bliżej niż Obłok Magellana. Numer galaktyki w tabeli odpowiada oznaczeniom na rys. 6, który przedstawia Układ Lokalny w rzucie na płaszczyznę Galaktyki. Ilustracja 230 (tabl. 63) przedstawia fotografię ciasnej grupy wzajemnie oddziałujących galaktyk zw. Kwintetem Stefana — inny przykład niestabilnego układu galaktyk. Jego składniki rozbiegają się z dużymi prędkościami.

Układ Lokalny Galaktyk

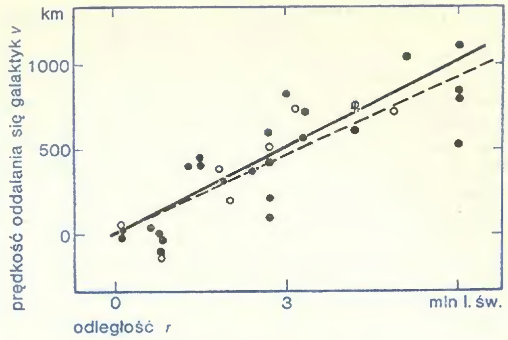
Numer	Nazwa	Typ	Odległość tys. l.św.	Jasność widoma	Wielkość gwiazdowa absolutna
1	Układ Drogi Mlecznej	*Sb*	0	—	−19,8
2	Wielki Obłok Magellana	*I1*	150	2,8	−17,4
3	Mały Obłok Magellana	*I1*	150	1,2	−16,0
4	System UMa	*I2*	220	10,0	−9,0
5	System Sculptor	*I2*	300	8,8	−12,1
6	System Draco	*I2*	330	10,0	−10,0
7	System Leo 1	*I2*	680	12,0	−9,7
8	System Leo 2	*I2*	680	12,0	−9,7
9	System Fornax	*I2*	910	9,1	−13,4
10	NGC 6822	*I1*	1100	9,1	−13,9
11	NGC 147	*E3*	1250	10,5	−13,4
12	NGC 185	*E1*	1250	10,2	−13,7
13	NGC 205	*E5*	1500	8,9	−15,0
14	NGC 221 (M 32)	*E2*	1500	9,1	−14,8
15	Mgławica Andromedy	*Sb*	1500	4,3	−19,8
16	IC 1613	*I1*	1500	10,0	−13,5
17	Mgławica Trójkąta	*Sc*	1600	6,0	−17,6

Galaktyki wchodzące w skład Układu Lokalnego porozrzucane są po całej sferze. By stwierdzić, czy dana galaktyka należy do Układu — trzeba wyznaczyć jej odległość. Najdokładniejszą z metod wyznaczania odległości galaktyk jest metoda cefeid zastosowana już przez Hubble'a. Zaobserwowawszy w galaktyce cefeidę (il. 226, tabl. 62), którą łatwo jest rozpoznać po charakterystycznej krzywej zmian blasku, mierzy się okres owych zmian, związany dość jednoznacznie z jasnością absolutną gwiazdy. Znając jasność widomą otrzymuje się już łatwo odległość. Niestety, metodą cefeid można wyznaczyć tylko odległości niezbyt duże, nie większe niż 7 mln l.św. Galaktyki z centralnych części gromady Virgo leżą już poza obszarem jej stosowalności. Inne metody pozwalają sięgnąć w Kosmos nieco dalej, dają jednak mniej dokładne wyniki. Należą do nich metoda gwiazd nowych, wykorzystująca założenie o stałej jasności absolutnej nowych w maksimum blasku, i metoda obszarów H II, wykorzystująca założenie o stałości rozmiarów liniowych mgławic emisyjnych. Obliczając opisanymi metodami odległości bliskich galaktyk E. Hubble dokonał w 1929 r. odkrycia, które pozwoliło wyznaczyć odległości nieporównywalnie większe. Stwierdził, że prędkości radialne galaktyk v, wyznaczone z przesunięć widm galaktyk ku czerwieni, rosną proporcjonalnie do odległości od obserwatora r:

wyznaczanie odległości galaktyk

$$v = H \cdot r.$$

prawo Hubble'a



Rys. 7. Zależność prędkości oddalania się galaktyk od ich odległości od obserwatora — oryginalny diagram Hubble'a z 1929 r. Różne symbole odpowiadają różnym metodom opracowania obserwacji. Miarą niedokładności obserwacji jest rozbieżność między prostą ciągłą i przerywaną

Zależność $v(r)$ otrzymana przez Hubble'a na podstawie obserwacji bliskich galaktyk przedstawiona jest na rys. 7. Zatem, aby zmierzyć odległość dzielącą nas od jakiejś galaktyki, wystarczy uzyskać jej widmo. Najdalsza z galaktyk, dla której otrzymano widma, leży w odległości kilku miliardów lat świetlnych.

Parametry fizyczne galaktyk

Galaktyki należące do różnych typów morfologicznych różnią się znacznie parametrami fizycznymi. Zależności te mają oczywiście charakter statystyczny i nie są związkami ścisłymi.

widma galaktyk

Widma galaktyk podobne są na ogół do widm pojedynczych gwiazd. Typ widmowy zmienia się od późnego *F* w typie *I1* do *G5–G6* w typie *E*. Różnice typów widmowych spowodowane są różnymi wkładami poszczególnych populacji. W galaktykach eliptycznych przeważają późne typy widmowe zaawansowanych ewolucyjnie gwiazd populacji II, zaś w galaktykach nieregularnych — młode gwiazdy populacji I o wczesnych typach widmowych. Jednak we wszystkich dokładniej zbadanych galaktykach zaobserwowano gwiazdy bardzo stare.

masy i jasności galaktyk

Średnia w danym typie morfologicznym masa galaktyki i średnia jasność rośnie wyraźnie z przejściem od typu *I* do *E*. Największe z galaktyk eliptycznych ważą kilkanaście razy więcej niż Układ Drogi Mlecznej, zaś karłowate galaktyki *I2* — do 10 000 razy mniej. Podobnie jest z jasnościami. Dwie stosowane obecnie metody wyznaczania mas (z obserwacji radiowych i optycznych) dają rażąco niezgodne wyniki (jest to tzw. problem „brakującej" masy — patrz str. 940). We wszystkich typach morfologicznych większość masy skupiona jest w gwiazdach. Na wodór neutralny H I (jedyną składową materii rozproszonej badaną powszechnie w innych galaktykach) przypada od 0,5% (typ *S0*) do 20% masy (typ *I1*). W galaktykach *I1* maksimum gęstości wodoru neutralnego pokrywa się ze środkiem geometrycznym układu. Galaktyki spiralne natomiast otoczone są pierścieniem wodorowym, tak że w środku geometrycznym wypada u nich minimum gęstości H I. Częstokroć pierścień ten jest dużo wyraźniejszy niż w naszej Galaktyce. Przyczyny tworzenia się pierścieni H I nie są znane. W galaktykach eliptycznych nie zaobserwowano wodoru neutralnego. Nie znaczy to jednak, że nie zawierają one materii rozproszonej w innych postaciach. Znaczny ich procent emituje promieniowanie synchrotronowe, świadczące o obecności zjonizowanego gazu, skupionego w obszarach centralnych. W niektórych widać ciemne obłoki pyłowe (il. 225j, k, tabl. 61).

prędkość rotacji

Największe prędkości rotacji obserwowano w systemach spłaszczonych, należących do typów *S0–Sc*.

Równowaga mechaniczna jest w nich utrzymywana dzięki równoważeniu się sił grawitacyjnych i odśrodkowych. Galaktyki *E*0 nie obracają się w ogóle, a równowagę mechaniczną utrzymują dzięki bezładnym ruchom gwiazd obiegających centrum po silnie spłaszczonych orbitach. Tak więc dość jednoznaczną funkcją typu morfologicznego jest rozkład momentu pędu (tzn. funkcja, która podaje procent masy galaktyki mający dany moment pędu). Galaktyki *S*0–*Sc* zawierają duże ilości materii z dużym momentem pędu (maksimum rozkładu — w obszarze dużych wartości), zaś galaktyki *E*0 zawierają wyłącznie materię z małym momentem pędu (maksimum rozkładu — w obszarze małych wartości).

Ostatnim parametrem związanym z typem morfologicznym jest liczba gromad kulistych towarzyszących galaktyce. Wybrawszy z typów *Sc*, *Sb*, *Sa*, *S*0 i *E* galaktyki o jednakowych masach można łatwo stwierdzić, że najwięcej gromad skupia się przy galaktykach eliptycznych.

Galaktyki karłowate są jeszcze na tyle słabo zbadane, że nie wiadomo, czy w ogóle zawierają materię rozproszoną. Tylko niektórym z nich towarzyszą nieliczne i niewielkie gromady kuliste. Większość galaktyk *I*2 zbudowana jest prawie wyłącznie z gwiazd populacji II, jednak np. galaktyki *I*2, leżące w gwiazdozbiorze Sekstansu składają się prawie wyłącznie z gwiazd młodych, należących do populacji I.

Jądra galaktyk

jądro a zagęszczenie centralne

Jeszcze kilkanaście lat temu wyrazem „jądro" określano centralną, najjaśniejszą część galaktyki. Zgodnie z ówczesnymi pojęciami galaktyki składały się wyłącznie z gwiazd i z materii rozproszonej.

Stopniowo zrozumiano jednak, że dwuskładnikowy model galaktyki jest zbyt prymitywny. Na trop tego wniosku naprowadziły znane od czasów Hubble'a galaktyki osobliwe (il. 227, tabl. 62, il. 231, 232, tabl. 63), wyłamujące się z ciągu klas morfologicznych. Wiele z nich nazywano galaktykami wybuchającymi. Proste oszacowania ilości energii potrzebnej do wprawienia w ruch ogromnych mas wyrzucanych z wybuchających galaktyk dawały wartości 10^{50} J, przekraczające wszystko, z czym się dotychczas we Wszechświecie spotkano. Dość oczywisty wniosek, iż energia wyzwalana jest w obiektach kosmicznych jeszcze nauce nie znanych, wyciągnięto jednak dopiero po wprowadzeniu nowych metod obserwacji w różnych zakresach fal radiowych i w podczerwieni.

galaktyki Seyferta

Obserwacje pewnej szczególnej klasy galaktyk osobliwych (galaktyk Seyferta) pozwoliły na ocenę rozmiarów obszaru, w którym wyzwalana jest energia. Znając całkowitą energię wytwarzaną przez hipotetyczny obiekt centralny oszacowano ilość energii wydzielanej w jednostce objętości. Ta ostatnia wartość była na tyle duża, że nie potrafiono przez długi czas podać mechanizmu produkcji energii ani wyobrazić sobie warunków fizycznych panujących w źródle. Własności nowych obiektów nie dawały się pogodzić z aktualnym stanem wiedzy. W ich istnienie uwierzono ostatecznie dopiero wtedy, gdy radioastronomowie dokonali niezależnych ocen gęstości produkcji energii i gdy zaczęto masowo odkrywać równie dziwne obiekty (→ Kwazary). Znaczenie nazwy jądro uległo zmianie. Większość autorów oznacza nią obecnie nowo odkryte obiekty centralne galaktyk i w takim też znaczeniu jest używana w niniejszym rozdziale. To, co dawniej nazywano jądrem, określa się dziś jako zagęszczenie centralne.

Na naświetlanych odpowiednio krótko zdjęciach centralnych części najbliższych galaktyki (il. 228, tabl. 62) można spostrzec punktowe lub niemal punktowe obiekty, różniące się od zwykłych gwiazd własnościami fotometrycznymi. Bezpośrednie ich obserwacje są niezmiernie utrudnione przez gwiazdy należące do zagęszczeń centralnych. Nie udało się jeszcze uzyskać takich widm tych obiektów, które nie byłyby zafałszowane przez widmo otoczenia. Mimo to uważane są one za jądra galaktyk znajdujące się w fazie spoczynku. Najprawdopodobniej wszystkie galaktyki masywne mają co najmniej jedno jądro, natomiast nie obserwuje się jąder w galaktykach karłowatych.

Galaktyki, których jądra znajdują się w fazie aktywności, nazywane bywają w skrócie aktywnymi. Oznakami aktywności galaktyki mogą być zmiany blasku, obecność linii emisyjnych w widmie, silne promieniowanie ultrafioletowe lub podczerwone, intensywna emisja radiowa, wreszcie — wypływ materii (il. 227, tabl. 62, il. 231, 232, tabl. 63). Różne rodzaje aktywności galaktyk mają to samo źródło — niestacjonarne procesy zachodzące w jądrach. Fazy aktywności najprawdopodobniej powtarzają się okresowo. Skala aktywności ma bardzo dużą rozpiętość, i to co w jednych galaktykach uznać należałoby za fazę aktywności — w innych byłoby fazą spoczynku.

jądro mgławicy Andromedy

Typowym przykładem fazy spoczynku jest obecny stan jądra Mgławicy Andromedy (M 31). Optyczne własności okolic jądra nie są w żadnym stopniu anomalne. Gęstość przestrzenna gwiazd zasłaniających samo jądro (il. 228, tabl. 62) oceniana jest na 30 000/l. św.[3] — przewyższa zatem wielokrotnie gęstości gromad kulistych, najbardziej zwartych z dotychczas znanych obiektów. Wszystkie te gwiazdy należą do populacji II, są zatem stare. W okolicach jądra nie zaobserwowano gwiazd młodych, choć znajdują się tam pewne ilości wodoru neutralnego. Nieznane czynniki, prawdopodobnie szybka rotacja i silne pole magnetyczne, zapobiegają kondensacji materii międzygwiazdowej i uniemożliwiają powstawanie nowych gwiazd. Jądro Mgławicy Andromedy można bezpośrednio obserwować dopiero w ultrafiolecie i na falach radiowych. Obszar emisji radiowej jest bardziej rozmyty niż centralne źródło naszej Galaktyki, zaś emitowane promieniowanie — znacznie słabsze.

jądro Układu Drogi Mlecznej

Jądro Układu Drogi Mlecznej znajduje się w fazie przejściowej po okresie wzmożonej aktywności, przez który przechodziło kilkanaście milionów lat temu. Jest ono utożsamiane z radioźródłem SGR A. Parę lat temu stwierdzono, że SGR A jest silnym źródłem promieniowania podczerwonego (emituje ok. 10^7 razy więcej energii niż Słońce we wszystkich długościach fal). Rozmiary obszaru aktywnego w podczerwieni wynoszą zaledwie 3–4 lata świetlne. Z analiz widma promieniowania podczerwonego wynika, iż źródło SGR A otoczone jest chmurą gwiazd o gęstości milion razy większej niż gęstość w okolicach Słońca. Pozostałością po fazie aktywności jest ramię ekspandujące — ogromny obłok wyrzucony z jądra o masie 10^7 mas Słońca. Wypływ masy z jądra trwa po dziś dzień. Dowodem tego jest emitowane przez jądro promieniowanie synchrotronowe, którego podtrzymywanie wymaga nieustannego dostarczania szybkich cząstek naładowanych. W epoce wybuchu nasza Galaktyka mogłaby zapewne być sklasyfikowana jako jedna ze wspomnianych galaktyk Seyferta. Ich bardzo jasne jądra przewyższają wielokrotnie swym blaskiem pozostałe części galaktyk. W ledwo widocznych dyskach można dostrzec ślady ramion spiralnych. Widmo optyczne jądra takiej galaktyki złożone jest z nietermicznej składowej ciągłej i wielu szerokich linii emisyjnych wzbronionych. Duże ilości energii emitowane są w ultrafioletowej i podczerwonej części widma.

Wiele galaktyk Seyferta wysyła także promieniowanie rentgenowskie i gamma (→ Astronomia promieni X i γ). Ich jasności optyczne zmieniają się nieregularnie, w ciągu kilku miesięcy mogą wzrastać lub maleć nawet kilkakrotnie. Nie jest jeszcze wyjaśnione, czy zmiany owe są wynikiem redystrybucji energii do innych obszarów widma, czy odbiciem zmian wydajności źródeł energii. Przypuszcza się, że nietermiczne kontinuum pochodzi z obszarów leżących najbliżej jądra. Nad nimi, tzn. nieco dalej od jądra, znajduje

się prawdopodobnie ośrodek dużo rzadszy, produkujący zwykłe linie emisyjne. Są one poszerzone w wyniku złożenia się przesunięć dopplerowskich wywołanych bardzo szybkimi ruchami obłoków gazu. Jeszcze wyżej, gdzie gęstość powinna być jeszcze mniejsza, powstają wąskie linie wzbronione. Masa całego gazu emitującego widmo liniowe nie jest zbyt wielka: nie przekracza 10^4 mas Słońca. Natomiast jądro, utrzymujące gaz siłami ciążenia mimo jego szybkich ruchów, musi mieć masę ok. 10^8–10^9 mas Słońca; jego rozmiary nie przekraczają przy tym 5 lat świetlnych.

Prócz galaktyk Seyferta istnieje wiele innych typów galaktyk aktywnych. Należą do nich galaktyki będące silnymi radioźródłami, galaktyki typu *N*, które różnią się od galaktyk Seyferta pewnymi cechami widmowymi i jeszcze większym stosunkiem jasności jądra do dysku, wreszcie — odkryte kilka lat temu lacertydy, nazywane czasem ogniwem pośrednim między galaktykami i kwazarami. Na zdjęciach kilku spośród ok. 30 znanych obecnie lacertyd można dostrzec ślady otoczki świecącej niezwykle słabo w porównaniu z jądrem. Pozostałe obiekty różnią się od kwazarów tylko własnościami spektralnymi: w dziedzinie radiowej wysyłają o wiele więcej promieniowania krótkofalowego, zaś ich widma optyczne są całkowicie pozbawione charakterystycznych dla kwazarów linii emisyjnych i absorpcyjnych. Jasność optyczna lacertyd zmienia się nieregularnie w dużym zakresie; w ciągu jednej nocy może wzrosnąć lub zmaleć nawet o kilkadziesiąt procent. Na tej podstawie przypuszcza się, iż jądro lacertydy jest mniejsze niż jądro kwazara, a jego rozmiary nie przekraczają paru dni świetlnych. Fale elektromagnetyczne wysyłane przez lacertydy są silnie spolaryzowane, przy czym zarówno stopień jak i i płaszczyzna polaryzacji zmieniają się w czasie. Własności widmowe jąder lacertyd są nadal nieznane (brak linii); natomiast uzyskane niedawno widma otoczek okazały się niemal takie same jak widma galaktyk. Fakt ten oraz cechy morfologiczne otoczek przesądziły o uznaniu lacertyd za olbrzymie galaktyki eliptyczne z aktywnymi jądrami. Galaktykami spiralnymi o aktywnych jądrach są natomiast obiekty klasyfikowane jako galaktyki Seyferta, galaktyki *N* oraz — prawdopodobnie — kwazary.

galaktyki *N*

lacertydy

Całkowitą ilość energii wyzwolonej w jądrach w fazach aktywności ocenia się na 10^{49}–10^{53} J. Jest to energia, jakiej mogłaby dostarczyć anihilacja kilkuset tysięcy Słońc. W fazie aktywności jądro musi więc utracić znaczną część masy. Prowadzi to zapewne do zaniku aktywności.

jądra w fazie aktywności

Wyjątkowo gwałtowne procesy zdają się prowadzić do rozpadu jąder: znane są galaktyki o jądrach podwójnych. Inne (wśród nich najokazalsza w gromadzie Virgo M 87 — il. 227, tabl. 62) ukazują obserwatorom wybiegające z centrów słupy materii, w których tkwią bardzo podobne do jąder zagęszczenia. Dobrym przykładem aktywności polegającej na rozpadzie jąder są radiogalaktyki klasyczne. Mają one po kilka zwartych centrów emisji radiowej leżących na jednej prostej. Należy do nich m.in. galaktyka NGC 5128 (Centaurus A; il. 232, tablica 63). Moc promieniowania radiogalaktyk jest dziesiątki i setki milionów razy większa niż moc promieniowania Słońca we wszystkich zakresach. Centra emisji radiowej, ułożone na jednej prostej, mogą oddalać się od galaktyki macierzystej na odległości wielokrotnie przekraczające jej rozmiary, zawsze jednak jedno z centrów pokrywa się ze środkiem galaktyki. Jest ono prawdopodobnie jądrem pierwotnym, które w fazie aktywności wyrzuciło szereg jąder wtórnych, obserwowanych dziś jako pozostałe centra radiowe. Przy znikomo małych rozmiarach jąder całkowitą tajemnicą pozostaje nadal mechanizm wyzwalania energii.

radiogalaktyki

Zgodnie z hipotezami roboczymi jądra galaktyk mogą być czarnymi dziurami, supermasywnymi dyskami bądź supermasywnymi gwiazdami mającymi silne pola magnetyczne (→ Czarne dziury i zapadanie grawitacyjne).

Współczesne poglądy na temat ewolucji galaktyk

Wszystkie galaktyki zawierają gwiazdy populacji II powstałe bardzo dawno. Nie zaobserwowano żadnego obiektu, który można by uznać za galaktykę powstającą. Oznacza to, iż epoka formowania się galaktyk już się skończyła i że wszystkie galaktyki mają mniej więcej ten sam wiek. Miejsce zajmowane przez galaktykę na ciągu Hubble'a zależy tylko od warunków panujących w czasie i miejscu jej powstania, a nie od czasu, jaki upłynął od chwili powstania. Jej typ morfologiczny zostaje określony przez warunki początkowe i pozostaje niezmieniony w trakcie ewolucji. Galaktyki powstały najprawdopodobniej wskutek istnienia niestabilności w jednorodnym pierwotnie ośrodku kosmicznym. Innej możliwości zaproponować obecnie nie można, chociaż żadna z licznych teorii kosmologicznych nie potrafiła dotychczas podać mechanizmu powstawania takich niejednorodności ośrodka, których masy byłyby równe masom galaktyk. Najmniejsze niejednorodności otrzymywane w różnych modelach Wszechświata mają masy bogatych gromad galaktyk. Problem powstawania niejednorodności o odpowiednich masach (tzw. protogalaktyk) jest nadal otwarty. Pomijając tę zasadniczą trudność dalszy ciąg ewolucji można opisać dość logicznie, choć na razie tylko jakościowo. Protogalaktyka jest opisana takimi parametrami i funkcjami jak: masa, rozkłady momentu pędu, gęstości i temperatury, pole magnetyczne oraz turbulencja (czyli mikroskopowe ruchy wirowe, powstające samorzutnie na pewnym etapie ewolucji ośrodka kosmicznego). Niestabilności turbulentne zabezpieczają protogalaktyki przed nieograniczonym zapadaniem się grawitacyjnym, mogącym doprowadzić do powstania czarnej dziury i dają początek wszechobecnemu w galaktykach ruchowi obrotowemu. Im większa jest masa protogalaktyki, tym szybsza jest kontrakcja i tym łatwiej dochodzi do rozpadu protogalaktyki na mniejsze obiekty, dające być może początek galaktykom satelitarnym i gromadom kulistym. Pole magnetyczne utrudnia rozpad protogalaktyki, zaś wysoka temperatura i duża turbulencja spowalniają kontrakcję Wyzwalająca się podczas kontrakcji energia grawitacyjna jest przekształcana w energię termiczną i promienistą. Część jej zostaje zużyta na dysocjację cząstek i jonizację atomów. Jednocześnie trwa proces separacji materii obdarzonej różnymi wartościami momentu pędu. W centralnych częściach protogalaktyki, które zapadają się najszybciej, grupuje się materia o małej energii i niewielkim momencie pędu. Materia o dużych wartościach momentu pędu (o ile w ogóle znajduje się w protogalaktyce) tworzy dysk otaczający kuliste centrum. W całej masie protogalaktyki lokalne niestabilności grawitacyjne prowadzą do powstania gwiazd o bardzo dużych masach i bardzo krótkim czasie życia. Promieniowanie ich jest dodatkowym czynnikiem hamującym kontrakcję i źródłem wtórnych niestabilności, które zapoczątkowują powstawanie nowych gwiazd, o znacznie mniejszych masach. Są to najstarsze z obecnie obserwowanych gwiazd II populacji.

protogalaktyki

Okres kontrakcji protogalaktyki i formowania się gwiazd II populacji jest bardzo krótki w porównaniu z resztą życia galaktyki i wynosi zaledwie 100 mln lat. Po tym okresie typ morfologiczny galaktyki jest już ustalony. Wiek galaktyk ocenia się na 10–15 miliardów lat.

wiek galaktyk

Dobierając różne zestawy parametrów początkowych można otrzymać modele galaktyk różniące się kształtem, zawartością materii międzygwiazdowej i składem populacyjnym, czyli odtworzyć wszystkie obserwowane typy morfologiczne. Przedstawiony obraz ewolucji ma jednak zasadniczą wadę: nie znajdują w nim miejsca jądra galaktyk. Okresy aktywności jąder i wytwarzana w nich energia nie mogą nie

zaważyć na przebiegu ewolucji. Dlatego też wszystko, co się obecnie na temat ewolucji galaktyk mówi, należy traktować jako pierwsze przybliżenie, z pewnością dalekie jeszcze od rzeczywistości.

T.A. Agiekjan *Zwiezdy, Galaktiki, Metagalaktika*, Moskwa 1970; F. Hoyle *Granice Astronomii*, Warszawa 1967; D. Mihalas, *Galactic Astronomy*, San Francisco 1968; T. Page, L. W. Page *Beyond the Milky Way*, New York 1969; G. Setti *Structure and Evolution of Galaxies*, Dordrecht 1975; G. L. Verschuur, K. I. Kellermann *Galactic and Extragalactic Radio Astronomy*, New York 1974; C. A. Whitney *The Discovery of Our Galaxy*, New York 1971.

Gwiazdy zmienne pulsujące

Kazimierz Stępień

Prowadząc systematyczne obserwacje gwiazd astronomowie zauważyli, że jasność wielu z nich zmienia się w sposób regularny. W początkach XIX w. znano już kilkanaście takich gwiazd i sądzono, że są to układy dwóch gwiazd krążących wokół siebie, i okresowo zasłaniających jedna drugą. Nazwano je gwiazdami podwójnymi zaćmieniowymi. W drugiej połowie XIX w. podjęto nowy typ obserwacji — fotograficzne badania widm gwiazdowych. Okazało się wówczas, że linie widmowe niemal wszystkich badanych gwiazd zmiennych ulegają okresowym przesunięciom wywołanym efektem Dopplera, potwierdzając hipotezę zaćmieniowości.

obserwacje δ Cephei

Obserwacje jednej gwiazdy — δ Cephei — wykazywały wprawdzie również okresowy efekt Dopplera, ale nie były zgodne z modelem gwiazdy podwójnej. W początkach XX w., gdy umiano już określać rozmiary gwiazd, okazało się z obliczeń, że hipotetyczna orbita δ Cephei musiałaby być 10 razy mniejsza od jej promienia. Ta sprzeczność ostatecznie obaliła hipotezę o podwójności δ Cephei. A więc przesunięcie linii widmowych oznaczało okresowe rozszerzanie się i kurczenie gwiazdy. W ten sposób odkryto nową klasę gwiazd zmiennych — gwiazdy pulsujące. Określimy je jako gwiazdy, które zmieniają swoją jasność oraz inne parametry fizyczne w sposób regularny, spowodowany zmianami swoich rozmiarów lub kształtu. W późniejszych latach okazało się, że istnieje wiele typów gwiazd pulsujących zasadniczo różniących się od siebie. Nazwę cefeidy z dodatkiem klasyczne zachowano dla gwiazd pulsujących podobnych do δ Cephei. Ponadto wprowadzono nazwy: cefeidy typu *W* Wirginis, gwiazdy typu *RR* Lyrae, cefeidy karłowate i gwiazdy typu δ Scuti. Wszystkie te gwiazdy leżą w tzw. klasycznym pasie niestabilności (o czym powiemy dalej), pulsują radialnie i one będą głównym tematem niniejszego artykułu. Oprócz nich znamy jeszcze gwiazdy typu β Cephei, których pulsacje polegają na zmianie kształtu (pulsacje nieradialne), a ich mechanizm jest zupełnie inny niż u klasycznych gwiazd pulsujących. Wspomnimy jeszcze krótko o gwiazdach typu *RV* Tauri i typu Mira Ceti, które też zmieniają swoje rozmiary, ale zmiany ich nie są regularne i powodowane mało znanym mechanizmem.

Dane obserwacyjne

Informacje zawarte w tym paragrafie pochodzą z obserwacji fotometrycznych i widmowych. Obserwacje fotometryczne polegają na pomiarach w różnych barwach strumienia promieniowania docierającego do nas od gwiazdy i jego zmian czasowych (krzywa zmian jasności). Światło gwiazdy można również rozszczepić za pomocą pryzmatu lub siatki dyfrakcyjnej i zarejestrować np. na kliszy fotograficznej. Otrzymując wiele takich widm dla gwiazdy pulsującej wyznaczymy skład chemiczny, temperaturę i krzywą zmian prędkości radialnej. Przy wykorzystaniu modeli gwiazd oba rodzaje obserwacji umożliwiają wyznaczenie promieni, mas i jasności absolutnych (całkowitej ilości energii wypromieniowanej przez gwiazdę w ciągu sekundy).

Omówimy teraz poszczególne typy gwiazd pulsujących. Ich średnie parametry fizyczne podaje pierwsza tabela. Jasność absolutna, masa i promień podane są w stosunku do jasności, masy i promienia Słońca, traktowanych jako jednostki. W drugiej tabeli podane są typowe amplitudy zmian parametrów fizycznych gwiazd pulsujących.

cefeidy klasyczne

Cefeidy klasyczne. Należą do najjaśniejszych gwiazd nieba. Są nadolbrzymami I populacji (→ Galaktyki), a więc zawartość pierwiastków ciężkich w ich atmosferach sięga paru procent. Amplituda zmian blasku wynosi najczęściej 1^m–$1\overset{m}{,}5$ czyli gwiazda jest w maksimum 3–4 razy jaśniejsza niż w minimum, chociaż istnieją też cefeidy o amplitudzie zaledwie kilku setnych wielkości gwiazdowej (np. Gwiazda Polarna). Druga tabela podaje również amplitudy zmian innych parametrów fizycznych, a rys. 1 przedstawia te zmiany w formie graficznej.

W 1912 r. amer. astronom H. S. Leavitt odkryła ważną zależność między okresem zmian blasku i jasności absolutnej cefeid. Zależność ta w swej współczesnej postaci jest przedstawiona na rys. 2. Jeżeli więc kiedykolwiek zaobserwujemy cefeidę i wyznaczymy dla niej okres (co można zrobić bez trudu z dużą dokładnością), to z tej zależności możemy od razu odczytać, jaką ma ona absolutną wielkość gwiazdową *M*. Z drugiej strony, łatwo możemy wyznaczyć widomą

Średnie parametry fizyczne gwiazd pulsujących radialnie

Wielkość	Cefeidy klasyczne	Cefeidy typu *W* Wirginis	*RR* Lyrae	Cefeidy karłowate	Gwiazdy typu δ Scuti
Okres (dób)	1–50	1–50	0,25–1	0,06–0,2	0,06–0,2
Jasność absolutna ($L/L_{\odot}$)	380–31000	100–8000	40	2–6	2–6
Masa ($M/M_{\odot}$)	3,7–14	0,5–1,0	0,5–0,7	?	1,5–2,5
Promień ($R/R_{\odot}$)	14–200	14–200	4–10	2–4	2–4
Temperatura (K)	6900–5400	6500	7400–6400	7500	7500

Typowe amplitudy zmian parametrów fizycznych gwiazd pulsujących radialnie

Amplituda	Cefeidy klasyczne	Cefeidy typu *W* Wirginis	*RR* Lyrae	Cefeidy karłowate	Gwiazdy typu δ Scuti
Jasności (wielkość gwiazdowa)	1–1,5	1–1,5	0,5–1,8	0,2–1,0	0,05–0,10
Prędkości radialnej (km/s)	40	60	20–60	20–40	10–20
Temperatury (K)	1000	1000	500–1500	500	< 500
Promienia (%)	10–15	20	5–15	5	< 5

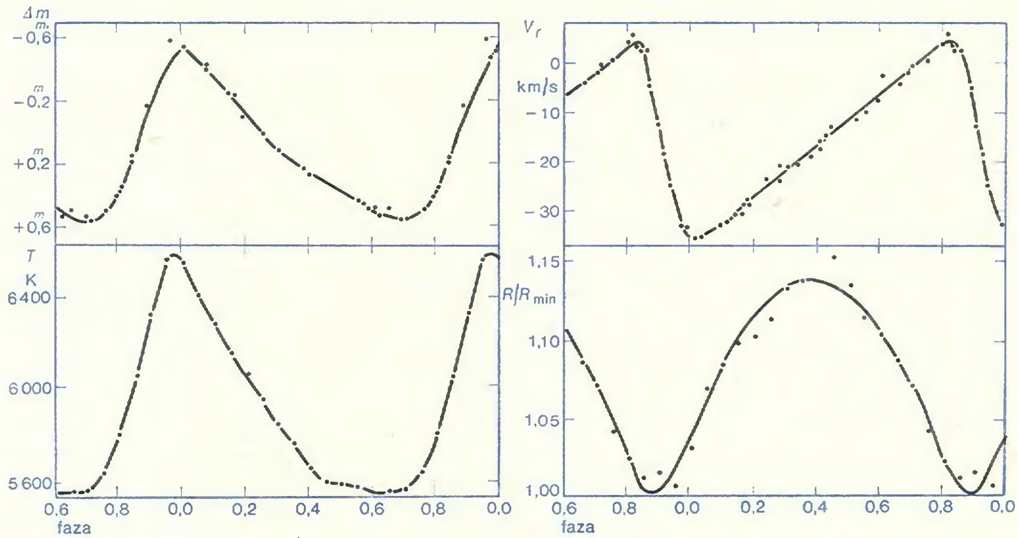
wielkość gwiazdową m. Obie te wielkości powiązane są ze sobą prostą zależnością:

$$m = M + 5\lg r - 5 + A(r).$$

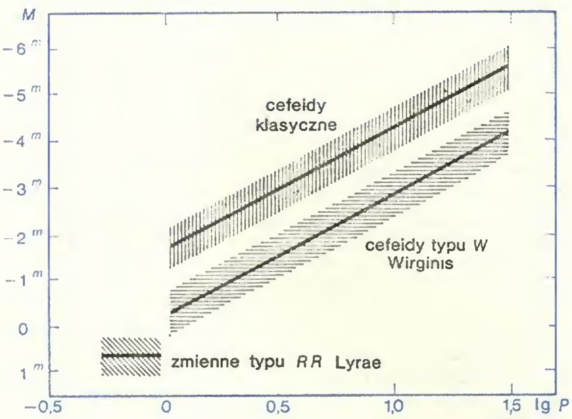
We wzorze tym r jest odległością do gwiazdy, a A jest wartością ekstynkcji międzygwiazdowej. Ekstynkcja ta wywołana jest przez materię międzygwiazdową, która pochłania „po drodze" część światła wysyłanego przez gwiazdę. Wielkość A wyznacza się indywidualnie dla każdej gwiazdy. Dla cefeid istnieje prosta i dokładna metoda wyznaczania A z pomiarów jasności gwiazdy w określonych obszarach widmowych wycinanych za pomocą filtrów barwnych. W ten sposób znając m, M i A możemy wyznaczyć odległość do badanej cefeidy. Ponieważ, jak wspominaliśmy, cefeidy należą do najjaśniejszych gwiazd nieba, widoczne są z olbrzymich odległości. Wiele z nich obserwujemy w sąsiednich galaktykach i dzięki zależności okres–jasność można było wyznaczyć ich odległości. Wszechświata. Zatem związek między jasnością i okresem cefeid jest podstawą wyznaczania skali odległości we Wszechświecie. W ciągu ostatnich kilkudziesięciu lat związek ten był wielokrotnie rewidowany i poprawiany w miarę napływu coraz dokładniejszych obserwacji, a wraz z nim odpowiednio „kurczyła się" i „rozszerzała" znana nam część Wszechświata.

cefeidy typu *W* Wirginis

Cefeidy typu *W* Wirginis. Przez długi czas cefeidy tego typu rozpatrywano łącznie z cefeidami klasycznymi. Spowodowane to było faktem, że obserwacyjnie wyglądają one bardzo podobnie. Okresy i amplitudy zmian blasku są takie same, a kształt krzywych blasku zbliżony, co przy fotograficznych metodach wykrywania cefeid uniemożliwiało rozróżnienie obydwu typów zmiennych. Łączne ich traktowanie spowodowało olbrzymi rozrzut na diagramie okres–jasność i wydawało się, że tak niedokładna zależność będzie bezużyteczna. Dopiero dokładniejsze obserwacje widmowe wykazały, że między cefeidami klasycznymi i cefeidami typu *W* Wirginis istnieją duże



Rys. 1. Zmiany jasności Δm (w wielkościach gwiazdowych), temperatury T (w kelwinach) i prędkości radialnej V_r (w km/s) oraz obliczone na ich podstawie zmiany promienia R gwiazdy δ Cephei. Punkty odpowiadają obserwacjom w różnych fazach okresu pulsacji. Faza podana jest w ułamkach okresu



Rys. 2. Zależność między logarytmem okresu P i absolutną wielkością gwiazdową M dla cefeid i gwiazd typu *RR* Lyrae. Zakreskowane są obszary, w których układają się punkty obserwacyjne

Obserwacje tych najbliższych galaktyk pozwoliły następnie znaleźć zależności między prędkością ucieczki galaktyki mierzonej z przesunięcia prążków w widmie gwiazdy ku czerwieni, a odległością do niej. Jest to tzw. prawo Hubble'a. Prawo Hubble'a stosuje się do wyznaczania odległości dalszych obiektów

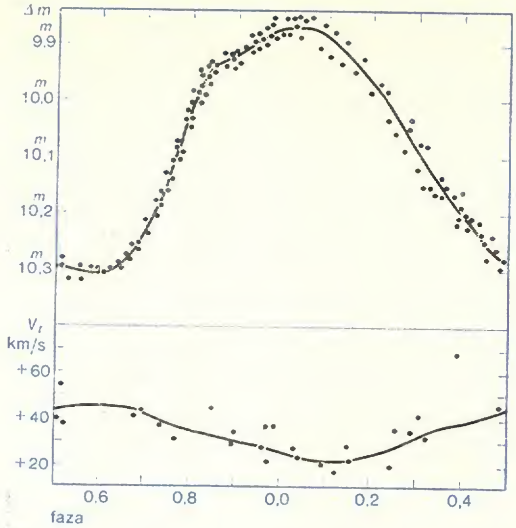
prawo Hubble'a

różnice. Krzywe prędkości radialnych tych ostatnich wyglądają zupełnie inaczej, gdyż mają nieciągłość spowodowaną okresowym rozszczepianiem się niektórych linii widmowych na dwa składniki o różnych prędkościach. Analiza widmowa oraz analiza rozkładu tych gwiazd w Galaktyce wykazały, że należą one do II populacji i są bardzo ubogie w pierwiastki ciężkie, gdyż ich zawartość w atmosferach gwiazd typu *W* Wirginis jest rzędu 0,1%. Po rozróżnieniu obu typów na diagramie okres–jasność okazało się, że dają one dwie zależności odległe o ok. $1^{m}.4$ i że wewnątrz każdej zależności rozrzut jest już niewielki (rys. 2).

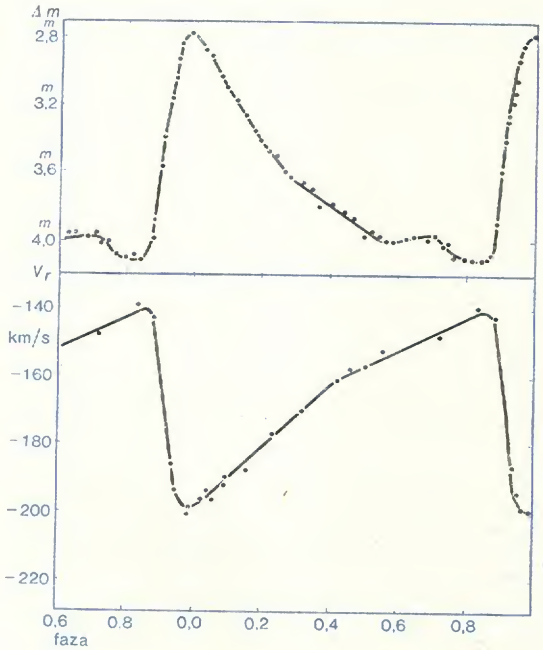
gwiazdy typu *RR* Lyrae

Gwiazdy typu *RR* Lyrae. Są to podolbrzymy II populacji. Mają znacznie krótsze okresy od cefeid i rozróżniamy wśród nich dwa podstawowe typy określane symbolami *ab* i *c*. Symbol *ab* pochodzi stąd, że pierwotnie rozróżniano trzy typy: *a*, *b* i *c*, lecz później okazało się, że gwiazdy typów *a* i *b* nie różnią się między sobą fizycznie. Zmienne typu *c* mają okresy zawarte pomiędzy $0^{d}.25$ i $0^{d}.4$ i niemal sinusoidalne krzywe blasku o amplitudzie około $0^{m}.5$ (rys. 3). Zmienne typu *ab* mają okresy między $0^{d}.3$ i $1^{d}.0$ i asymetryczne krzywe blasku o amplitudach od $0^{m}.7$ do $1^{m}.8$ (rys. 4).

Niektóre z gwiazd typu *ab* wykazują długookresowe zmiany kształtu krzywej blasku, a nawet okresu podstawowego (tj. okresu pulsacji). Przykładowo, gwiazda *RR* Lyrae — prototyp omawianej grupy zmiennych — ma podstawowy okres zmian nieco ponad pół



Rys. 3. Krzywa zmian jasności i prędkości radialnej *T* Sextansis, gwiazdy typu *RRc*



Rys. 4. Krzywa zmian jasności i prędkości radialnej *SU* Draconis, gwiazdy typu *RRab*

doby, lecz długość tego okresu oraz kształt i amplituda krzywych blasku i prędkości radialnej zmieniają się z okresem około 41^d. Przyczyny tych długookresowych zmian, określanych mianem efektu Błażko, są dotychczas zupełnie nieznane.

cefeidy karłowate

Cefeidy karłowate albo gwiazdy typu *AI* Velorum. Zmienne te mają jeszcze krótsze okresy niż gwiazdy typu *RR* Lyrae. Są gwiazdami populacji pośredniej. Wykazują silny efekt Błażko. Niestety, znamy stosunkowo niewiele gwiazd należących niewątpliwie do tego typu i dlatego nasza wiedza o parametrach fizycznych tych gwiazd jest dość fragmentaryczna. Najbardziej niepewne są wartości ich mas, gdyż zależą one od modeli teoretycznych i przyjętej fazy ewolucyjnej. Wielu astronomów uważa, że cefeidy karłowate są fizycznie identyczne z gwiazdami typu δ Scuti i podział na typy jest sztuczny. Inni jednak uważają, że są one w różnych fazach ewolucyjnych i mają różne masy. Gdyby tak było, masy cefeid karłowatych wynosiłyby zaledwie około $0,2M_\odot$.

gwiazdy typu δ Scuti

Gwiazdy typu δ Scuti. Charakterystyki obserwacyjne tych gwiazd są bardzo podobne do cefeid karłowatych — różni je tylko amplituda zmian blasku, która jest rzędu kilku procent, choć zdarzają się wyjątki. Gwiazd typu δ Scuti znamy niewiele i dlatego dane o nich są dość niepewne. Co do ich mas panuje jednak raczej zgodna opinia, że wynoszą one $1,5$–$2,5M_\odot$.

Gwiazdy typu β Cephei albo β Canis Majoris. Wszystkie omówione dotychczas typy gwiazd pulsujących tworzą jakby jedną wspólną klasę ze względu na identyczny mechanizm pulsacji, natomiast gwiazdy typu β Cephei należy rozpatrywać oddzielnie. Wiemy, że zachodzą w nich pulsacje nieradialne (polegające nie na kurczeniu się i rozszerzaniu, lecz na okresowym odkształcaniu się gwiazd). Niestety wszelkie próby wskazania mechanizmu, który mógłby takie pulsacje podtrzymywać kończyły się dotychczas niepowodzeniem. Wiemy jednak na pewno, że nie może to być ten sam mechanizm, co we wcześniej omawianych gwiazdach.

gwiazdy typu β Cephei

Gwiazdy typu β Cephei mają okresy rzędu kilku godzin (od $0^d,15$–$0^d,25$) i amplitudy zmian blasku rzędu kilku procent. Krzywe blasku zmieniają się w dłuższej skali czasowej, a zmiany te można interpretować jako nałożenie się dwu bliskich okresowości (np. w przypadku β Canis Majoris — o okresach $6^h\,00^m$ i 6^h02^m). Są one raczej gorące, gdyż ich temperatury sięgają 25 000 K, i jasne — parę tysięcy jaśniejsze od Słońca. Należą do I populacji i mają masy kilkanaście razy większe od Słońca. Absolutna wielkość gwiazdowa gwiazdy typu β Cephei rośnie przy dłuższych okresach, podobnie jak w przypadku cefeid.

gwiazdy typu *RV* Tauri i Mira Ceti

Gwiazdy typu *RV* Tauri i Mira Ceti. Jasności gwiazd należących do tych typów nie zmieniają się regularnie i dlatego nie są objęte zakresem tematycznym artykułu. Ponieważ jednak zmienność ich wiąże się zapewne ze zmianami promienia, przedstawimy krótko ich charakterystyki obserwacyjne. Są to chłodne olbrzymy i nadolbrzymy o temperaturach 2000–4000 K należące do populacji pośredniej, lub nawet II o kwaziokresach rzędu 100 dni (*RV* Tauri), czy rzędu roku lub więcej (Mira Ceti), dużej amplitudzie zmian blasku i masach prawdopodobnie ok. $1,5$–$2M_\odot$.

Teoria pulsacji

Postaramy się wyjaśnić dlaczego niektóre gwiazdy zmieniają okresowo swoje rozmiary, a wraz z nimi inne parametry fizyczne. Rozpatrzmy w tym celu zewnętrzne warstwy typowej gwiazdy. Na powierzchni jej panuje temperatura kilku tysięcy stopni. W tej temperaturze praktycznie cały wodór i hel są niezjonizowane. Natomiast w coraz głębszych warstwach temperatura gazu jest coraz większa. Wraz z temperaturą rośnie stopień jonizacji gazu, aż w temperaturze ok. trzydziestu tysięcy stopni cały wodór i hel jest zjonizowany. Przeszliśmy więc przez warstwę częściowej jonizacji wodoru i pierwszej jonizacji helu. Przesuwając się dalej w głąb gwiazdy natrafiamy na jeszcze wyższe temperatury, aż w okolicy 50 000 K, przechodzimy przez warstwę drugiej jonizacji helu. Oczywiście przy przesuwaniu się w głąb rośnie również, i to szybko, gęstość. Wskutek tego w warstwie drugiej jonizacji helu znajduje się znacznie więcej materii niż w rzadkiej warstwie jonizacji wodoru i pierwszej helu. W jeszcze głębszych warstwach cały wodór i hel są całkowicie pozbawione elektronów.

Temperatura każdej kolejnej warstwy określona jest przez równowagę między energią pochłoniętą od dołu i wysłaną do góry. Odbywa się to tak: we wnętrzu gwiazdy produkowana jest w reakcjach jądrowych energia w postaci wysokoenergetycznych kwantów promieniowania elektromagnetycznego. Warstwa leżąca bezpośrednio nad jądrem jest nieprzezroczysta

dla tego promieniowania i pochłania je. Aby nie ogrzewać się w nieskończoność, musi dokładnie tyle samo energii wypromieniować ile pochłonęła. Promieniuje zaś zgodnie z prawem promieniowania dla ciała doskonale czarnego, przyjmując taką temperaturę, jaka zapewnia jej równowagę cieplną. Następna warstwa jest też nieprzezroczysta i pochłania promieniowanie warstwy spodniej emitując je w nieco innym zakresie widmowym (gdyż jej temperatura jest nieco niższa niż warstwy leżącej pod nią). Proces pochłaniania i reemisji zachodzi w ten sposób w kolejnych warstwach gwiazdy mających coraz to niższą temperaturę, aż wreszcie ostatnia warstwa wysyła promieniowanie w przestrzeń.

przemiana termodynamiczna gazu zjonizowanego

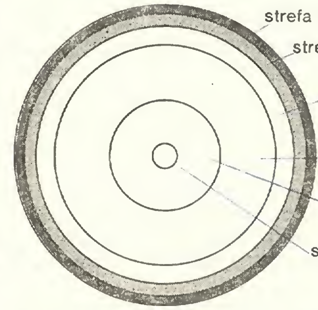
Zastanówmy się, co będzie, jeśli część gazu gwiazdowego znajdującego się w różnych miejscach gwiazdy poddamy przemianie termodynamicznej polegającej na ściśnięciu i następnie rozprężeniu. Gorący i zjonizowany gaz leżący głęboko można traktować jak gaz niemal doskonały, tak że ściśnięcie i rozprężenie go jest przemianą adiabatyczną. Przy takiej przemianie nie tracimy ani nie zyskujemy energii mechanicznej. Zachowanie się gazu z warstw zewnętrznych jest inne. Z dokładnych badań wynika, że nieprzeźroczystość maleje ze wzrostem temperatury, a ponieważ ściśnięcie gazu spowoduje jego ogrzanie, tym samym będzie on pochłaniał mniej energii niż poprzednio, natomiast emitować będzie więcej. Stan taki nie może trwać długo — gaz emitując ochłodzi się, ale do temperatury wyższej niż miał przed ściśnięciem, i przejdzie w nowy stan równowagi cieplnej, w którym tyle emituje, co pochłania. Jeżeli teraz pozwolimy mu się rozprężyć, odda nam mniej energii niż włożyliśmy w jego sprężenie. Przy wielokrotnym ściskaniu i rozprężaniu gaz będzie pochłaniał energię mechaniczną zamieniając ją na ciepło.

przemiana termodynamiczna gazu częściowo zjonizowanego

Rozpatrzmy podobnie gaz znajdujący się w warstwie częściowej jonizacji. Gdy ściśniemy go adiabatycznie, temperatura wzrośnie o wiele mniej niż w poprzednim przykładzie. Energia, którą włożyliśmy ściskając gaz, zostanie w znacznej części zamieniona na energię jonizacji wielu dodatkowych atomów, a tylko niewielka jej część wywoła wzrost temperatury. Nawet gdyby nieprzezroczystość tej warstwy malała wraz z temperaturą w tym samym stopniu co dla gazu zjonizowanego, to przebieg przemiany termodynamicznej byłby inny. Tymczasem zależność ta jest odwrotna — im gaz częściowo zjonizowany jest gorętszy, tym silniej pochłania promieniowanie. Ściśnięty gaz będzie pochłaniał coraz więcej promieniowania, będzie ogrzewał się (równocześnie będzie rósł stopień jonizacji), co wywoła jeszcze silniejsze pochłanianie itd. W efekcie energia dostarczana od dołu będzie gromadzona w warstwie częściowej jonizacji. Proces taki nie może przebiegać nieskończenie długo — ograniczą go różne zjawiska dodatkowe (tzw. efekty nieliniowe), np. fakt, że w pewnym momencie gaz zjonizuje się całkowicie. Gdy pozwolimy, by ściśnięty gaz rozprężył się i wykonał pracę, jego temperatura w trakcie rozprężania będzie niemal stała, a więc i ciśnienie będzie znacznie większe niż w przypadku gazu spoza strefy częściowej jonizacji. Dzięki temu otrzymamy więcej pracy mechanicznej niż jej włożyliśmy. Nadwyżka powstanie wskutek przemiany: energia promienista — energia jonizacji — energia wewnętrzna — energia mechaniczna. Wystarczy więc, by warstwę częściowej jonizacji ścisnąć dowolnie mało, a rozpręży się znacznie silniej. W gwieździe, jak w każdym układzie fizycznym, wszystkie wielkości fizyczne ulegają stale drobnym wahaniom, czyli tzw. fluktuacjom — nietrudno zatem o powstanie „pierwszego pchnięcia". Rozprężająca się warstwa częściowej jonizacji ciśnie na warstwy leżące pod nią i unosi w górę warstwy wierzchnie (też zresztą je ściskając). Obydwie te warstwy pochłaniają część przekazanej im energii, a jednocześnie rozprężając się cisną ponownie na warstwę częściowej jonizacji, która oddaje im energię mechaniczną z nadwyżką powiększając amplitudę wahań. W ten sposób będą stopniowo narastać pulsacje. Oczywiście ich amplituda nie będzie wzrastać nieskończenie. Wystąpią różne efekty nieliniowe i w końcu ustali się taka amplituda pulsacji warstw zewnętrznych gwiazdy, przy której energia generowana w warstwie częściowej jonizacji będzie dokładnie równa energii pochłanianej w innych warstwach biorących udział w ruchu.

warstwy pulsujące

W pulsacjach radialnych biorą udział tylko warstwy zewnętrzne (rys. 5). Amplituda drgań szybko maleje przy przechodzeniu do głębszych warstw, tak że wnętrze gwiazdy nie bierze w ogóle udziału w ruchu. We wnętrzu jest skupiona niemal cała masa gwiazdy, a w warstwach biorących udział w ruchu znajduje się zaledwie 0,1% masy całej gwiazdy, mimo iż rozciągają się one do połowy promienia gwiazdy.



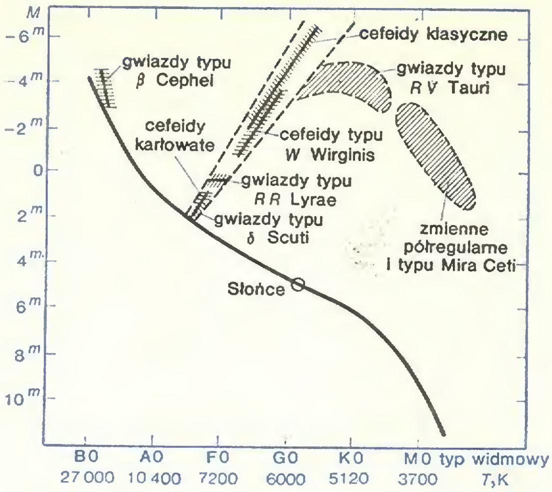
Rys. 5. Rozmieszczenie warstw biorących udział w pulsacjach gwiazdy pulsującej

warunki pulsowania gwiazd

Narzuca się pytanie, dlaczego nie pulsują wszystkie gwiazdy? Przecież wszystkie mają warstwy częściowej drugiej jonizacji helu, a większość (takie, które nie są zbyt gorące na powierzchni) warstwę częściowej jonizacji wodoru i pierwszej helu. Jeżeli warstwy te są niestabilne pulsacyjnie, to powinny pobudzić do pulsacji każdą gwiazdę. Odpowiedź na to pytanie dają dokładne obliczenia modeli gwiazd pulsujących. Wynika z nich, że decydująca dla niestabilności pulsacyjnej gwiazdy jest głębokość, na jakiej występuje warstwa drugiej jonizacji helu. W gwiazdach gorących leży ona płytko pod powierzchnią i obejmuje niewielką ilość masy, dlatego jej pojemność cieplna (określona przede wszystkim energią jonizacyjną wszystkich atomów tej warstwy) jest zbyt mała, by móc wytworzyć dostateczną ilość energii mechanicznej zdolnej do „rozhuśtania" sporej części gwiazdy i zrównoważenia strat energii. Im gwiazda jest chłodniejsza, tym głębiej znajduje się jej warstwa jonizacyjna i zdolność gwiazdy do wytwarzania energii mechanicznej szybko rośnie. Przy pewnej temperaturze energia ta wystarcza do wywołania pulsacji gwiazdy. Przy jeszcze niższej temperaturze gwiazdy warstwa jonizacyjna znajduje się jeszcze głębiej, ale ilość energii mechanicznej generowanej przez nią rośnie znacznie wolniej, natomiast straty energii stają się wyraźnie większe. Warstwa jonizacyjna znów nie jest w stanie wywołać pulsacji gwiazdy. A więc pulsują tylko te gwiazdy, w których warstwa częściowej jonizacji helu znajduje się w ściśle określonym zakresie głębokości, czyli tylko w ściśle określonym zakresie temperatur gwiazd. Zakres ten różni się nieco dla różnych wartości jasności absolutnych, co daje na tzw. diagramie Hertzsprunga–Russella, na którego osiach mamy jasność i temperatury gwiazd, skośnie nachylony pas niestabilności (rys. 6).

pas niestabilności

Gdy w czasie ewolucji gwiazda osiągnie wartość temperatury i jasności z pasa niestabilności, zacznie pulsować. Wychodząc z tego pasa stanie się znów gwiazdą o stałym blasku. Należy dodać, że warstwa częściowej jonizacji wodoru i pierwszej helu znajduje się zawsze bardzo płytko, o ile w ogóle istnieje, i dlatego jej rola w podtrzymywaniu pulsacji jest niewielka. Być może w niektórych typach gwiazd (typu *RV* Tauri,



Rys. 6. Diagram Hertzsprunga–Russella z zaznaczonymi granicami pasa niestabilności oraz położeniem głównych typów gwiazd pulsujących

czy Mira Ceti) warstwa ta ma decydujące znaczenie dla ich zmienności, ale przypuszczenie to nie jest dotychczas poparte odpowiednimi obliczeniami modelowymi.

okres pulsacji

Zastanówmy się jeszcze, od czego zależy okres pulsacji. Wiemy, że dla struny drgającej okres zależy od jej długości i siły naciągu. Analogami tych wielkości fizycznych dla gwiazdy jest jej promień R i siła grawitacji wywołana jej masą M. Z obliczeń wynika, że okres zależy od stosunku M/R^3, który z dokładnością do stałej jest równy średniej gęstości gwiazdy $\bar{\varrho}_*$. Mamy więc:

$$P \sim 1/\bar{\varrho}_*,$$

gdzie P jest okresem pulsacji. Wyznaczając go w dniach i odnosząc średnią gęstość gwiazdy do średniej gęstości Słońca dostajemy:

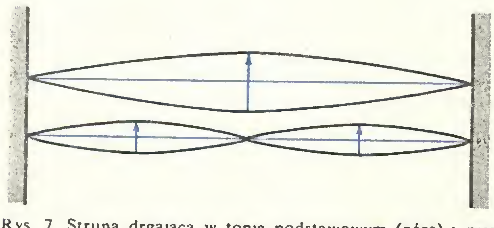
$$P = Q\sqrt{\bar{\varrho}_\odot/\bar{\varrho}_*}.$$

stała pulsacyjna

Stała proporcjonalności Q wyrażona jest w dniach i nazwana stałą pulsacyjną. Zależy ona do wewnętrznej struktury gwiazdy, ale zmienia się niewiele od gwiazdy do gwiazdy.

Patrząc na ten wzór łatwo zrozumiemy, dlaczego cefeidy — silnie rozdęte nadolbrzymy o małej gęstości średniej — mają tak długie okresy pulsacji, a np. gwiazdy typu δ Scuti o rozmiarach kilkadziesiąt razy mniejszych, a więc i odpowiednio większej gęstości średniej, mają okresy zaledwie rzędu kilku godzin. Gdybyśmy do ostatniego wzoru wprowadzili średnią gęstość białego karła — gwiazdy o masie porównywalnej ze Słońcem i promieniu porównywalnym z Ziemią — to okazałoby się, że okres jego pulsacji należy mierzyć w sekundach.

Przypomnijmy sobie jeszcze jedną własność drgającej struny. Może ona drgać w tonie podstawowym, gdy na strunie mamy dwa węzły w punktach zamocowania i strzałkę pośrodku, oraz w tonie wyższym (tzw. pierwszym tonie harmonicznym), gdy mamy



Rys 7. Struna drgająca w tonie podstawowym (góra) i pierwszym tonie harmonicznym (dół)

trzeci węzeł pośrodku i dwie strzałki (rys. 7). Okres drgań będzie wtedy oczywiście krótszy. Gwiazdy również mogą pulsować zarówno w tonie podstawowym, jak i w pierwszym, a nawet wyższych tonach harmonicznych. Jeżeli okres podstawowy wynosi dla gwiazdy P_0, to okres drgań w pierwszym tonie harmonicznym wynosi w przybliżeniu:

$$P = {}^3/_4 P_0 .$$

Który ton jest najbardziej niestabilny (i który tym samym „wybierze" gwiazda), zależy od jej budowy wewnętrznej. Gwiazdy *RR* Lyrae typu *c* pulsują w pierwszym tonie harmonicznym, podczas gdy typu *ab* w tonie podstawowym. Oczywiście amplitudy wszystkich zmian w gwiazdach typu *c* są mniejsze niż w gwiazdach typu *ab*. Podany powyżej związek między okresem pulsacji i średnią gęstością jest słuszny dla pulsacji w dowolnym tonie harmonicznym, tylko stałą Q należy odpowiednio przeskalować.

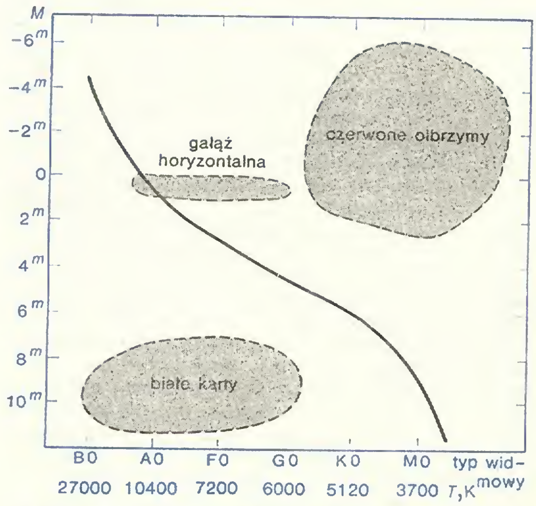
Faza ewolucyjna gwiazd pulsujących

Drogę ewolucyjną gwiazdy najlepiej można prześledzić na diagramie Hertzsprunga–Russella (H–R). Na osi odciętych tego diagramu mamy temperaturę gwiazdy (w skali logarytmicznej) lub inny parametr związany z logarytmem temperatury (np. typ widmowy). Na drugiej osi mamy jasność absolutną gwiazdy, również w skali logarytmicznej, lub absolutną wielkość gwiazdową, proporcjonalną do logarytmu jasności gwiazdy. Jasność i temperatura dobrze opisują gwiazdę, a równocześnie najłatwiej porównać je z obserwacjami. Gwiazda o danej jasności i temperaturze jest przedstawiona na diagramie H–R przez punkt. Gdy w czasie ewolucji gwiazdy parametry te zmieniają się, punkt zatacza krzywą zwaną torem ewolucyjnym.

tor ewolucyjny

Typowa gwiazda „spędza" większość swego życia w fazie, w której w jej wnętrzu zachodzi przemiana wodoru w hel. W tej fazie jasność i temperatura niemal nie zmieniają się i są jednoznacznie określone przez masę gwiazdy. Punkty odpowiadające gwiazdom o różnych masach nie są chaotycznie rozsypane po całym diagramie, lecz układają się w jedną linię nazwaną ciągiem głównym (rys. 8). Czas przebywania gwiazdy na ciągu głównym zależy od jej masy i wynosi

ciąg główny



Rys. 8. Diagram Hertzsprunga–Russella

np. od 100 000 lat dla gwiazd masywnych (rzędu $20M_\odot$) do 10 miliardów lat dla gwiazd o masie rzędu masy Słońca.

Gdy we wnętrzu gwiazdy wypali się wodór, wchodzi ona w burzliwy okres ewolucji, w którym kolejne

fazy trwają znacznie krócej niż czas życia na ciągu głównym. Losy gwiazdy zależą teraz bardzo od jej masy. Gwiazdy masywne żyją znacznie krócej niż gwiazdy lżejsze, zatem gwiazdy o masach równych wielu masom Słońca, które powstały wiele miliardów lat temu, są już bardzo stare w sensie ewolucyjnym i dawno zakończyły swą drogę rozwoju stając się białym karłem, gwiazdą neutronową lub czarną dziurą (→ Czarne dziury i zapadanie grawitacyjne). Jeżeli obecnie obserwujemy gwiazdy masywne (a do takich należą np. cefeidy klasyczne), to na pewno są to gwiazdy, które powstały stosunkowo niedawno (ok. kilkudziesięciu milionów lat temu). Mimo to są one znacznie bardziej zaawansowane ewolucyjnie niż np. Słońce mające wiek rzędu 5 miliardów lat.

obszar czerwonych olbrzymów

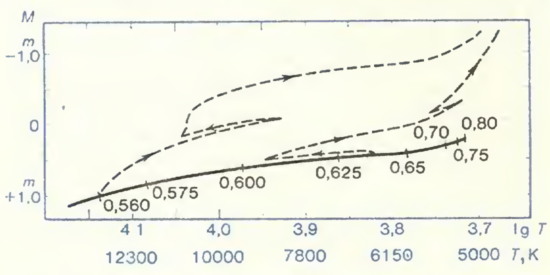
Wszystkie gwiazdy, niezależnie od masy, po wypaleniu wodoru w centrum „wędrują" na diagramie H–R na prawo od ciągu głównego, do obszaru czerwonych olbrzymów. W tym czasie następuje przebudowa wnętrza gwiazdy — jądro pozbawione wodoru kurczy się, a warstwy zewnętrzne rozszerzają się. Spalanie wodoru zachodzi w tym czasie w warstwie otaczającej jądro. Czas przejścia do fazy czerwonego olbrzyma jest bardzo krótki dla gwiazdy masywnej i dlatego szansa zaobserwowania jej w tej fazie jest znikoma. Czas ten jest znacznie dłuższy dla gwiazd o masach równych 1–2 masom Słońca. Zwróćmy uwagę, że w czasie swej wędrówki ku obszarowi diagramu H–R, zajmowanemu przez czerwone olbrzymy, gwiazdy przecinają pas niestabilności, czyli powinny przechodzić przez fazę gwiazdy pulsującej. Nie udało się nam dotychczas zaobserwować masywnych gwiazd pulsujących w tym stadium ewolucyjnym, natomiast wszystkie dane obserwacyjne i rachunki modelowe wskazują na to, że gwiazdy typu δ Scuti są gwiazdami odchodzącymi od ciągu głównego i spalającymi wodór w grubej warstwie otaczającej bezwodorowe jądro. Byłyby to więc najmłodsze ewolucyjnie gwiazdy pulsujące.

ewolucja gwiazd mało masywnych

Przyjrzyjmy się teraz dalszej ewolucji mało masywnych gwiazd, o masach nie przewyższających jednej masy Słońca. Gdy znajdą się w obszarze czerwonych olbrzymów powiększają swe rozmiary dalej, podczas gdy jądro kurczy się, aż w centrum powstanie temperatura dostatecznie wysoka do zapoczątkowania reakcji przemiany helu w węgiel. W fazie ewolucyjnej czerwonego olbrzyma siła przyciągania na powierzchni gwiazdy jest bardzo mała, a obserwacje wskazują na silne ruchy chaotyczne w atmosferze i stąd przypuszcza się, że gwiazda traci w tej fazie sporo masy „posiewając" ją ze swej powierzchni. Nie mamy wprawdzie ścisłych obliczeń, które powiedziałyby nam dokładnie, ile masy i w jakim tempie będzie tracone, ale sam fakt utraty masy nie budzi obecnie wątpliwości.

gałąź horyzontalna

W momencie zapalenia się helu (błysk helowy) gwiazda robi nagły zwrot na diagramie H–R, kurczy się nieco i trafia na tzw. gałąź horyzontalną (rys. 9). Gałąź horyzontalna rozciąga się od obszaru czerwonych olbrzymów, przecina pas niestabilności i sięga daleko na lewo od tego pasa. Miejsce, na które trafia



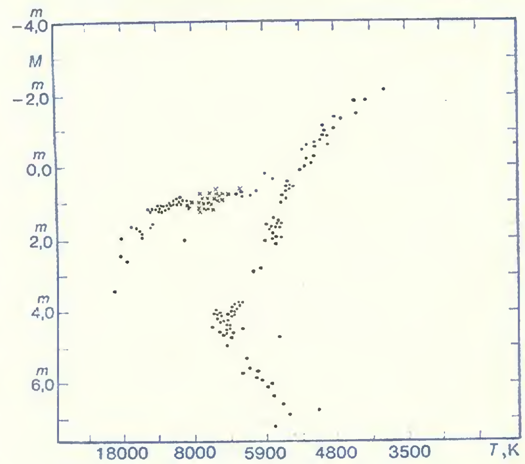
Rys. 9. Fragment diagramu H–R z teoretyczną gałęzią horyzontalną (linia ciągła) oraz torami (linie przerywane) pokazującymi dalszą ewolucję gwiazd. Liczby oznaczają masy gwiazd (w jednostkach masy Słońca) lądujących w różnych miejscach gałęzi horyzontalnej

gwiazda zależy od wielu czynników: masy jądra helowego, całkowitej masy gwiazdy i składu chemicznego (a szczególnie zawartości pierwiastków ciężkich). Gdy utworzymy diagram H–R dla gromady kulistej (gęstego skupiska setek tysięcy gwiazd, które powstały równocześnie wiele miliardów lat temu) widzimy, że część gwiazd rzeczywiście układa się w gałąź horyzontalną (rys. 10). Porównanie obserwowanej pozycji gwiazd gałęzi horyzontalnej na diagramie H–R z modelami teoretycznymi wskazuje, że masy tych gwiazd wynoszą 0,6–0,7$M_\odot$.

diagram H–R dla gromady M 15

Gwiazdy znajdujące się na gałęzi horyzontalnej spalają we wnętrzu jądra hel, a w warstwie otaczającej jądro — wodór. Ta faza ewolucyjna jest względnie spokojna — gwiazda przez dłuższy czas niemal nie zmienia swej jasności i temperatury.

Jeżeli gwiazda z gałęzi horyzontalnej znajdzie się w pasie niestabilności, powinna pulsować. Istotnie obserwujemy takie gwiazdy. Są to gwiazdy typu *RR* Lyrae. Z rys. 10 wynika, że przejście od gwiazd



Rys. 10. Diagram Hertzsprunga–Russella dla gromady kulistej *M* 15. Krzyżykami zaznaczone są gwiazdy typu *RR* Lyrae

zmiennych do stałych jest ostro zarysowane. Stąd dla wielu gromad kulistych możemy wyznaczyć obserwacyjne granice pasa niestabilności. Z danych teoretycznych wynika, że lewa granica pasa (wysokotemperaturowa) jest na danym poziomie jasności absolutnej silnie zależna od ilości helu znajdującego się w otoczce gwiazdy pulsującej. Porównując dane teoretyczne i obserwacyjne można wyznaczyć zawartość helu w gromadzie. Otrzymany wynik obarczony jest dość dużym błędem, ale nie znamy obecnie lepszych metod. Dla wszystkich zbadanych gromad otrzymano, że zawartość helu wynosi ok. 25%. Przemawia to za powstaniem helu na początku istnienia Wszechświata, a nie za stopniowym zwiększaniem się jego zawartości w miarę ewolucji.

wyznaczanie zawartości helu

wyznaczanie masy

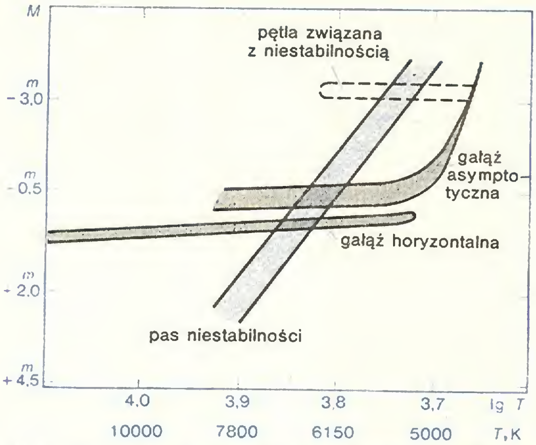
Drugim ważnym osiągnięciem teorii pulsacji jest opracowanie metody wyznaczania z obserwacji masy gwiazdy zmiennej. Daje to możliwość sprawdzenia poprawności teoretycznie wyznaczonych torów ewolucyjnych na diagramie H–R dla gwiazd o różnych masach. Okazuje się, że zgodność jest dobra, chociaż miejsce zajmowane przez gwiazdy typu *RR* Lyrae powinny zajmować gwiazdy o nieco większej masie niż to obserwujemy. Może to być potwierdzeniem faktu utraty masy w fazie czerwonego olbrzyma.

gałąź asymptotyczna

Gwiazda z gałęzi horyzontalnej przesuwa się na diagramie H–R w trakcie spalania helu w jądrze ku górze, a gdy cały hel w środku spali się i pali się tylko w warstwie otaczającej węglowe jądro, gwiazda wraca do obszaru czerwonych olbrzymów. Wędruje w tym czasie wzdłuż tzw. gałęzi asymptotycznej znajdującej się na prawo od pasa niestabilności. Z dokładnych obliczeń wynika, że w tej fazie ewolucyjnej tempo

przemian jądrowych nie jest stałe, lecz podlega okresowym niestabilnościom. Jednak obliczenia są ciągle jeszcze niewystarczające, ze względu na ograniczone możliwości maszyn matematycznych, ale z już istniejących wiemy, że mniej więcej co kilkaset tysięcy lat tempo produkcji energii w warstwie spalającej hel gwałtownie narasta (tzw. puls helowy), co ma silny wpływ na jasność i temperaturę gwiazdy. Wskutek tego gwiazda zatacza pętlę na diagramie H–R rozciągającą się poza pas niestabilności (rys. 11). Powinniśmy więc obserwować gwiazdy pulsujące o niedużych

puls helowy



Rys. 11. Schemat pętli powstającej na diagramie H–R wskutek niestabilności spalania się helu w gwiazdach z gałęzi asymptotycznej

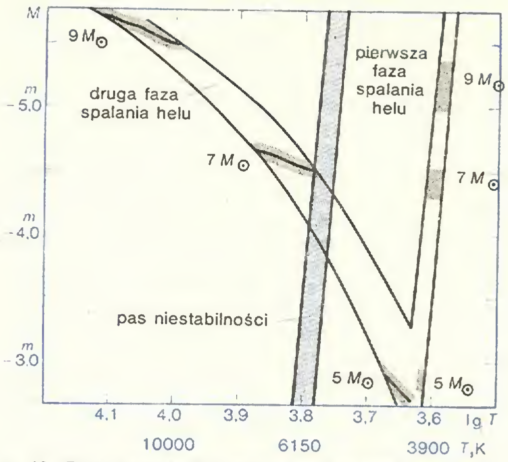
masach leżące w tym obszarze diagramu. Gwiazdami takimi są cefeidy typu *W* Wirginis i dlatego przypuszczamy, że są one w trakcie pulsu helowego. Byłyby więc ewolucyjnie starsze od gwiazd typu *RR* Lyrae. W tym miejscu warto podkreślić, że niestabilność modeli spalających wodór i hel w dwu warstwach została odkryta zaledwie ok. 10 lat temu. Do tego czasu faza ewolucyjna, w jakiej gwiazda staje się cefeidą typu *W* Wirginis była jeszcze bardziej niejasna.

Gwiazdy przechodzące do obszaru czerwonych olbrzymów wzdłuż gałęzi asymptotycznej tracą zapewne jeszcze część masy i w końcu przechodzą do obszaru białych karłów. Nie jest wykluczone, że niektóre gwiazdy typu *W* Wirginis znajdują się w trakcie tego przejścia.

Niektóre mało masywne gwiazdy mają prostszą drogę ewolucyjną. Aby w jądrze nastąpiło zapalenie się helu, musi ono mieć minimalną masę rzędu 0,45 $M_\odot$. Jeżeli gwiazda wypali w centrum wodór i przejdzie do obszaru czerwonych olbrzymów, mając masę jądra wyraźnie mniejszą niż ta minimalna wartość (np. wskutek małej masy początkowej lub silnej utraty masy w trakcie ewolucji), to hel nie zapala się, a po zakończeniu kurczenia się jądra gwiazda ewoluuje do obszaru białych karłów spalając wodór w cienkiej warstwie otaczającej helowe jądro. W tym czasie musi ponownie przeciąć pas niestabilności. Przypuszcza się, że w tej fazie ewolucyjnej znajdują się cefeidy karłowate. Jak już wspominaliśmy, wielu astronomów uważa, że podział na gwiazdy typu *δ* Scuti i cefeidy karłowate jest niepotrzebny. Jeżeli jednak różnią się one, to w tym obszarze diagramu H–R mogą wystąpić obok gwiazd typu *δ* Scuti tylko gwiazdy o masach około 0,2$M_\odot$ spalające wodór w cienkiej warstwie i ewoluujące do stadium białego karła.

Pozostało nam jeszcze wyjaśnienie fazy ewolucyjnej masywnych gwiazd pulsujących. Gwiazdy o masach równych kilku masom Słońca po przejściu do obszaru czerwonych olbrzymów również zapalają hel w jądrze. Jednak sam proces spalania helu przebiega nieco inaczej. Gwiazda najpierw spala hel będąc

ewolucja gwiazd masywnych



Rys. 12. Dwie główne fazy spalania się helu w gwiazdach masywnych

w obszarze czerwonych olbrzymów, po czym zmienia swoją budowę wewnętrzną wędrując równocześnie w lewo na diagramie H–R i wchodzi w drugą „niebieską" fazę spalania helu. Rysunek 12 pokazuje rozkład tych dwóch głównych faz dla określonego składu chemicznego odpowiadającego gwiazdom I populacji. Należy dodać, że dokładna lokalizacja „niebieskiej" fazy spalania helu zależy od wielu założeń przyjętych podczas rachunków i dlatego nie jest jeszcze zbyt dokładnie znana. Będąc w tej fazie (lub w czasie przejścia do niej) gwiazdy mogą się znaleźć w pasie niestabilności. Stają się wtedy cefeidami klasycznymi.

Jeżeli chodzi o stan zaawansowania ewolucyjnego innych gwiazd, o których wspominaliśmy, to gwiazdy typu *β* Cephei są to gwiazdy, które właśnie kończą (albo dopiero co skończyły) spalanie wodoru w centrum i „szykują się" do wędrówki do obszaru czerwonych olbrzymów. Są więc gwiazdami bardzo młodymi, również ewolucyjnie. Gwiazdy typu Mira Ceti są już w końcowych stadiach ewolucji. Wzdłuż gałęzi asymptotycznej zawędrowały do obszaru czerwonych olbrzymów i obecnie spalają wodór i hel w dwu warstwach, tracąc jednocześnie masę. Czeka je jeszcze przejście do obszaru białych karłów.

Najmniej możemy powiedzieć na temat tego, w jakiej fazie ewolucyjnej znajdują się gwiazdy typu *RV* Tauri.

test dla teorii ewolucji

Sumując, należy podkreślić, że obserwacje gwiazd pulsujących były od początku jednym z najważniejszych testów dla teorii ewolucji. Dzięki swej zmienności pozwalały na dokładne wyznaczenie parametrów fizycznych i dopiero w ostatnich latach zbudowano wystarczająco dobre modele ewolucyjne pozwalające śledzić dzieje gwiazdy krok po kroku, od narodzin aż do końca drogi ewolucyjnej, i z dostateczną dokładnością odtwarzające obserwowane wielkości gwiazd pulsujących. Jest to duży triumf teorii ewolucji. Zgodność tej teorii z obserwacjami w wypadkach, w których porównanie jest możliwe, utwierdza nas w przekonaniu, że opisuje ona również dobrze sytuacje trudniejsze do porównania z obserwacjami.

W. P. Cesewicz *Zwiozdy tipa RR Liry*, Kijew 1966; J. S. Glasby *Variable Stars*, Cambridge, Massachussetts 1969; B. W. Kukarkin (red.) *Pulsirujuszczije zwiozdy*, Moskwa 1970; W. Strohmeier *Variable Stars*, Oxford 1972; O. Struve, V. Zebergs *Astronomia XX wieku*, Warszawa 1967.

Kwazary

Marcin Kubiak

Odkrycie kwazarów było jednym z największych osiągnięć współczesnej astronomii obserwacyjnej. W końcu lat pięćdziesiątych w wielu obserwatoriach radioastronomicznych na świecie realizowane były programy obserwacyjne mające na celu sporządzenie możliwie kompletnych katalogów pozaziemskich źródeł promieniowania radiowego. Powstały wówczas katalogi oznaczane symbolami: MSH — od nazwisk australijskich radioastronomów: B. Y Millsa, O. B. Slee i E. Hilla, którzy zestawili katalog źródeł występujących na południowej półkuli nieba; PKS — katalog radioźródeł na niebie południowym sporządzony na podstawie obserwacji za pomocą radioteleskopu Parkesa; seria katalogów 1C, 2C, 3C, 4C i 5C — sporządzonych przez grupę radioastronomów z angielskiego uniwersytetu w Cambridge; CTA i CTD — katalogi zestawione przez radioastronomów z California Institute of Technology; AO — katalog sporządzony za pomocą radioteleskopu Arecibo w Puerto Rico, NRAO — katalog sporządzony w National Radio Astronomy Observatory, BI — katalog zestawiony w Bolonii.

katalogi radioźródeł

W miarę zwiększania dokładności określania współrzędnych źródeł radiowych możliwe stało się dokonywanie ich identyfikacji z obiektami optycznymi. W dużych szerokościach galaktycznych, tzn. w dużych odległościach kątowych od płaszczyzny Galaktyki, radioźródła identyfikowano zazwyczaj z galaktykami. Jednak w wielu wypadkach w miejscach radioźródeł brak było galaktyk, natomiast w pobliżu znajdowano słabe obiekty gwiazdowe. Pierwszej niewątpliwej identyfikacji źródła promieniowania radiowego z obiektem gwiazdowym dokonano w 1960 r. Źródłem tym było źródło z katalogu 3C oznaczone symbolem 3C 48, a jego odpowiednikiem optycznym był obiekt podobny do gwiazdy o jasności 16^m. Obserwacje spektroskopowe wykazały, że obiekt optyczny wysyła widmo ciągłe, na którego tle występują szerokie linie emisyjne nie dające zidentyfikować się z żadnymi liniami występującymi zazwyczaj w widmach gwiazd. Wkrótce potem dokonano identyfikacji optycznej źródła 3C 273, któremu odpowiadał obiekt o jasności $12^m_{\cdot}8$ (najjaśniejszy spośród znanych obiektów tego rodzaju). Odkryte obiekty nazwano kwazarami od angielskich słów quasi-stellar radio sources (kwazi gwiazdowe radioźródła).

Zasadniczym przełomem w historii kwazarów było zidentyfikowanie w 1963 r. linii emisyjnych w widmie kwazara 3C 273. Okazało się, że są to linie serii Balmera wodoru oraz linia magnezu o laboratoryjnej długości fali 279,8 nm, znacznie przesunięte ku czerwieni. (Przesunięciem ku czerwieni nazywamy wielkość $z = (\lambda - \lambda_0)/\lambda_0$, gdzie λ — długość fali mierzona w widmie, a λ_0 — długość fali tej samej linii mierzona w laboratorium ziemskim). Dla kwazara 3C 273 przesunięcie ku czerwieni okazało się równe 0,158. Podobna identyfikacja linii w widmie kwazara 3C 48 dała przesunięcie ku czerwieni $z = 0{,}368$. Od tej chwili liczba kwazarów o znanych przesunięciach ku czerwieni zaczęła szybko rosnąć, przy czym okazało się, że kwazary mogą mieć przesunięcia ku czerwieni większe niż obserwowano poprzednio u innych obiektów znanych astronomii. Kwazary stały się też przedmiotem intensywnych i wszechstronnych badań zarówno obserwacyjnych, jak i teoretycznych.

przesunięcie ku czerwieni

Obserwacyjne cechy kwazarów

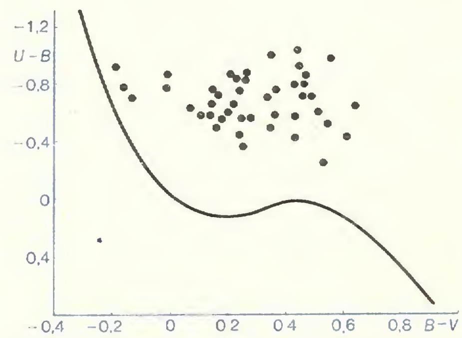
Ponieważ nie wiadomo jeszcze, czym w istocie są kwazary (znamy ich obecnie ok. kilkuset), nie można podać ich ścisłej definicji. Tym większe znaczenie ma wyodrębnienie tych cech obserwacyjnych, które odróżniają kwazary od innych obiektów astronomicznych. Obecnie znamy następujące cechy kwazarów:

cechy kwazarów

1) są to obiekty, których obraz optyczny jest taki jak gwiazdy i które są często identyfikowane z radioźródłami;
2) jasność optyczna jest zmienna;
3) promieniowanie ich ma, w porównaniu z promieniowaniem gwiazd, znaczną nadwyżkę nadfioletową i podczerwoną;
4) w ich widmach występują szerokie linie emisyjne oraz niekiedy linie absorpcyjne;
5) linie są znacznie przesunięte ku czerwieni.

Początkowo kwazary były wykrywane jako źródła promieniowania radiowego odznaczające się stosunkowo małymi rozmiarami kątowymi i nie dające zidentyfikować się ze zwykłymi radiogalaktykami. Istotną rolę w identyfikacji optycznej odgrywała dokładność, z jaką możliwe było określenie współrzędnych radiowych. Fotoelektryczne pomiary promieniowania widzialnego zidentyfikowanych kwazarów wykazały, że na diagramie dwuwskaźnikowym w systemie *UBV* (→ Astronomia w podczerwieni) zajmują one obszar leżący powyżej ciągu normalnych gwiazd Galaktyki (rys. 1), co oznacza istnienie w ich promieniowaniu dużej nadwyżki promieniowania nadfioletowego. Wykorzystując tę właściwość podjęto próby

wykrywanie i identyfikacja kwazarów



Rys. 1. Diagram dwuwskaźnikowy dla kwazarów (punkty). Linią ciągłą przedstawiona jest standardowa zależność dla gwiazd Galaktyki

wykrycia kwazarów metodami wyłącznie optycznymi przez poszukiwanie na dwubarwnych zdjęciach nieba słabych obiektów gwiazdowych mających dużą nadwyżkę nadfioletową. Należy jednak dodać, że duże nadwyżki nadfioletowe mogą mieć niektóre gwiazdy szczególnego rodzaju, takie jak białe karły, stare gwiazdy nowe lub inne gwiazdy będące w późnych stadiach ewolucji. Ostateczne potwierdzenie, czy dany obiekt jest kwazarem, mogą przynieść tylko obserwacje spektroskopowe, ujawniające ewentualną obecność linii o dużym przesunięciu ku czerwieni. Mimo to, pierwsze próby ograniczone do niewielkich obszarów nieba doprowadziły do odkrycia kwazarów metodami wyłącznie optycznymi, przy czym okazało się, że niektóre spośród nich nie dają się zidentyfikować ze znanym źródłem promieniowania radiowego. Obiekty te nazwano kwazarami spokojnymi radiowo. Oczywiście brak promieniowania radiowego może być tylko wynikiem ograniczonej czułości stosowanych obecnie odbiorników; niektóre spośród kwazarów spokojnych radiowo zostały z czasem zidentyfikowane z bardzo słabymi źródłami promieniowania radiowego. Jeżeli wyniki te są reprezentatywne dla całej populacji kwazarów, to można oczekiwać, że kwazarów spokojnych radiowo jest ok. 100 razy więcej niż kwazarów radiowych. Możliwość dalszego udoskonalenia tej metody wynika z faktu, że kwazary

kwazary spokojne radiowo

oprócz nadwyżki fioletowej mają również znaczną nadwyżkę promieniowania podczerwonego. Metoda trójbarwna, polegająca na wykrywaniu obiektów o dużych nadwyżkach nadfioletowych i podczerwonych, zmniejszy znacznie prawdopodobieństwo fałszywych identyfikacji, a jednocześnie pozwoli na ewentualne wykrycie kwazarów pozbawionych nadwyżki fioletowej wskutek dużego przesunięcia ku czerwieni. (Nadwyżka ta w wyniku efektu Dopplera przejawi się w promieniowaniu odbieranym jako nadwyżka promieniowania widzialnego lub nawet podczerwonego). Metoda ta nie przyniosła jeszcze oczekiwanych wyników.

W tym miejscu należy jeszcze dodać, że w katalogach istnieje dość dużo radioźródeł, których nie udało się dotychczas zidentyfikować z jakimikolwiek obiektami optycznymi. Przebieg zależności lg N–lg$S$ (paragraf: Interpretacja teoretyczna) dla tych źródeł sugeruje ich podobieństwo do kwazarów wysyłających promieniowanie radiowe.

optyczne obrazy kwazarów

Optyczne obrazy kwazarów są w zasadzie punktowe i nie dają odróżnić się od obrazów słabych gwiazd. Wprawdzie istnieją powody, by przypuszczać, że liniowe rozmiary kwazarów są niewielkie, to jednak trudno jest obecnie powiedzieć, w jakim stopniu brak jakiejkolwiek struktury obrazu optycznego jest rzeczywistą cechą kwazarów, a w jakim — jest wynikiem dużej odległości. Na przykład w obrazie optycznym kwazara 3C 48 (rys. 2) można wyróżnić niewielką



Rys. 2. Optyczny obraz kwazara 3C 48 (wg A. Sandage and W. C. Miller, Astrophys. J. 144, 1238, 1966)

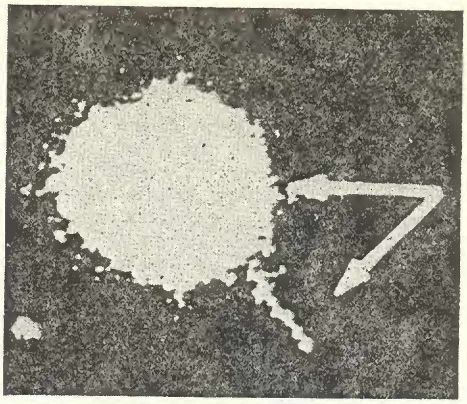
otoczkę mgławicową, a kwazar 3C 273 ma w swej budowie wyraźną strugę (rys. 3). Z drugiej strony,

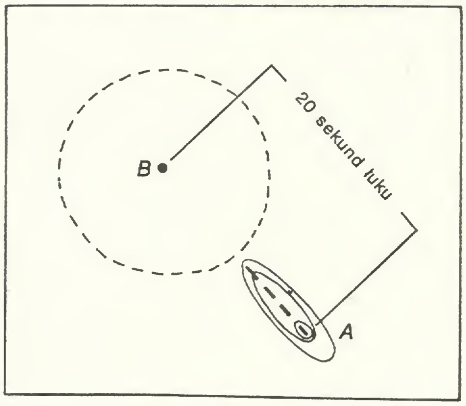
galaktyki typu N

tzw. galaktyki typu N (→ Galaktyki), będące źródłami promieniowania radiowego i wykazujące w widmach linie emisyjne przesunięte ku czerwieni (choć o mniejszą wartość niż kwazary) są zbudowane z małego i jasnego jądra otoczonego bardzo słabą otoczką gwiazd i materii rozrzedzonej. Nie jest wykluczone, że galaktyki typu N oglądane z dużej odległości, z której widoczne byłyby tylko jasne jądro, wyglądałyby podobnie jak kwazary. Właściwości jąder galaktyk typu N są pod wieloma względami pośrednie między właściwościami zwykłych galaktyk i kwazarów.

radiowe obrazy kwazarów

Obrazy radiowe kwazarów są natomiast znacznie bardziej złożone. Obserwacje interferometryczne, pozwalające osiągnąć kątową zdolność rozdzielczą rzędu sekundy łuku, wykazały, że obrazy radiowe mogą zawierać następujące elementy: dwa dobrze rozdzielone obszary emitujące promieniowanie radiowe położone symetrycznie po obu stronach obiektu optycznego; dwa obszary radiowe, z których jeden pokrywa się z obiektem optycznym; kilka obszarów radiowych. Pod tym względem kwazary nie różnią się zasadniczo od radioźródeł innych rodzajów. Ostatnio





Rys. 3. Optyczny (u góry) i radiowy (u dołu) obraz kwazara 3C 273. Zdjęcie jest przeeksponowane w celu uwidocznienia słabej strugi. A i B oznaczają źródła radiowe. Obserwacje interferometryczne wykazały, że radioźródło B ma składową o rozmiarach mniejszych od 10^{-3} sekundy łuku (wg F. D. Kahn and H. Palmer, *Quasars*, Manchester 1967)

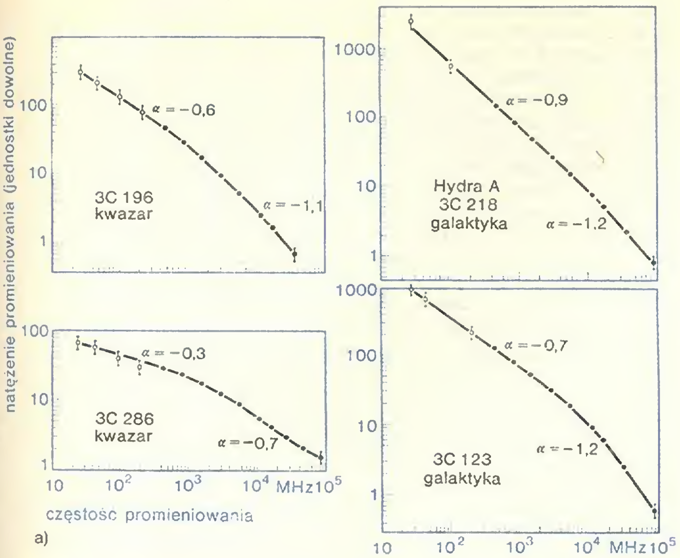
udało się również wykryć istnienie interesującej struktury subtelnej w obszarach emisji radiowej oraz wykryć różnice w widmach radiowych wysyłanych przez elementy radioźródeł zidentyfikowanych z kwazarami.

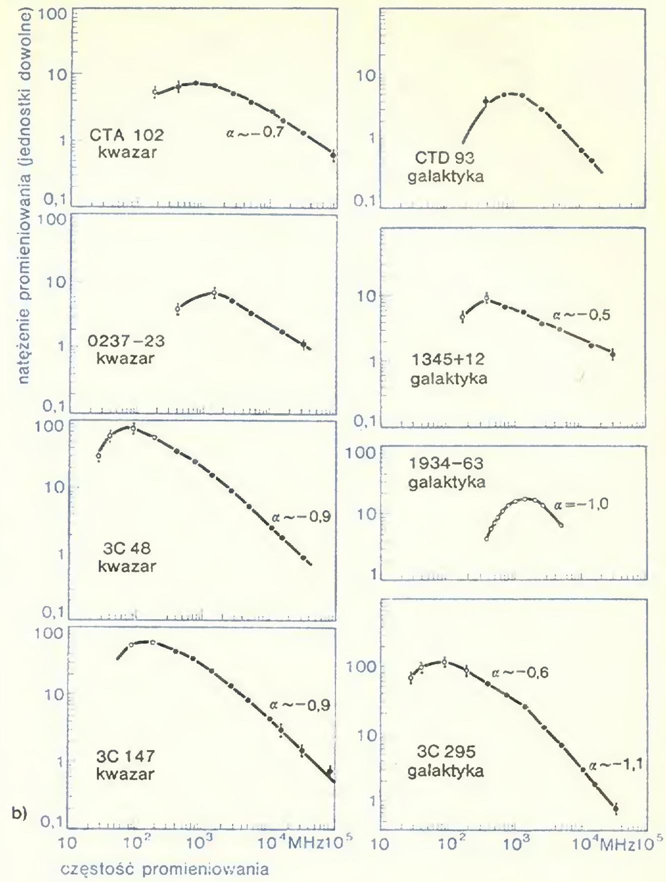
widma kwazarów

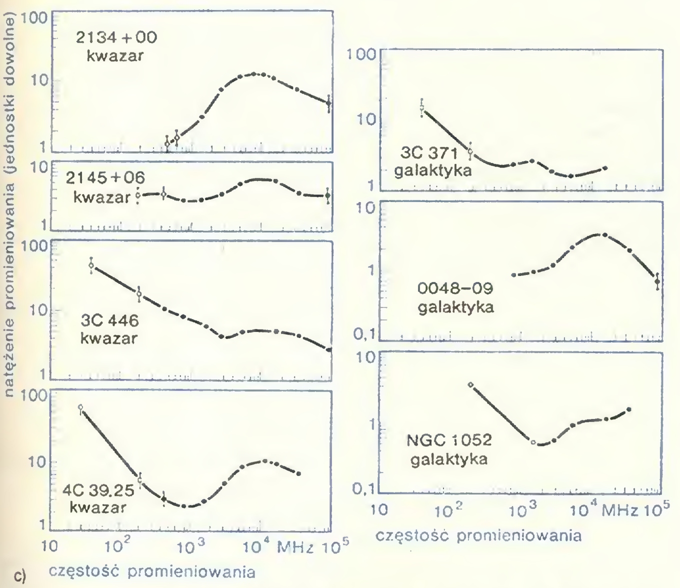
Promieniowanie ciągłe kwazarów we wszystkich zakresach długości fal, w których obserwacje były wykonywane, można opisać w przybliżeniu prawem potęgowym $F(\nu) \sim \nu^{-\alpha}$, gdzie F — strumień promieniowania, ν — częstość. W dziedzinie radiowej α jest równe średnio 0,7 i słabo zależy od częstości. Widma poszczególnych kwazarów mogą jednak różnić się znacznie między sobą. Niektóre kwazary wykazują szybki spadek natężenia w widmie w obszarze fal dłuższych (tzw. obcięcie długofalowe, interpretowane jako wynik samoabsorpcji promieniowania synchrotronowego) lub mają widma złożone, nie dające opisać się prostym prawem potęgowym. Jak widać na rys. 4, widma kwazarów nie różnią się od widm radiowych zwykłych galaktyk. Widma niektórych kwazarów zmieniają się w czasie, podobnie jak to obserwuje się u innych radioźródeł. Odróżnienie kwazara od radiogalaktyki na podstawie tylko obserwowanych cech promieniowania radiowego jest więc niemożliwe.

W dziedzinie widzialnej i w bliskiej podczerwieni wykładnik α jest średnio równy ok. 1, co wyraźnie odróżnia kwazary (oraz jądra niektórych galaktyk zwartych, w tym również galaktyk typu N) od gwiazd i wyraźnie wskazuje na nietermiczny rodzaj promieniowania. Przebieg widma w podczerwieni nie jest jeszcze dokładnie znany, wiadomo tylko, że natężenie widma ciągłego w tym zakresie silnie wzrasta.

Widmo rentgenowskie jest znane tylko dla najjaśniejszego kwazara 3C 273. W przedziale od 1 do 15 keV wykładnik α ma wartość 0,89. Po stronie ma-







Rys. 4. Przykłady radiowych widm kwazarów i galaktyk: a) typowe widma, b) widma z obcięciem długofalowym (samoabsorpcja?), c) widma złożone

łych energii obserwuje się obcięcie widma, będące najprawdopodobniej wynikiem absorpcji promieniowania X w materii rozproszonej znajdującej się na drodze między Ziemią i kwazarem.

linie emisyjne

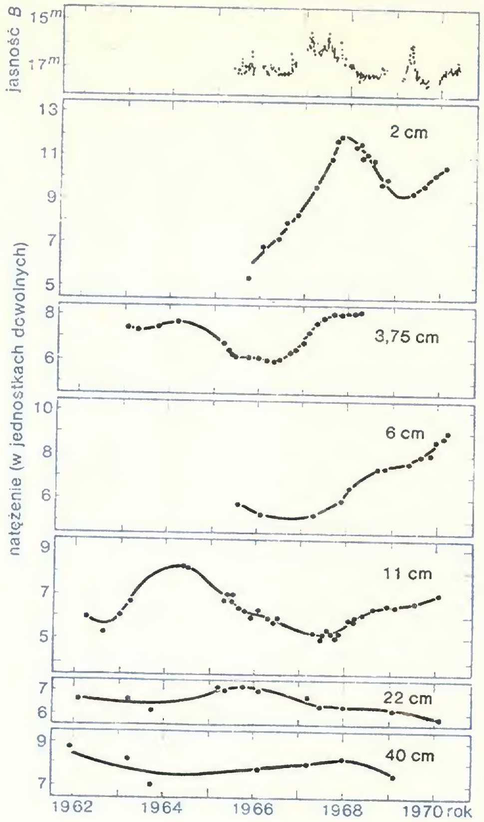
Charakterystyczną cechą kwazarów jest występowanie w ich widmach optycznych silnych linii emisyjnych, których szerokości zawierają się w przedziale od 0,1 nm do ponad 10 nm, ze średnią wartością ok. 5 nm. Wśród linii emisyjnych obserwuje się linie wodoru, jednokrotnie zjonizowanego magnezu (Mg II) oraz wzbronione linie tlenu (O II i O III) i neonu (Ne III i Ne IV). Wszystkie linie emisyjne są przesunięte ku czerwieni o taką samą wartość z.

linie absorpcyjne

Niezwykle interesującą cechą kwazarów jest występowanie niekiedy w ich widmach stosunkowo wąskich i słabych linii absorpcyjnych. Liczba linii absorpcyjnych jest na ogół tym większa, im większe jest przesunięcie ku czerwieni (wyznaczone z linii emisyjnych). Przesunięcie ku czerwieni wyznaczone z linii absorpcyjnych jest zwykle mniejsze od przesunięcia ku czerwieni wyznaczonego z linii emisyjnych. W widmach wielu kwazarów o dużym z linie absorpcyjne mają kilka składowych, które mają różne przesunięcia ku czerwieni. Kwazar 4C 05.34, mający największe znane obecnie przesunięcie ku czerwieni $z = 2{,}877$, ma w widmie ok. 100 linii absorpcyjnych rozpadających się na co najmniej pięć układów o różnych z_{abs}, z których największe jest równe 2,785, a najmniejsze 2,473. Wśród kwazarów mających kilka układów linii znanych jest kilka obiektów, dla których obok układów o $z_{abs} < z_{em}$ występuje również układ o $z_{abs} > z_{em}$. Przypuszcza się, że linie absorpcyjne powstają w warstwach chłodnej i rzadkiej materii wyrzucanych z kwazarów z dużymi prędkościami. Istnienie wielu układów linii absorpcyjnych wskazywałoby na możliwość wielokrotnego wyrzucania materii przez kwazary. Niektóre układy linii absorpcyjnych powstają zapewne w obłokach materii międzygalaktycznej nie związanej bezpośrednio z kwazarami; w ten sposób można wyjaśnić istnienie przypadków $z_{abs} > z_{em}$.

zmienność kwazarów

Charakterystyczną cechą kwazarów jest zmienność ich jasności zarówno w dziedzinie optycznej, jak i radiowej. Określenie najkrótszego czasu t, w którym zmiany te mogą zachodzić, pozwala ocenić górną granicę rozmiarów R obszaru, z którego wysyłane jest promieniowanie. Pomijając pewne szczególne przypadki, możemy oczekiwać, że $R \lesssim ct$, gdzie c jest prędkością światła. (Nawet nagła zmiana jasności rozległego źródła będzie przez nas obserwowana jako zmiana powolna, zachodząca w czasie, w ciągu którego światło zdąży dobiec do nas od wszystkich punktów źródła). Ponieważ w dziedzinie widzialnej t jest rzędu tygodni lub dni, górna granica rozmiarów obszarów wysyłających promieniowanie widzialne jest rzędu dziesiątych, a nawet setnych części parseka. Zmiany jasności radiowej są powolniejsze i następują z okresem charakterystycznym rzędu lat. Obszary wysyłające promieniowanie radiowe mogą mieć rozmiary rzędu parseków. Na rys. 5 przedstawione są



Rys. 5. Optyczne i radiowe zmiany kwazara 3C 345 w latach 1962–1970

wyniki jednoczesnych obserwacji kwazara 3C 345 w dziedzinie widzialnej (w barwie niebieskiej *B*) oraz w różnych długościach fal radiowych. Podobne zmiany jasności wykazują również inne radiogalaktyki, w tym galaktyki typu *N*.

Interpretacja teoretyczna

Najbardziej charakterystyczną cechą kwazarów są ich duże przesunięcia ku czerwieni. W zasadzie istnieją dwie następujące możliwości wyjaśnienia tych przesunięć:

1. Są one przesunięciami dopplerowskimi, tzn. odpowiadają dużym prędkościom v_r oddalania się kwazarów od obserwatora zgodnie z zależnością $z+1 = \sqrt{(v_r+c)/(v_r-c)}$. (Dla małych przesunięć ku czerwieni, czyli dla prędkości radialnych dużo mniejszych od c spełniona jest przybliżona zależność klasyczna $z = v/c$).

2. Są one pochodzenia grawitacyjnego, tzn. że obserwowane przez nas światło zostało wysłane przez ciała o bardzo dużej masie, a opuszczając ich silne pola grawitacyjne doznało dużego przesunięcia ku czerwieni, które zgodnie z ogólną teorią względności jest określone wzorem $z_g+1 = \sqrt{1-2(GM/Rc^2)}$, gdzie G — stała grawitacji, M — masa, R — promień ciała wysyłającego kwanty.

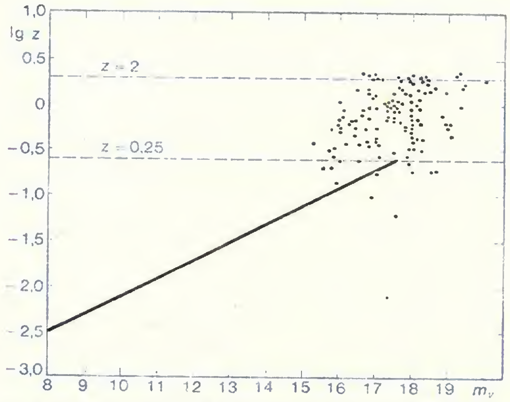
Obecnie istnieje wiele argumentów za tym, że przesunięcia ku czerwieni w widmach kwazarów są przesunięciami dopplerowskimi. Zakładając, że jest tak naprawdę, stajemy wobec dwu możliwości:

1. Duża prędkość ucieczki jest wynikiem ogólnej ekspansji Wszechświata. Zgodnie z prawem Hubble'a, prędkość ucieczki v jest wprost proporcjonalna do odległości, co zapisujemy jako $v = Hr$, gdzie r — odległość, H — stała Hubble'a, równa według ostatnich wyznaczeń ok. 50 km/s·Mps. Innymi słowy, kwazary są obiektami kosmologicznymi, tzn. znajdują się w ogromnych odległościach rzędu tysięcy Mps (hipoteza kosmologiczna).

hipoteza lokalna

2. Duże prędkości oddalania się kwazarów wynikają z wielkich prędkości własnych kwazarów w stosunku do ich najbliższego otoczenia. Duże prędkości mogą oznaczać, że kwazary zostały wyrzucone w wyniku potężnej eksplozji, która nastąpiła w naszym pobliżu (hipoteza lokalna). Ponieważ nie obserwuje się wśród kwazarów przesunięć ku fioletowi (choć kwazary takie byłyby łatwiejsze do wykrycia niż kwazary z przesunięciem ku czerwieni), hipoteza ta w sposób niemożliwy do przyjęcia wyróżnia naszą Galaktykę lub jej najbliższe otoczenie. Hipotezę kosmologiczną wzmocniło jeszcze odkrycie kwazara PKS 225+11 leżącego w kierunku jednej z gromad galaktyk i mającego takie samo przesunięcie ku czerwieni, jakie ma jedna z galaktyk gromady.

Przekonanie, że kwazary są obiektami kosmologicznymi, jest przyczyną szczególnego zainteresowania nimi współczesnej atronomii. Biorąc pod uwagę, że rekordowa wartość przesunięcia ku czerwieni wśród radiogalaktyk jest równa 0,46, łatwo można zrozumieć, że kwazary o przesunięciach ku czerwieni większych od 2 dostarczają nieoczekiwanej możliwości sięgnięcia do odległości ponad dwukrotnie większych z jednoczesnym „cofnięciem się" w czasie o kilka miliardów lat (tyle czasu biegnie bowiem światło od najodleglejszych kwazarów). W związku z tym z badaniami kwazarów wiązano duże nadzieje na rozstrzygnięcie wielu zasadniczych problemów kosmologicznych. Niestety nadzieje te zawiodły po stwierdzeniu, że kwazary wykazują bardzo duży rozrzut jasności absolutnych i to zarówno optycznych, jak i radiowych. Wykres zależności jasności obserwowanych od przesunięcia ku czerwieni, będący odbiciem tych różnic, jest przedstawiony na rys. 6.



Rys. 6. Zależność przesunięcia ku czerwieni od jasności obserwowanej m_v dla 136 kwazarów (punkty). Linia prosta przedstawia tę samą zależność dla najjaśniejszych galaktyk w gromadach

wnioski na temat ewolucji Wszechświata

Mimo to, obserwacje kwazarów przyczyniły się do wysunięcia pewnych interesujących wniosków na temat ewolucji Wszechświata. Istotną rolę odegrała tu zależność $\lg N - \lg S$, czyli wykres przedstawiający liczbę źródeł o jasnościach większych od danej jasności S w funkcji S. W przypadku źródeł rozłożonych w przestrzeni w sposób jednorodny (nawet przy uwzględnieniu ekspansji Wszechświata) nachylenie zależności $\lg N - \lg S$ powinno być bliskie $-1{,}5$. (Wartość $-1{,}5$ odpowiada jednorodnemu rozkładowi źródeł statycznych). Nachylenie takie stwierdzono dla zidentyfikowanych radiogalaktyk, których przesunięcia ku czerwieni są na ogół niewielkie. Taki sam wykres dla kwazarów zidentyfikowanych optycznie ma nachylenie wyraźnie większe, ok. $-1{,}8$. Z drugiej strony, znając przesunięcia ku czerwieni kwazarów oraz przyjmując któryś z modeli kosmologicznych (przy czym wybór konkretnego modelu nie ma za-

sadniczego znaczenia), możemy określić ich odległości, a następnie ocenić przestrzenną gęstość kwazarów w funkcji odległości (przesunięcia ku czerwieni). Z rozważań takich wynika, że istnieje nadmiar kwazarów o dużych przesunięciach ku czerwieni w stosunku do tego, czego powinniśmy oczekiwać dla rozkładu jednorodnego. Wielkość i kierunek tej rozbieżności jest zgodny z zależnością $\lg N - \lg S$. Ich gęstość przestrzenna w odległości odpowiadającej $z = 1$ jest ok. 100 razy większa niż w naszym najbliższym otoczeniu. Wniosek ten przemawia silnie za stałą ewolucją we Wszechświecie.

Niezależnie od znaczenia jakie kwazary mogłyby mieć dla kosmologii, są one obiektami interesującymi również z innych powodów. Jeżeli znajdują się one rzeczywiście w odległościach kosmologicznych, tj. zgodnych z prawem Hubble'a, wówczas ich jasność absolutna musi być ok. 100 lub więcej razy większa od jasności absolutnej najjaśniejszych spośród znanych galaktyk. Biorąc zaś pod uwagę niewielkie rozmiary liniowe obszarów wysyłających te ilości energii, stajemy wobec pasjonującego problemu wyjaśnienia procesów fizycznych, które mogą być jej źródłem. Nie ulega wątpliwości, że promieniowanie kwazarów ma charakter nietermiczny; jest ono najprawdopodobniej promieniowaniem synchrotronowym z domieszką promieniowania powstającego w tzw. odwrotnym efekcie Comptona (▸ Astronomia promieni X i γ). W obu wypadkach źródłem kwantów promieniowania są bardzo szybkie, relatywistyczne elektrony. Pochodzenie szybkich strumieni elektronów u kwazarów nie zostało jeszcze wyjaśnione. Do bardziej interesujących hipotez należy zaliczyć hipotezę anihilacji materii i antymaterii oraz hipotezę wybuchów supernowych, będących następstwem „zlepiania" się gwiazd w obszarach, w których przestrzenna gęstość gwiazd jest rzędu 10^{10} na ps^3. Kwazary nie wniosły więc istotnego wkładu do rozwiązania problemów kosmologicznych, ale okazały się obiektami niezwykle interesującymi z punktu widzenia procesów fizycznych, które mogą w nich przebiegać.

problem źródła promieniowania kwazarów

G. Burbidge and M. Burbidge *Quasi-Stellar Objects*, San Francisco 1976; D. Evans (ed.) *External Galaxies and Quasi-Stellar Objects*, Dordrecht 1972.

Pulsary

Marek Demiański

Latem 1967 r. grupa radioastronomów pod kierunkiem Anthony Hewisha, prowadziła obserwacje dalekich → Kwazarów. Ich celem było zbadanie szybkich zmian sygnałów fal radiowych, emitowanych przez te źródła. Spójne fale radiowe przechodzące przez chmury plazmy ulegają zniekształceniom wywołującym charakterystyczne, szybkie i nieregularne fluktuacje w ich natężeniu. To „migotanie" obserwuje się tylko w wypadkach zwartych źródeł promieniowania, których rozmiary kątowe są mniejsze od jednej sekundy. Zwykłe radiogalaktyki są znacznie większe, natomiast kwazary mają małe rozmiary kątowe. Migotanie było więc pierwszym testem na to, czy dane radioźródło może być kwazarem. Podobnie jak jonosfera, obszary wypełnione plazmą uginają bardziej fale długie niż krótkie. Zatem migotanie jest wyraźniejsze dla fal o metrowej długości niż dla fal centymetrowych. Do tych badań używano radioteleskopu, który pracował na długości fali 3,7 metra (rys. 1).



Rys. 1. Radioteleskop używany do badania kwazarów i pulsarów

Analiza danych obserwacyjnych wskazywała na istnienie bardzo dziwnego obiektu. Pojawiał się on wówczas, kiedy migotanie innych źródeł było małe a natężenie promieniowania bardzo nieregularne. Paralaksa tego źródła była zbyt mała, aby ją zmierzyć, co oznaczało, że znajduje się ono daleko poza granicami Układu Słonecznego. Najwłaściwsze wydawało się przypuszczenie, że jest to wybuchająca gwiazda. Zaczęto więc badać, czy sygnały pochodzące od tego źródła są podobne do sygnałów, jakie wysyła Słońce w okresie wzmożonej aktywności radiowej. Najważniejsze było zatem zbadanie zmienności sygnału. Tymczasem aktywność źródła znacznie zmalała i trzeba było czekać dwa miesiące, aby pojawiło się znowu. 28 IX 1967 r. przeprowadzono najważniejsze obserwacje. Okazało się, że źródło wysyła sygnały w postaci krótkich, powtarzających się co 1,3 s impulsów, jakich dotychczas nie obserwowano.

Na podstawie zmienności sygnału można ocenić rozmiary źródła. Jeżeli okres zmienności sygnału wynosi t, wówczas rozmiary źródła nie mogą być mniejsze od ct, gdzie c — prędkość światła. Źródłem zaobserwowanych impulsów mógł być zatem bardzo mały obiekt, raczej planeta niż gwiazda. Wówczas powstała fantastyczna hipoteza, że sygnały te są wysyłane przez pozaziemską cywilizację. Jednak hipotezę tę odrzucono, gdy okazało się, że sygnały nie zawierają żadnej informacji — nie doszukano się bowiem jakiejkolwiek korelacji między sygnałami ani prawidłowości w ich przebiegu prócz niezwykłej ich periodyczności. W tym czasie odkryto również inne źródła tego typu i w ten sposób przekonano się, że są to obiekty astronomiczne zupełnie nowego rodzaju.

Dane obserwacyjne

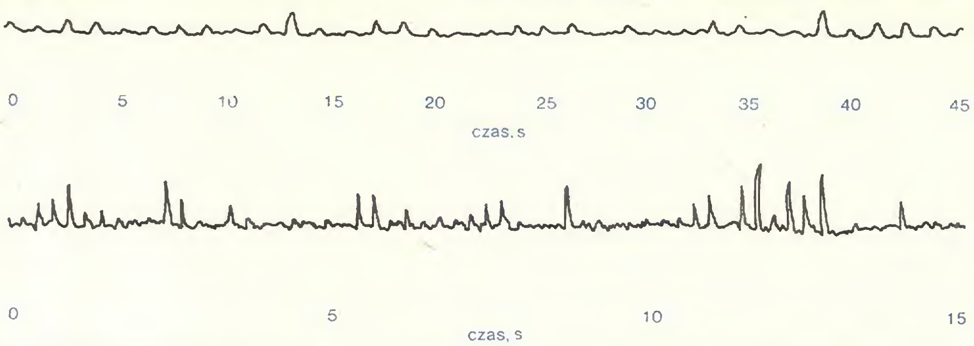
Obecnie znamy ok. trzystu dwudziestu pulsarów. Najbardziej charakterystyczną cechą ich promieniowania jest zmienność impulsów przy zachowaniu niezwykle stabilnej periodyczności. Amplituda oraz kształt impulsu, jak to widać na rys. 2, ulegają zmianom, zarówno w ciągu krótkich odstępów czasu, a więc nawet od impulsu do impulsu, jak też w ciągu dłuższych okresów czasu. Dużą stabilnością cechuje się natomiast średni odstęp pomiędzy impulsami, który nazywa się okresem pulsara. Poszczególne impulsy mają złożoną strukturę i bardzo często składają się

okres pulsara

struktura impulsów

z dwóch a nawet kilku podimpulsów. Ze względu na strukturę impulsu pulsary można podzielić na dwie grupy: takie, u których zmiany impulsu są przypadkowe, i takie, gdzie zmiany te są mniej lub bardziej wyraźnie periodyczne.

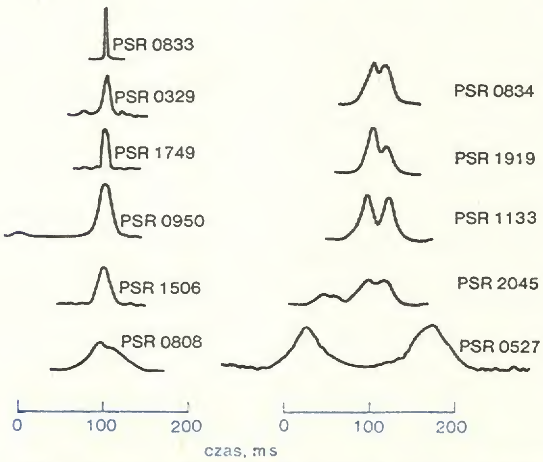
niego impulsu Δt jest w prosty sposób związany z okresem pulsara P. Empirycznie po uśrednieniu otrzymuje się przybliżoną zależność $\Delta t = 0{,}04P$. Dla różnych pulsarów stosunek Δt do P zawiera się w granicach od 0,01 do 0,12. Przy interpretacji tych wyników



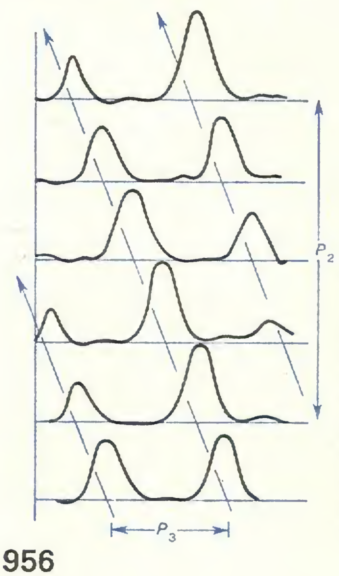
Rys. 2. Pierwsze obserwacje impulsów pulsarów. U góry — fragment sygnału wysyłanego przez pulsara CP 1919. W każdym dwudziestosekundowym odstępie można wyróżnic 15 impulsów. U dołu — zmiany amplitudy pulsara CP 0950

kształty impulsów

Analiza wielu (kilkuset) impulsów prowadzi do wniosku, że każdy pulsar ma charakterystyczny średni kształt impulsu. Niektóre z nich przedstawiono na rys. 3. Podobnie jak pojedynczy impuls — średni impuls może mieć jedno lub kilka maksimów. Z danych obserwacyjnych wynika, że czas trwania średniego impulsu trzeba pamiętać, że Δt zależy od częstości, na jakiej prowadzone są obserwacje.



Rys. 3. Średnie kształty impulsów niektórych pulsarów. Z boku podano numery pulsarów



Rys. 4. Kształt impulsu niektórych pulsarów ulega periodycznym zmianom. Dotyczy to głównie pulsarów o złożonym kształcie impulsu. Prócz podstawowego okresu pulsara P można wyróżnić okresy P_2 i P_3. P_2 jest to najmniejszy odstęp czasu, po którym impuls przyjmuje pierwotny kształt. Odstęp czasu pomiędzy podimpulsami oznaczono przez P_3

F.D. Drake i H.D. Craft zauważyli periodyczne zmiany kształtu impulsu. Śledzili oni dokładnie moment pojawienia się podimpulsu w pulsarze CP 1919 (początkowo pulsary oznaczano podając miejsce odkrycia oraz rektascensję pulsara, np.: CP 1919 to pulsar odkryty w Cambridge, o rektascensji 19 h i 19 min, MP 0034 to pulsar odkryty w Mullard w Australii, o rektascensji 34 min; obecnie powszechnie stosuje się oznaczenie PSR 1919, co oznacza pulsar o rektascencji 19 h i 19 min) i stwierdzili, że okres tych zmian jest bardzo krótki, rzędu 10 ms (rys. 4). Ponadto doszli do wniosku, że szybkie oscylacje są niezależne od głównego sygnału i występują jak gdyby na tle głównego impulsu, co sugeruje, że zachodzą w sposób ciągły. Stąd wynika, że mechanizm odpowiedzialny za powstawanie szybkich zmian jest różny od tego, który wywołuje podstawową okresowość pulsara. Okres szybkich zmian nie jest stały, a zatem co pewien czas sytuacja może się powtarzać. Istotnie periodyczność taka występuje i okres tych zmian waha się od kilku do kilkunastu sekund.

Podstawowy okres pulsara jest bardzo stabilny. Dla większości znanych pulsarów wyznaczono dP/dt, przy czym wielkość ta jest mniejsza od $4{,}2 \cdot 10^{-13}$, a na ogół jest rzędu 10^{-15}. W ciągu ostatnich lat okres pulsarów zmienił się co najwyżej o 10^{-6} swej pierwotnej wartości. Są one zatem bardzo dobrymi zegarami. Prócz powolnych zmian okresu zauważono kilka szybkich zmian, w czasie których okres gwałtownie malał. Nie udało się jednak dokładnie prześledzić tych zmian, ale przypuszcza się, że zachodzą one w ciągu kilku godzin.

Sygnały wysyłane przez pulsary były obserwowane w szerokim zakresie fal od 40 MHz do 8,1 GHz, a pulsary występujące w mgławicach Krab i Żagiel prócz sygnałów radiowych wysyłają również sygnały w obszarze widzialnym, rentgenowskim i w obszarze promieni γ.

cechy impulsów pulsara

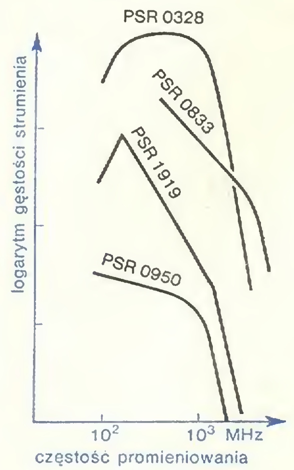
Zauważono trzy główne efekty, które zależą od częstości, mianowicie: dyspersję impulsów, rozmywanie impulsów i zmiany natężenia impulsów. Pierwsze dwa efekty są związane z tym, że sygnał na swej drodze napotyka na materię międzygwiezdną, a trzeci efekt jest zapewne właściwością samego źródła.

Dyspersja impulsów, tzn. zależność prędkości rozprzestrzeniania się sygnałów od częstości, wywołana przejściem impulsów przez obszary plazmy międzygwiezdnej, umożliwia oszacowanie odległości od pulsarów. Znajdują się one daleko poza granicami Układu Słonecznego i większość pulsarów jest w odległości kilkuset parseków. Plazma międzygwiezdna może wywoływać zmianę natężenia impulsów. Trud-

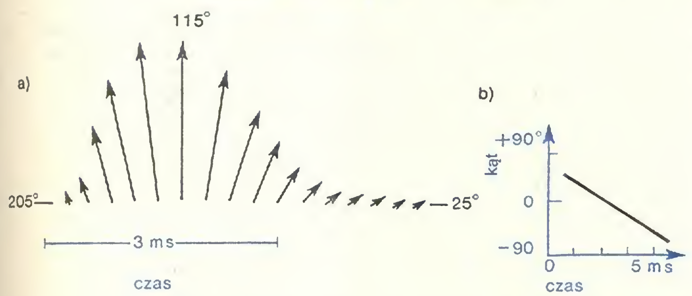
no sobie jednak wyobrazić, aby plazma międzygwiezdna powodowała zmiany amplitudy kolejnych sygnałów. Muszą one być wywoływane przez procesy fizyczne zachodzące w źródle.

widmo energii pulsarów

Widma energii pulsarów są różne, mają jednak wspólną cechę. Wydaje się, że dla wszystkich pulsarów energia wypromieniowywana gwałtownie maleje zarówno dla bardzo dużych, jak też i dla małych częstości (rys. 5).



Rys. 5. Widmo energii niektórych pulsarów. Energia wypromieniowana maleje zarówno dla małych jak i dla dużych częstości



Rys. 6. Zmiany polaryzacji impulsów: a) zmiana polaryzacji impulsu pulsara PSR 0833, b) zmiana kąta polaryzacji pulsara PSR 0833

polaryzacja sygnałów

Wiele danych o mechanizmie promieniowania pulsarów można otrzymać badając polaryzację sygnałów. Stopień polaryzacji jest duży i dochodzi do 95%. Niektóre impulsy są spolaryzowane kołowo. W wielu wypadkach stopień polaryzacji zmienia się w czasie trwania impulsu. Własności polaryzacyjne sygnałów bardzo nieznacznie lub w ogóle nie zależą od częstości. Zmiany polaryzacji impulsów pulsara PSR 0833 przedstawia rys. 6.

pulsar PSR 0532

Najdokładniej zbadany jest pulsar PSR 0532 w Mgławicy Krab (rys. 7), która powstała w wyniku wybuchu supernowej w 1054 r. Pulsar ten wysyła sygnały w zakresie fal radiowych, w widzialnej części widma (rys. 8) oraz w obszarze rentgenowskim. Stwierdzono, że kształty impulsów rentgenowskie i optyczne pokrywają się z dokładnością do jednej milisekundy. Impulsy występują też w obszarze promieni γ.

Jak dotychczas nie stwierdzono impulsów w obszarze γ o energii większej od 50 MeV. Bardzo dokładne obserwacje nie doprowadziły do odkrycia żadnych linii ani pasm w widmie pulsarów. Próby odkrycia krótkotrwałych periodycznych zmian impulsów w obszarze widzialnym też zawiodły.

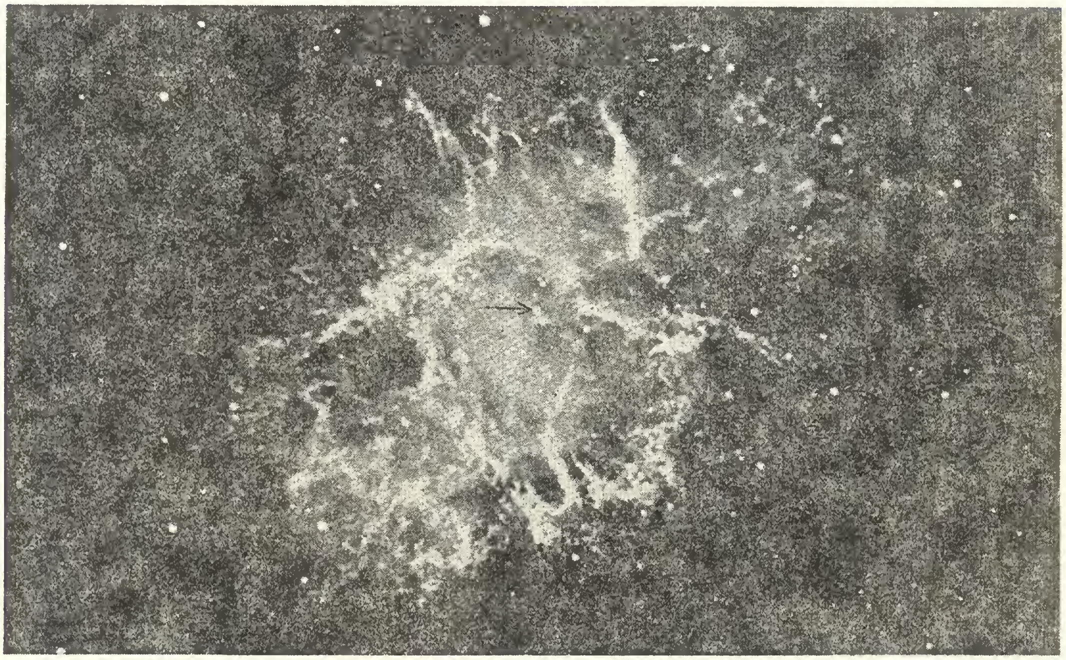
Pulsar w Mgławicy Krab jest jedynym jak do tej pory pulsarem, dla którego stwierdzono sporadyczne (raz na ok. 10^4 impulsów) krótkotrwałe wzrosty energii impulsu o mniej więcej 1000 razy w porównaniu ze średnią energią impulsów. Impulsy są spolaryzowane. W obszarze widzialnym stopień polaryzacji liniowej zmienia się od 20 do 30%, a stopień polaryzacji kołowej jest mniejszy od 10%. Płaszczyzna polaryzacji szybko się obraca, stale w tym samym kierunku. Okres pulsara PSR 0532 zmierzono bardzo dokładnie; jego zmiany można opisać wzorem empirycznym:

$$P = b_0 + b_1 t + b_2 t^2, \tag{1}$$

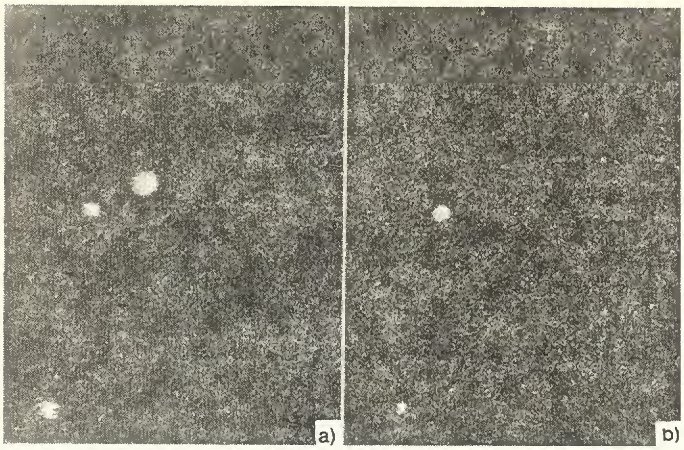
gdzie b_0, b_1 i b_2 są stałymi. Za początek liczenia czasu wybrano północ 17 III 1969 r. (czas t mierzony jest w tygodniach). Dane obserwacyjne prowadzą do

mgławica Krab

$$P = [33095563{,}9268 + (3652256 \pm 0{,}00050)\,t - (0{,}55 \pm 0{,}13)\,10^{-4} t^2] \pm 0{,}037\ \text{ns}. \tag{2}$$



Rys. 7. Mgławica Krab. Strzałką zaznaczono pulsara PSR 0532



Rys. 8. Pulsar w Mgławicy Krab periodycznie „mruga" i w widzialnej części widma. Na zdjęciu wykonanym w momencie maksimum amplitudy jest widoczny (a), natomiast nie pojawia się na zdjęciu wykonanym o pół okresu później (b)

Proszę zwrócić uwagę na niezwykłą dokładność, z jaką znamy okres tego pulsara. Przyjmuje się, że zmiany prędkości kątowej ω pulsarów są opisywane zależnością:

$$\frac{d\omega}{dt} = -K\omega^n, \qquad (3)$$

gdzie K i n — stałe; dla pulsara w Mgławicy Krab $n = 2{,}515 \pm 0{,}005$.

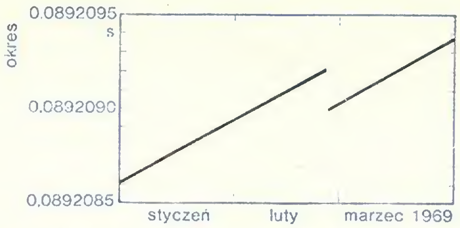
straty energii

Wydłużenie się okresu oznacza, że pulsar PSR 0532 traci część swojej energii rotacyjnej. Przyjmując, że moment bezwładności I nie ulega zmianie, szybkość strat energii można zapisać w postaci:

$$\frac{dE_{\rm rot}}{dt} = 4\pi^2 \frac{I}{P^3} \frac{dP}{dt}. \qquad (4)$$

Oszacowanie strat energii wymaga znajomości momentu bezwładności pulsara. Do tych rozważań wrócimy, gdy będziemy rozpatrywali modele pulsarów.

Dla kilku pulsarów zauważono gwałtowną zmianę okresu P, po czym dP/dt znowu przyjmowało stałą wartość, choć nieco różną od wartości przed skokiem. Zjawisko to obserwowano wielokrotnie. Na przykład zmiany takie wystąpiły u pulsara PSR 0833 (przedstawiono je na rys. 9).



Rys. 9. Gwałtowna zmiana okresu pulsara PSR 0833. Między 24 II a 3 III 1969 r. okres zmniejszył się o 200 ns

Dotychczas tylko dla pulsarów w mgławicach Krab i Żagiel i pulsara rentgenowskiego Her X-1 zarejestrowano impulsy w obszarze widzialnym. Poszukiwania pulsarów w obszarze rentgenowskim też zakończyły się powodzeniem. Znaleziono kilka takich obiektów. Prócz PSR 0532 impulsy w obszarze rentgenowskim wysyłają między innymi Her X-1 i Cen X-3. Rentgenowski pulsar Her X-1 ma okres 1,24 s i jest najszybciej zmiennym pulsarem rentgenowskim. Okres pulsara Cen X-3 wynosi 4,8 s i 90% energii wypromieniowuje w postaci impulsów. Oba te pulsary wchodzą w skład układów podwójnych (→ Ewolucja gwiazd) o okresach zmienności odpowiednio dla Her X-1 1,7 dnia i dla Cen X-3 2,08 dnia.

pulsary rentgenowskie

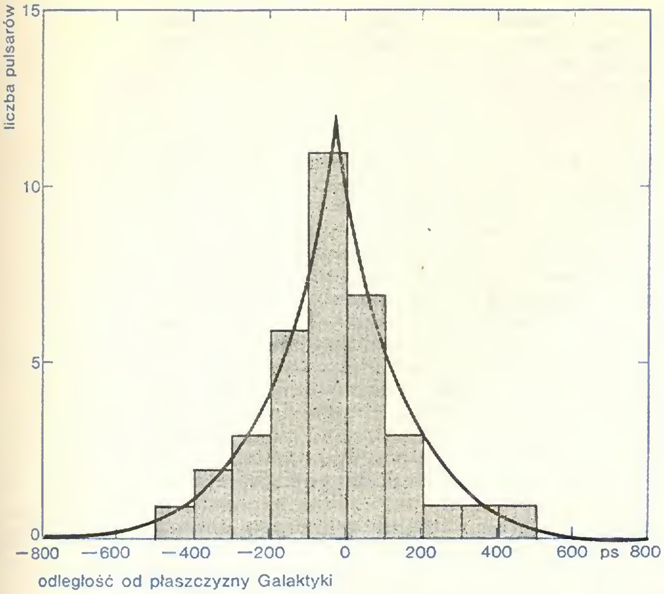
własności pulsarów

Obecnie znamy ok. 320 pulsarów. Można już więc próbować przeprowadzić statystyczną analizę ich własności, niektóre z nich zestawiono w tabeli. Okresy pulsarów radiowych zawierają się w zakresie od 0,033 s do 4,3 s, przy czym najwięcej pulsarów ma okresy od 0,5–1,0 s. Odległości pulsarów od Ziemi wynoszą najczęściej kilkaset parseków i zawierają się w granicach 60–11 000 ps. Interesujące jest zestawienie odległości pulsarów od płaszczyzny Galaktyki (rys. 10), z którego wynika, że grupują się one w jej płaszczyźnie. Przestrzenny rozkład pulsarów w Galaktyce, który przedstawia rys. 11, jest podobny do rozkładu gwiazd pierwszej populacji (→ Galaktyka) i supernowych drugiego rodzaju.

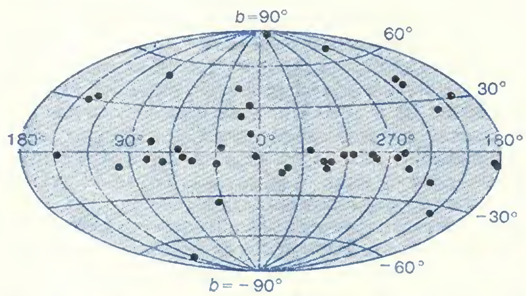
Wiosną 1974 r. grupa radioastronomów korzystając z największego obecnie na świecie radioteleskopu w Arecibo (Porto Rico) odkryła kilkadziesiąt nowych

Własności niektórych pulsarów

Nazwa pulsara	Rektascensja (1950,0)			Deklinacja (1950,0)			Okres P, s	Zmiana okresu $\frac{dP}{dt}\cdot 10^{1\ldots}$	Wiek pulsara T, lata	Miara dyspersyjna cm^{-3} ps
	h	m	s	°	′	″				
MP	00	31	37	−07	37	—	0,940	—	—	12
MP	02	54	24	−54	—	—	0,448	—	—	10
CP	03	29	11,17	54	24	37	0,714 518 625	2,03	$1{,}1\cdot 10^7$	26,75
MP	04	50	22	−18	00	—	0,54978	—	—	25
NP	05	25	45	21	58	—	3,745 491 780	—	—	50,2
NP	05	31	31,46	21	58	54,8	0,033 099 324(28–6–69)	422,66	$2{,}5\cdot 10^3$	56,88
PSR	06	28	53	−28	33	—	1,244 436	—	—	34,4
MP	07	36	51	−40	35	—	0,375	—	—	100
CP	08	08	58,00	74	38	10	1,292 241 315	<0,07	$>6\cdot 10^8$	5,77
MP	08	18	6	−15	—	—	1,237	—	—	25
AP	08	23	52	26	48	00,0	0,530 659 625	—	—	19,4
PSR	08	33	39	−45	00	05	0,089 209 298(24–3–69)	124,26	$23\cdot 10^4$	63
CP	08	34	26,3	06	20	47,0	1,273 763 256	6,81	$5{,}9\cdot 10^6$	12,8
MP	08	35	34	−40	—	—	0,765	—	—	120
HP	09	04	—	77	40	—	1,579 05	—	—	—
MP	09	40	40	−56	—	—	0.662	—	—	145
PP	09	43	19,6	10	05	33	1,097 707	—	—	15,35
CP	09	50	30,85	08	09	49,8	0,253 065 046	0,238	$3{,}4\cdot 10^7$	2,98
MP	09	59	51	−54	37	—	1,436 551	—	—	90
CP	11	33	27,39	16	07	30,4	1,187 911 129	3,75	$1{,}0\cdot 10^7$	4,87
MP	11	54	45	−62	—	—	0,400	—	—	270
AP	12	37	17	25	09	30	1,382 451	—	—	8,5
MP	12	40	21	−63	36	—	0,388	—	—	220
MP	13	59	43	−50	—	—	0,690	—	—	20
MP	14	26	35	−66	30	—	0,788	—	—	60
MP	14	49	22	−65	—	—	0,180	—	—	90
PSR	14	51	29	−68	32	—	0 264	—	—	8,6

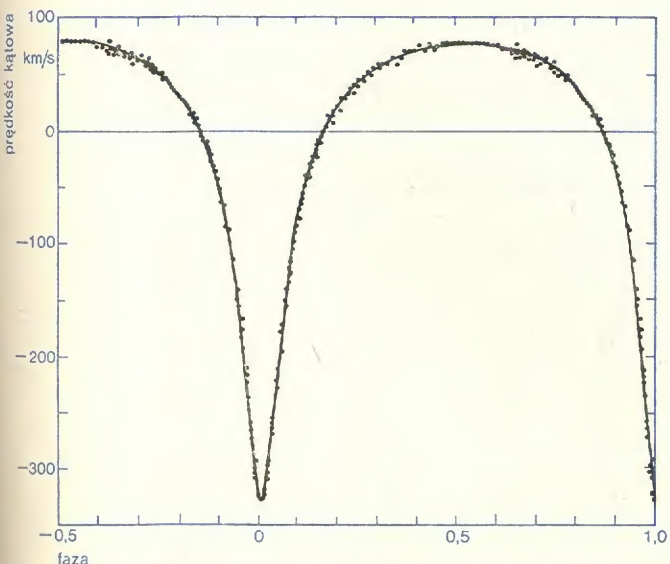


Rys. 10. Rozkład pulsarów w zależności od ich odległości od płaszczyzny Galaktyki



Rys. 11. Przestrzenny rozkład pulsarów. Mają one wyraźną tendencję do skupiania się w płaszczyźnie Galaktyki; *b* szerokość galaktyczna

pulsarów. Okres jednego z nich (pulsar PSR 1923, odkryty przez J. H. Taylora i R. A. Hulse'a) zmieniał się periodycznie co 0,3230 dnia pomiędzy 0,058967 a 0,059045 s (rys. 12). Z danych tych wynika, że pulsar



Rys. 12. Periodyczna zmiana prędkości kątowej pulsara PSR 1923 oznacza, że wchodzi on w skład układu podwójnego

pulsar radiowy w układzie podwójnym

ten wchodzi w skład układu podwójnego. Jest to pierwszy pulsar radiowy występujący w układzie podwójnym. Jego odkrycie ma duże znaczenie. Być może korzystając z dokładniejszych danych obserwacyjnych na temat tego układu będzie można wyznaczyć masę gwiazdy neutronowej. Byłaby to pierwsza bezpośrednia obserwacyjna informacja o podstawowych parametrach charakteryzujących gwiazdy neutronowe. Układ ten można również stosować do sprawdzenia różnych efektów szczególnej i ogólnej teorii względności.

Dane obserwacyjne pozwalają na podanie wielu ograniczeń na modele pulsarów. Krótki czas trwania impulsów oznacza, że źródło promieniowania jest małych rozmiarów. Niezwykła stabilność okresu pulsarów wskazuje na to, że proces odpowiedzialny za periodyczność powinien być związany z mechanicznymi własnościami układu.

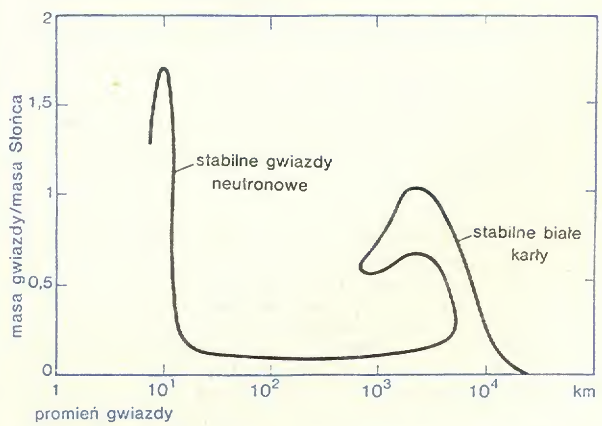
Początkowo przypuszczano, że pulsary są pulsującymi białymi karłami, stąd zresztą myląca nieco ich nazwa. Jednak z dokładnej analizy wynika, że białe karły nie mogą pulsować z tak krótkim okresem. Obrót białego karła, jako mechanizm zapewniający periodyczność, należało też odrzucić, gdyż wymagana prędkość kątowa obrotu dla krótkookresowych pulsarów byłaby za duża i siły grawitacyjne nie mogłyby równoważyć sił odśrodkowych.

Gwiazdy neutronowe

Bardzo szybko po odkryciu pulsarów zdano sobie sprawę z tego, że jedynie gwiazda neutronowa może być podstawą wszelkich modeli mechanizmu zapewniającego periodyczność pulsarów. Jedynego konkurencyjnego kandydata — białego karła — trzeba było odrzucić. Białe karły nie mogą ani pulsować, ani obracać się z tak krótkimi okresami. Obecnie nie ulega już wątpliwości, że pulsar to obracająca się gwiazda neutronowa. Nie można wprawdzie podać bezpośredniego dowodu, ale inne modele musiałyby być znacznie bardziej skomplikowane.

hipoteza gwiazd neutronowych

Hipotezę o istnieniu gwiazd neutronowych zaproponował w latach trzydziestych L. L. Landau, a niezależnie od niego W. Baade i F. Zwicky. Baade i Zwicky przeprowadzali systematyczne obserwacje wybuchów supernowych. Obliczając energię, która zostaje wydzielona w czasie wybuchu, doszli do wniosku, że stanowi ona tylko pewien procent całkowitej energii i po wybuchu powinna jeszcze pozostawać centralna, bardzo gęsta część tworząca gwiazdę neutronową. Zainteresowanie gwiazdami neutronowymi wzrosło, gdy zaczęto badać ostatnie etapy ewolucji gwiazd. Okazało się wówczas, że zależnie od masy mogą istnieć dwie różne stabilne konfiguracje (rys. 13).



Rys. 13. Zależność masy gwiazdy od promienia dla gwiazd o dużej gęstości centralnej. Występują dwie stabilne konfiguracje; białe karły i gwiazdy neutronowe

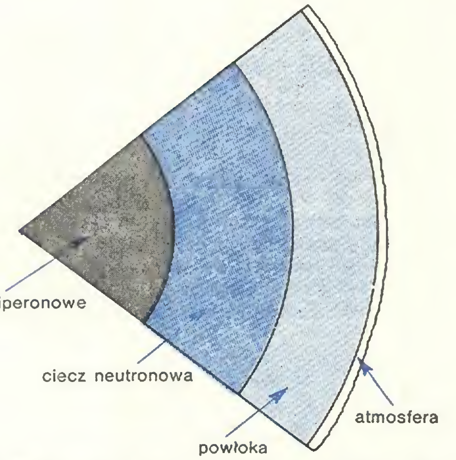
Gwiazda, która wyczerpała już swój zapas paliwa jądrowego i kurcząc się powoli ostyga, stanie się w końcu białym karłem, jeżeli jej masa nie przekracza krytycznej wartości $1{,}4 M_{\odot}$. Gwiazdy bardziej masywne przechodzą przez stadium wybuchu supernowej i z ich centralnych części w wyniku dalszego kurczenia może powstać gwiazda neutronowa. W białych karłach siła grawitacyjnego przyciągania jest równoważona przez ciśnienie gazu elektronowego. W gwiazdach neutronowych siły grawitacyjne są równoważone przez ciśnienie gazu neutronowego. Należy podkreślić, że ciśnienie zarówno w gazie elektronowym jak i neutronowym jest wywołane nie przez zderzenia, jak w przypadku zwykłego gazu, lecz przez efekty kwantowe. Elektrony i neutrony są fermionami i dla nich obowiązuje zakaz Pauliego. Dwa elektrony lub dwa neutrony nie mogą znajdować się w takim samym stanie kwantowym.

ciśnienie gazu neutronowego

Gdy gęstość materii przekracza $5 \cdot 10^{11}$ g/cm³ następuje gwałtowna zmiana składu chemicznego materii; jądra rozpadają się i jako składniki materii pozostają protony, neutrony i elektrony. Przy mniejszych gęstościach charakterystyczny czas tego procesu zależy od temperatury. Temperatura w obszarach centralnych gwiazdy neutronowej, powstałej w procesie wybuchu supernowej, jest bardzo wysoka, często wyższa od 10^{10} K. Reakcje jądrowe sprowadzające układ do stanu równowagi przebiegają wówczas bardzo szybko. Gdy gwiazda neutronowa ostygnie, jej skład chemiczny nie ulega już zmianie, reakcje jądrowe przebiegają bowiem tak wolno, że można je pominąć. Skład chemiczny materii w gwieździe neutronowej w różnych jej obszarach powinien zatem odpowiadać stanowi równowagowemu.

Dzięki znacznym różnicom gęstości pomiędzy centrum 10^{15} g/cm³ a powierzchnią, gwiazda neutronowa powinna być podzielona na obszary wyraźnie rozgraniczone i różniące się własnościami fizycznymi. Można wyróżnić trzy takie obszary (rys. 14).

obszary w gwieździe neutronowej



Rys. 14. Przekrój przez gwiazdę neutronową. Centralne jądro hipernowe jest otoczone cieczą neutronową. Dalej na zewnątrz znajduje się krystaliczna powłoka, a nad nią bardzo cienka warstwa gęstej atmosfery

a) Powłoka. Jej grubość zależy od całkowitej masy gwiazdy i zmienia się od ok. 100 m dla gwiazd o dużej masie do ok. 10 km dla lekkich gwiazd. W powłoce gęstość stopniowo wzrasta i na wewnętrznej granicy osiąga wartość $4{,}3 \cdot 10^{11}$ g/cm³. Poza tym powłoka jest złożona z elektronów i jąder tworzących strukturę krystaliczną. W głębszych warstwach, gdzie gęstość jest większa od 10^{6} g/cm³, najbardziej stabilne są jądra żelaza, na dolnej zaś granicy jądra cyrkonu $Z = 40$, $A = 127$. W normalnych warunkach jądra te były niestabilne ze względu na rozpad β.

b) Ciecz neutronowa. W głębszych warstwach w miarę zbliżania się do centrum, wzrasta gęstość materii. Z energetycznego punktu widzenia korzystna jest w tych warunkach reakcja odwrotna do rozpadu β swobodnego neutronu:

$$p + e \to n + \nu,$$

prowadząca do jąder o coraz to większej zawartości neutronów. W końcu przy gęstości 10^{12} g/cm³ jądra rozpadają się i przy większych gęstościach występuje już tylko mieszanina neutronów, protonów i elektronów. W stanie równowagi stosunek ilości neutronów do protonów i do elektronów wynosi 8 : 1 : 1. Warunki fizyczne panujące w tym obszarze, a więc zakres gęstości i temperatur, jest taki, że neutrony stanowią ciecz neutronową w stanie nadciekłym.

c) Hiperonowe jądro. Gdy masa gwiazdy neutronowej przewyższa $0{,}5 M_{\odot}$, to gęstości w centralnych częściach mogą przewyższać 10^{15} g/cm³. Wtedy prócz neutronów, protonów i elektronów mogą występować w równowadze hiperony. W takich warunkach fizycznych konfiguracją o najniższej energii jest układ krystaliczny. Należy zauważyć, że rozważania dotyczące własności tego obszaru są dość spekulatywne. Nie mamy bowiem pewności, czy równanie stanu, za pomocą którego opisujemy własności materii przy tak dużych gęstościach, znamy dostatecznie dokładnie.

własności gwiazd neutronowych

Zależnie od postaci równania stanu, a więc od tego, jakie procesy fizyczne uwzględnia się przy bardzo dużych gęstościach, otrzymuje się różne wartości parametrów charakteryzujących gwiazdę neutronową. Na ogół przyjmuje się, że masy gwiazd neutronowych wynoszą $0{,}1\text{–}2 M_{\odot}$, a ich promienie mogą być zawarte pomiędzy 10 a 1000 km. Pole grawitacyjne na powierzchni gwiazdy neutronowej jest bardzo silne i konstruując modele takich gwiazd stosuje się równania ogólnej teorii względności. Poprawki relatywistyczne mogą stanowić aż kilkanaście procent. Podstawowe parametry gwiazd neutronowych zestawiono w tabeli (masa gwiazdy neutronowej nie jest sumą mas wszystkich cząstek. Sumę mas wszystkich cząstek wchodzących w skład gwiazdy neutronowej nazywamy jej masą właściwą).

Gwiazdy neutronowe mogą wykonywać radialne drgania i początkowo przypuszczano, że właśnie one będą odpowiednim mechanizmem zapewniającym periodyczność pulsarów. Przypuszczenie to jednak odrzucono. Okazało się bowiem, że typowy okres pulsacji gwiazd neutronowych jest bardzo krótki i wynosi kilka milisekund.

Maksymalna wartość prędkości kątowej, z jaką może się obracać gwiazda neutronowa $\omega_{kr} = \sqrt{G\varrho} \sim 3000$ rad/s, gdzie ϱ — gęstość materii jądrowej równej $2 \cdot 10^{14}$ g/cm³. Wartość ta jest znacznie większa od prędkości kątowej najszybszych pulsarów; na przykład dla PSR 0532 wynosi ona 190 rad/s.

Podstawowe parametry modeli gwiazd neutronowych

lg gęstości centralnej	Masa 10^{30} kg	Masa właściwa 10^{30} kg	Promień km	Energia wiązania 10^{30} kg
15,6	3,45	4,69	8,74	1,24
15,5	3,44	4,56	9,13	1,12
15,4	3,35	4,31	9,54	0,96
15,3	3,17	3,95	9,92	0,78
15,2	2,88	3,46	10,2	0,58
15,1	2,48	2,87	10,5	0,40
15,0	2,01	2,25	10,6	0,25
14,9	1,53	1,68	10,7	0,14
14,8	1,09	1,16	10,7	0,07
14,7	0,73	0,77	11,0	0,03
14,6	0,52	0,54	11,6	0,02
14,5	0,37	0,38	12,5	0,01
14,4	0,25	0,25	15,1	0,004
14,3	0,14	0,14	47,9	0,001
14,2	0,37	0,45	341	0,03
14,1	0,95	1,03	401	0,03
14,0	1,10	1,17	349	0,02

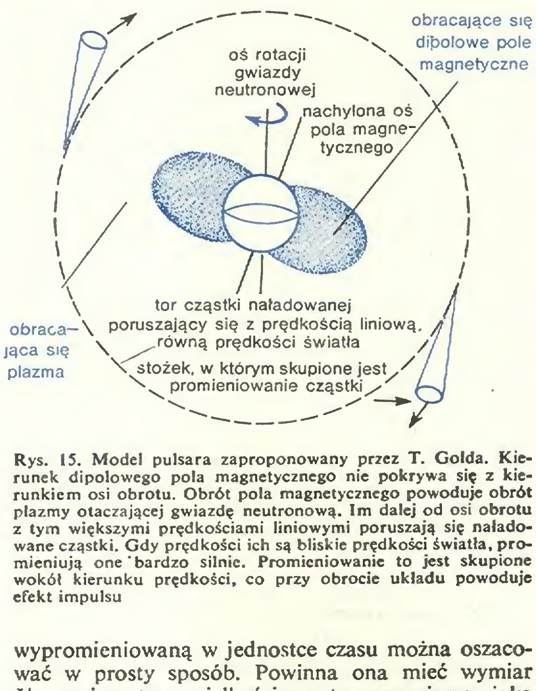
Gwiazdy neutronowe mogą mieć silne pola magnetyczne. Oszacowanie natężeń tych pól można przeprowadzić stosując zasadę zachowania strumienia. Pole magnetyczne B na powierzchni typowej gwiazdy wynosi kilka mikrotesli, zatem korzystając z zasady zachowania strumienia $BR^2 = \text{const}$ można przypuszczać, że pole magnetyczne na powierzchni gwiazdy neutronowej będzie bardzo silne, rzędu 10^8 T.

Model mechanizmu zegarowego

Zmierzono dokładnie prędkość wydłużenia się okresu wielu pulsarów. Stosując wzór (4) można oszacować straty energii rotacyjnej w czasie najdokładniej zbadanego pulsara w Mgławicy Krab. Moment bezwładności I gwiazd neutronowych o masach zbliżonych do masy Słońca szacuje się na 10^{44} g/cm². Podstawiając dane dla pulsara PSR 0532 otrzymujemy $dE_{rot}/dt = 5\cdot10^{30}$ J/s $= 2\cdot10^4\ L_\odot$, gdzie $L_\odot$ jest jasnością Słońca. Stąd wynika, że energia kinetyczna ruchu obrotowego może być źródłem energii pulsarów. Problem polega teraz na tym, aby energię mechaniczną rotacji zamienić na energię promieniowania pola elektromagnetycznego.

Golda model pulsara

W jednej z pierwszych prac teoretycznych, jakie się pojawiły na temat pulsarów, T. Gold zaproponował prosty model takiego mechanizmu. Według Golda pulsar to obracająca się gwiazda neutronowa mająca dipolowe pole magnetyczne. Oś obrotu gwiazdy tworzy pewien kąt z kierunkiem dipola magnetycznego, dzięki czemu pole magnetyczne zmienia się periodycznie w czasie (rys. 15). Energię pola magnetycznego



Rys. 15. Model pulsara zaproponowany przez T. Golda. Kierunek dipolowego pola magnetycznego nie pokrywa się z kierunkiem osi obrotu. Obrót pola magnetycznego powoduje obrót plazmy otaczającej gwiazdę neutronową. Im dalej od osi obrotu z tym większymi prędkościami liniowymi poruszają się naładowane cząstki. Gdy prędkości ich są bliskie prędkości światła, promieniują one bardzo silnie. Promieniowanie to jest skupione wokół kierunku prędkości, co przy obrocie układu powoduje efekt impulsu

wypromieniowaną w jednostce czasu można oszacować w prosty sposób. Powinna ona mieć wymiar J/s; najprostszą wielkością o tym wymiarze, jaką można zbudować z momentu magnetycznego d, prędkości kątowej ω i prędkości światła w próżni c jest $d^2\omega^4/c^3$. Dokładne obliczenia prowadzą do wzoru:

$$\frac{dE}{dt} = -\frac{2}{3}\frac{d^2\omega^4}{c^3}. \tag{5}$$

Podstawiając tutaj dane dla typowej gwiazdy neutronowej, a więc średnie natężenie pola magnetycznego $B = 10^8$ T, promień gwiazdy $R = 10^6$ cm i $\omega = 200\,\text{s}^{-1}$ oraz pamiętając o tym, że możemy przyjąć $d = BR^3$, otrzymamy:

$$dE/dt = 2\cdot10^{31}\ \text{J/s} = 10^5 L_\odot. \tag{6}$$

Wartość ta jest zgodna z obserwowanym strumieniem energii wysyłanym przez pulsar z Mgławicy Krab, która promieniuje $4\cdot10^{30}$ J/s w obszarze rentgenowskim i $2\cdot10^{29}$ J/s w widzialnej części widma. Wartość ta zgadza się też dobrze z oszacowanymi stratami energii rotacyjnej. Różnicy o czynnik 5 nie należy się dziwić, dokonywaliśmy bowiem jedynie grubszych oszacowań.

Model obracającej się, namagnesowanej gwiazdy neutronowej zapewnia periodyczność oraz wyjaśnia pochodzenie energii promieniowania pulsarów. Istnieje duże prawdopodobieństwo, że przestrzeń wokół obracającej się namagnesowanej gwiazdy neutronowej jest wypełniona plazmą. W pobliżu gwiazdy pole magnetyczne będzie dostatecznie silne i plazma będzie się obracała wraz z gwiazdą. Jednak w pewnej odległości od powierzchni gwiazdy wpływ sił zewnętrznych zacznie dominować i plazma nie będzie już obracać się z gwiazdą. Dla szybko obracającej się gwiazdy neutronowej z silnym polem magnetycznym obszar, w którym plazma obracać się będzie wraz z gwiazdą, rozciągać się może niemal do promienia, na którym liniowa prędkość cząstek osiąga prędkość światła. Promieniowanie takiej relatywistycznej plazmy ma być według Golda źródłem sygnałów. Model ten nazwano modelem latarni morskiej. Wyjaśnia on w ogólnych zarysach ideę mechanizmu promieniowania, nie prowadząc do żadnych szczegółowych przewidywań. Teraz należy jeszcze udowodnić, że namagnesowana, szybko rotująca gwiazda neutronowa może być otoczona plazmą.

model latarni morskiej

Jeżeli pominąć rotację gwiazdy neutronowej i jej pole magnetyczne, to gęstość atmosfery gwiazdy opisuje znany wzór barometryczny:

$$\varrho(z) = \varrho_0 e^{-mgz/kT}. \tag{7}$$

Wysokość atmosfery H można określić jako wysokość, na której gęstość jest e-krotnie mniejsza, niż przy powierzchni (wykładnik potęgowy równy 1). Mamy zatem:

$$H = kT/mg, \tag{8}$$

gdzie k — stała Boltzmanna, m — masa cząsteczek atmosfery, a g — przyspieszenie grawitacyjne. Podstawiając dane dla gwiazdy neutronowej $T = 10^6$ K, $m = 3\cdot10^{-24}$ g oraz $g = GM/R^2 = 3\cdot10^{13}$ cm/s² do wzoru (8) otrzymujemy $H = 1$ cm. Żadna gwiazda neutronowa nie powinna mieć atmosfery, sądzono więc, że jest po prostu otoczona próżnią. P. Goldreich i W. Julian wykazali, że tak nie jest. Rozpatrzyli oni prosty model obracającej się gwiazdy neutronowej z polem magnetycznym o dipolu ustawionym równolegle do osi obrotu. Założyli przy tym, że zarówno materia wewnątrz gwiazdy jak i atmosfera są doskonałymi przewodnikami. Analizując własności pola elektromagnetycznego w otoczeniu gwiazdy neutronowej, a szczególnie warunki, jakie muszą być spełnione na granicy między atmosferą a pustą przestrzenią na zewnątrz, zauważyli, że gwiazda neutronowa nie może być otoczona przez próżnię. Pomimo, że w pobliżu powierzchni istnieje silne pole grawitacyjne, gwiazda neutronowa musi być otoczona przez obłoki zjonizowanego gazu, tworząc tzw. magnetosferę. Gaz tworzący magnetosferę jest zjonizowany. Zatem struktura magnetosfery zależy od własności pola grawitacyjnego i pola magnetycznego wytwarzanego przez gwiazdę neutronową.

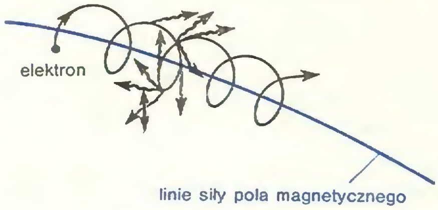
magnetosfera gwiazdy neutronowej

Jak powstają impulsy?

Na to pytanie do dziś nie ma zadowalającej odpowiedzi. Istnieje wiele konkurencyjnych teorii, ale każda z nich ma jeszcze słabe strony. Prawdopodobnie mechanizm powstawania impulsów jest złożony i wiele procesów fizycznych jest odpowiedzialnych za ich powstawanie. Zaproponowane modele, wyjaśniające powstawanie impulsów, wykorzystują następujące

procesy fizyczne: a) promieniowanie relatywistycznych naładowanych cząstek, b) efekty kwantowe występujące w silnych polach magnetycznych, c) laserowe promieniowanie plazmy. Można je zatem podzielić na trzy grupy. Promieniowanie relatywistycznych naładowanych cząstek jako źródło sygnałów wykorzystał T. Gold w swoim modelu latarni morskiej. Przyjął on, że pole magnetyczne obracające się wraz z gwiazdą neutronową unosi ze sobą naładowane cząstki. W obszarze, gdzie prędkości liniowe cząstek są porównywalne z prędkością światła, a więc w odległości R od pulsara, takiej że $c = \omega R$ (na powierzchni tak zwanego cylindra świetlnego), wysyłają one silne promieniowanie elektromagnetyczne skoncentrowane wzdłuż kierunku ruchu cząstki. Promieniowanie to charakteryzuje duży stopień polaryzacji (rys. 16). Trudności w dopasowaniu tego modelu do faktów obserwacyjnych są dwojakiego rodzaju. Po pierwsze, nie wiadomo, w jaki sposób byłaby zapewniona dostatecznie duża koncentracja cząstek w pobliżu cylindra świetlnego, a tym samym uzyskane odpowiednie natężenie promieniowania. Po drugie, model ten nie tłumaczy złożonej struktury sygnałów pulsarów.

promieniowanie relatywistycznych cząstek naładowanych



Rys. 16. Mechanizm promieniowania relatywistycznych naładowanych cząstek. Elektron poruszający się z bardzo dużą prędkością w polu magnetycznym zakreśla linię śrubową wzdłuż linii sił pola magnetycznego. Emitowane fale radiowe, przedstawione liniami falistymi, są spolaryzowane w płaszczyźnie prostopadłe do linii sił pola magnetycznego

efekty kwantowe

Analizując zachowanie się plazmy w silnych polach magnetycznych H. Chiu i V. Canuto zauważyli, że promień linii śrubowej, po której zwykle porusza się naładowana cząstka w polu magnetycznym, nie może być dowolny, lecz jest skwantowany, to znaczy może przyjmować tylko określone wartości. Uważają oni, że promieniowanie pulsarów powstaje w pobliżu powierzchni gwiazdy neutronowej w okolicach jej biegunów magnetycznych. W tych obszarach mogłyby powstawać silne impulsy promieniowania przy jednoczesnym przeskoku elektronów lub protonów z jednej dopuszczalnej linii śrubowej na inną o mniejszym promieniu. Ze względu na trudności z pogodzeniem przewidywań tego modelu z danymi obserwacyjnymi był on wielokrotnie ulepszony, ale nadal nie wyjaśnia wielu ważnych obserwowanych własności pulsarów. Zgodnie z przewidywaniami tego modelu sygnały powstają w pobliżu powierzchni gwiazdy, ale gęsta plazma otaczająca gwiazdę neutronową może silnie zaburzyć sygnały.

promieniowanie laserowe

Aby zrozumieć nadzieje, jakie wiązano z możliwością laserowego promieniowania plazmy, przypomnijmy równanie opisujące transport promieniowania elektromagnetycznego w ośrodku materialnym. Jeżeli przez I oznaczamy natężenie promieniowania, wówczas:

$$dI/dx = A+(B-\mu)I, \qquad (9)$$

gdzie x — odległość mierzona wzdłuż kierunku rozchodzenia się fali, A i B — współczynniki spontanicznej i wymuszonej emisji, μ — współczynnik pochłaniania. Gdy przyjmujemy, że wszystkie te wielkości są stałe, to rozwiązanie tego równania można zapisać w postaci:

$$I = \frac{A}{\mu-B}[1-e^{(B-\mu)x}]+I_0e^{(B-\mu)x}, \qquad (10)$$

gdzie $I = I(x = 0)$. Zauważmy, że jeżeli $B \leqslant 0$, co odpowiada absorpcji, to natężenie będzie asymptotycznie malało do wartości $A/(\mu-B)$. Jeżeli natomiast $B > \mu$, występuje efekt laserowy, i natężenie wzrasta wykładniczo. Gdyby zatem istniały takie warunki fizyczne, które zapewniałyby, że $B > \mu$, wówczas można by wyjaśnić wiele własności promieniowania pulsarów. Takie warunki można zapewnić jednak co najwyżej w pobliżu powierzchni gwiazdy i wobec tego napotykamy tutaj takie same trudności jak poprzednio.

Ostatnio dużą popularnością cieszy się model zaproponowany przez M. Rudermana i P. Sutherlanda. Zbadali oni własności magnetosfer pulsarów i zauważyli, że w pobliżu biegunów magnetycznych wzdłuż linii sił pola magnetycznego powstaje różnica potencjałów uwożliwiająca kreację par elektron–pozyton. Elektrony i pozytony są przyspieszane do relatywistycznych prędkości, a następnie ulegają grupowaniu dzięki niestabilnościom plazmy. Takie grupy ładunków, poruszające się po zakrzywionych liniach sił pola magnetycznego, emitują fale radiowe.

model Rudermana i Sutherlanda

Jak dotychczas nikomu nie udało się podać pełnego i zadowalającego modelu promieniowania pulsarów i zapewne nie nastąpi to szybko. Nie są też jeszcze dokładnie zbadane właściwości magnetosfery otaczającej gwiazdę neutronową. Poznanie ich jest pierwszym krokiem na długiej drodze do zrozumienia mechanizmu promieniowania pulsarów.

mechanizm zmian okresu

Omawiając dane obserwacyjne wspomnieliśmy o dziwnym zjawisku zaobserwowanym u kilku pulsarów, a mianowicie — o raptownej zmianie okresu. U niektórych pulsarów nastąpiło to już kilkakrotnie, przy czym okres pulsara zawsze uległ skróceniu. Stosując model obracającej się gwiazdy neutronowej F. Dyson podał bardzo przekonujące wyjaśnienie. Przypomnijmy, że w gwieździe neutronowej można wyróżnić trzy obszary — powłokę, ciekły obszar pośredni i krystaliczne jądro. Choć ciecz wypełniająca obszar pośredni znajduje się w stanie nadciekłym, to jednak warstwa cieczy granicząca z powłoką obraca się z taką samą prędkością kątową co i powłoka. Kiedy w wyniku straty energii rotacyjnej gwiazda zaczyna się obracać wolniej, maleje też wartość siły odśrodkowej i powierzchnia swobodna cieczy przyjmuje nowy, bardziej sferyczny kształt. W płaszczyźnie równikowej pomiędzy powłoką a powierzchnią swobodną cieczy wytwarza się próżnia. Narastają również naprężenia w powłoce. W końcu struktura krystaliczna powłoki nie może zrównoważyć naprężeń, pęka i przyjmuje nowy kształt, pokrywający się z powierzchnią swobodną cieczy. W czasie tego procesu maleje moment bezwładności. Zgodnie jednak z zasadą zachowania momentu pędu iloczyn momentu bezwładności i prędkości kątowej powinien być stały, a więc prędkość kątowa wzrasta, co odpowiadałoby gwałtownemu skróceniu okresu impulsu. Korzystając z tego modelu można przewidzieć, kiedy nastąpi kolejny taki wstrząs. Przewidywania są w zasadzie zgodne z obserwacjami.

Pulsary rentgenowskie

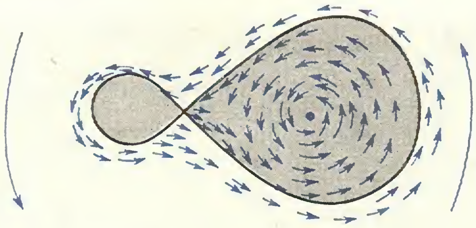
Bardzo interesującą grupą pulsarów są pulsary rentgenowskie. Satelita UHURU, który okrążał Ziemię (→ Astronomia promieni X i γ), dostarczył wiele informacji o źródłach promieniowania rentgenowskiego. Niektóre z nich okazały się źródłami pulsującymi. Ich natężenie zmienia się o czynnik dwa lub więcej z okresem krótszym od kilku sekund. Widmo energii jest płaskie i występują obcięcia od strony niskich energii. Kilka z tych źródeł udało się zindetyfikować z układami gwiazd podwójnych.

Najdokładniej zbadanymi pulsującymi źródłami rentgenowskimi są Cyg X-2, Cen X-3 i Her X-1. Energia wypromieniowywana w jednostce czasu przez te źródła jest zawarta pomiędzy 10^{29}–10^{31} J/s, czyli pomiędzy 10^2–$10^5 L_\odot$ i jest tak duża, że trudno przy-

puszczać, aby ich źródłem mogły być reakcje jądrowe. Słońce na przykład wysyła mniej niż 10^{-6} całkowitej wypromieniowywanej energii w rentgenowskim obszarze widma.

model pulsara rentgenowskiego

Dane obserwacyjne wskazują na to, że pulsary rentgenowskie są układami gwiazd podwójnych. Model takiego pulsara powinien wyjaśniać, skąd się biorą znaczne ilości energii wypromieniowywane przez te obiekty. Obecnie powszechnie przyjęty jest następujący model takiego pulsara. Gwiazda neutronowa lub czarna dziura wchodzi w skład układu podwójnego ze zwykłą gwiazdą ciągu głównego lub czerwonym olbrzymem (rys. 17). Część materii ze zwykłej gwiazdy przepływa i zostaje wychwytywana przez



Rys. 17. Gwiazda neutronowa w układzie podwójnym. Materia przepływa od zwykłej gwiazdy ku gwieździe neutronowej. W silnym polu grawitacyjnym gwiazdy neutronowej cząstki są przyspieszane do bardzo dużych prędkości. W wyniku zderzeń ulegają jonizacji i gdy osiągają prędkości zbliżone do prędkości światła wysyłają promieniowanie rentgenowskie

gwiazdę neutronową lub czarną dziurę. Ten proces wychwytu nazwano akrecją. Przyspieszona do znacznych prędkości materia przyciągana przez gwiazdę neutronową lub czarną dziurę wypromieniowuje znaczne ilości energii. Natężenie promieniowania będzie takie jak obserwowane, jeżeli strumień masy spadającej na gwiazdę neutronową w ciągu roku wynosi ok. 10^{-11} masy Słońca. Strumień ten może być jeszcze mniejszy w przypadku czarnej dziury. Przewidywania tego modelu jak dotychczas są zgodne z danymi obserwacyjnymi.

Astrofizyczne znaczenie pulsarów

Odkrycie pulsarów miało ogromne znaczenie przede wszystkim jako ostateczny argument świadczący o istnieniu gwiazd neutronowych. Teoria ostatnich faz ewolucji gwiazd uzyskała pierwsze obserwacyjne potwierdzenie. Rozważania dotyczące procesów prowadzących do powstania gwiazd neutronowych są dzięki temu jednym z głównych zagadnień astrofizyki. Bardzo ważne są badania struktury gwiazd neutronowych i procesów fizycznych w pobliżu ich powierzchni, choć nadal fascynującym jest problem mechanizmu promieniowania pulsarów, który nie znalazł dotychczas zadowalającego rozwiązania. Dzięki odkryciu pulsarów zaczęto badać własności plazmy w silnych polach magnetycznych, własności pola magnetycznego wokół rotującej namagnesowanej gwiazdy neutronowej i procesy fizyczne zachodzące w pobliżu cylindra świetlnego.

informacje otrzymywane z obserwacji pulsara

Za pomocą pulsarów można dokonywać nowych obserwacji dostarczających informacji o stanie i własnościach materii międzygwiezdnej. Porównując kształt impulsu dla różnych częstości można otrzymać informacje o rozkładzie elektronów wzdłuż promienia widzenia. W ten sposób oceniono rozmiary obszarów turbulentnych w kilku obłokach gazu międzygwiezdnego. Wysuwane są propozycje pomiaru koncentracji cząsteczek i rodników, takich jak H_2, SH, OH i CH, występujących w materii międzygwiezdnej, przez powiązanie jej z czasem przybycia tego samego sygnału na różnych częstościach, czyli z dyspersją. Mierząc różnicę czasów przybycia sygnału na różnych częstościach można obliczyć średnią ilość elektronów, jaką napotkał on na swej drodze. Z pomiarów radiowych można też wyznaczyć koncentrację wodoru w tych samych obszarach. Znając te dwie wielkości łatwo już wyznaczyć stopień jonizacji, a tym samym temperaturę gazu międzygwiezdnego. Zawiera się ona pomiędzy 10^3 a 10^4 K.

Wkrótce po odkryciu pulsarów stwierdzono, że mierząc kąt polaryzacji sygnału na różnych częstościach można wyznaczyć średnią wartość pola magnetycznego wzdłuż promienia widzenia. Okazało się, że w przestrzeni międzygwiezdnej istnieje niemal jednorodne pole magnetyczne o natężeniu $3 \cdot 10^{-10}$ T.

Pulsary mogą być również źródłem promieniowania kosmicznego. Obecnie trudno jeszcze mierzyć anizotropię tego promieniowania. Promieniowanie kosmiczne jest bowiem bardzo silnie zaburzone przez pole magnetyczne Ziemi. Na dokładne dane obserwacyjne trzeba będzie zatem jeszcze poczekać, a tymczasem pulsary można traktować jako gigantyczne akceleratory zapełniające przestrzeń międzygwiezdną wysokoenergetycznymi cząstkami promieniowania kosmicznego.

Odkryte niedawno pulsary, wchodzące w skład układów podwójnych, mogą dostarczyć wielu informacji o ewolucji takich układów. Za ich pomocą można też będzie sprawdzić przewidywania ogólnej teorii względności.

M. Demiański *Astrofizyka relatywistyczna*, Warszawa 1978; A. Hewish *Pulsars*, Ann. Rev. of Astr. and Astrophys., *Palo Alto* 1970; J. P. Lasota *Magnetosfery pulsarów*, Post. Astr. 24, 173 (1976); *Pulsating Stars — a Nature Reprints*, London 1969; M. Ruderman *Pulsars: Structure and Dynamics*, Ann. Rev. of Astr. and Astrophys., *Palo Alto* 1972.

Czarne dziury i zapadanie grawitacyjne

Marek Demiański

Ciała, które nas otaczają, zawdzięczają swoje własności oddziaływaniom elektrycznym między elementarnymi składnikami — elektronami, jądrami, atomami i cząsteczkami. Siły grawitacyjne, które zgodnie z prawem powszechnego ciążenia są siłami przyciągającymi, można w odniesieniu do ciał o małej masie całkowicie pominąć. Dają one o sobie znać dopiero wówczas, gdy masy ciał są odpowiednio duże. Sił grawitacyjnych nie można pominąć, jeżeli chce się wyjaśnić strukturę ciał niebieskich. Ziemia, planety, Słońce i pozostałe gwiazdy mają kształt zbliżony do kuli dzięki równowadze pomiędzy siłami grawitacyjnymi starającymi się ścisnąć te ciała i siłami ciśnienia wewnętrznego przeciwdziałającymi ściskaniu. Ciśnienie wewnętrzne może być wywołane naprężeniami w ciele stałym, może to być ciśnienie cieczy, jak w przypadku Ziemi i innych planet, lub ciśnienie plazmy — w przypadku gwiazd. Temperatury i gęstości panujące we wnętrzach gwiazd są bowiem tak duże, że atomy są częściowo lub całkowicie zjonizowane i tworzą plazmę — mieszaninę dodatnio i ujemnie naładowanych cząstek. W niektórych typach gwiazd

warunki we wnętrzu gwiazd

istotną rolę odgrywa też ciśnienie promieniowania wywołane strumieniami termicznych fotonów przenikających z centralnych części gwiazdy ku powierzchni. Planety nie mają wewnętrznych źródeł energii i ich struktura nie ulega zmianie od chwili ich uformowania. Gwiazdy natomiast promieniują w otaczającą przestrzeń ogromne ilości energii i trwają przez bardzo długi czas w nie zmienionym stanie dzięki temu, że w ich centralnych częściach przebiegają termojądrowe reakcje spalania wodoru i helu, które uzupełniają wypromieniowaną energię (→ Ewolucja gwiazd). Co się dzieje z gwiazdą, która wyczerpała cały zapas wodoru i helu? Zachwiany jest wówczas jej bilans energii: energia wypromieniowana z powierzchni nie jest uzupełniana i gwiazda zaczyna stygnąć. Temperatura, a więc i ciśnienie we wnętrzu gwiazdy zmniejsza się, co powoduje jej kurczenie się wywołujące z kolei wzrost temperatury, ciśnienia i gęstości w centrum. Tak więc gwiazda kurcząc się zamienia grawitacyjną energię potencjalną na energię cieplną. Ciśnienie plazmy nie narasta wystarczająco szybko na to, aby przeciwdziałać kurczeniu się. Powolne kurczenie się może zostać powstrzymane przez odpowiednio szybki przyrost ciśnienia w centrum gwiazdy w wyniku jakiegoś innego procesu.

gwiazda po wypaleniu H i He

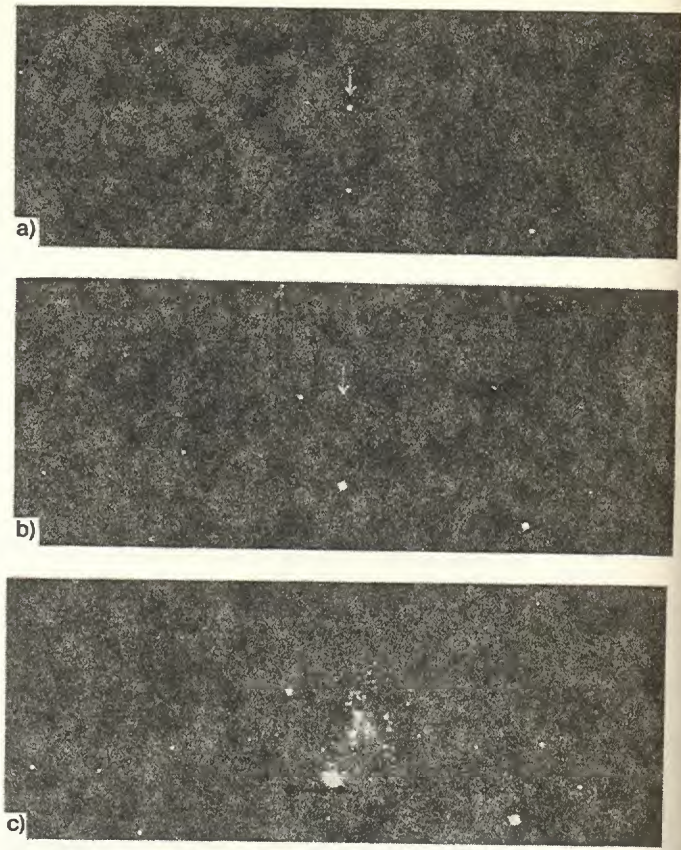
Ostatnie fazy ewolucji gwiazd

Kolejne stadia ewolucji gwiazdy zależą od jej masy. Dla gwiazd o masach mniejszych od 1,4 masy Słońca gęstość w centrum wzrasta na tyle, że elektrony zostają oderwane od atomów i tworzą gaz elektronowy. Siła grawitacyjna jest wówczas równoważona przez ciśnienie gazu elektronowego. Ciśnienie to nie jest wynikiem zderzeń między cząsteczkami, ale tzw. zakazu Pauliego, według którego dwa elektrony nie mogą się znajdować w takim samym stanie kwantowym. W miarę ściskania ciśnienie gazu elektronowego rośnie znacznie szybciej niż ciśnienie plazmy, natomiast temperatura gazu prawie nie ulega zmianie. Gwiazda w takim stadium, zwana białym karłem, stygnąc nie kurczy się, gdyż wraz z obniżeniem się temperatury nie zmniejsza się ciśnienie.

białe karły

Dla gwiazd o masach większych od 1,4 masy Słońca ciśnienie gazu elektronowego nie wystarcza, aby zahamować proces kurczenia. Możliwe są tu trzy warianty ewolucji, które zależą od masy gwiazdy i jej układu chemicznego: a) gwiazda zostaje całkowicie rozerwana przez potężny wybuch wywołany gwałtownym wydzieleniem energii przy spalaniu cięższych pierwiastków np. tlenu, b) gwiazda przechodzi przez stadium supernowej (rys. 1). Zewnętrzne części wraz z otoczką zostają odrzucone, a bardzo gęste i gorące jądro podlega dalszej ewolucji, c) gwiazda katastroficznie się kurczy, gęstość materii w centrum wzrasta na tyle, że do dalszego opisu procesu kurczenia trzeba stosować relatywistyczną teorię grawitacji. W dalszym ciągu zajmiemy się przypadkami b) i c).

Jeżeli masa gorącego i bardzo gęstego jądra pozostającego po wybuchu supernowej nie przekracza krytycznej wartości, która zawiera się w granicach od 1,6 do 3 mas Słońca, to proces dalszego kurczenia się zostaje zatrzymany przez ciśnienie gazu neutronowego (cieczy neutronowej). Należy dokładniej wyjaśnić, w jaki sposób we wnętrzu gwiazdy powstają swobodne neutrony. Proces syntezy termojądrowej prowadzi do powstania ciężkich jąder o liczbach atomowych zbliżonych do żelaza. Przy dalszym ściskaniu takiego układu z energetycznego punktu widzenia korzystny jest wychwyt jednego elektronu przez jądro i utworzenie neutronu z elektronu i protonu. Powstające w ten sposób jądra są w normalnych warunkach nietrwałe ze względu na rozpad β, jednak przy bardzo dużych gęstościach (10^9 g/cm^3) nie rozpadają się, gdyż poziomy energetyczne, które mogłyby zająć elektrony emitowane przy rozpadzie β, są już zajęte. Przy jeszcze większych gęstościach ($3 \cdot 10^{11}$ g/cm^3) energia wiązania kolejnego neutronu maleje do zera i zostaje on oderwany od jądra. Powstają w ten sposób jądra o coraz mniejszej zawartości neutronów



Rys. 1. Wybuch supernowej. Początkowo gwiazda rozbłyskuje (rys. a). Po kilku dniach na zdjęciach o krótkim czasie naświetlania nie jest już widoczna (rys. b), natomiast na zdjęciach o długim czasie naświetlania pojawia się rozszerzająca się otoczka (rys. c)

i coraz mniejszej energii wiązania, a w końcu jądra rozpadają się na swobodne protony i neutrony. W rezultacie materia w centralnych częściach gwiazdy składa się z elektronów, protonów i neutronów, przy czym główny wkład do ciśnienia pochodzi od neutronów. Takie gwiazdy nazwano gwiazdami neutronowymi. Podobnie jak białe karły gwiazdy neutronowe mogą stygnąć nie ulegając już dalszemu kurczeniu się.

gwiazdy neutronowe

Jeżeli masa jądra gwiazdy po wybuchu supernowej przekracza 3 masy Słońca lub gdy masywna gwiazda kurczy się katastroficznie, to gęstości w centrum wzrastają na tyle, że do opisu dalszego procesu kurczenia należy stosować ogólną teorię względności. Aby wyjaśnić, jakich efektów można wówczas oczekiwać, porównamy warunki równowagi sferycznych gwiazd wg teorii grawitacji Newtona i ogólnej teorii względności.

Według Newtona sferyczna gwiazda pozostaje w równowadze, jeżeli w każdym punkcie wewnątrz gwiazdy siła grawitacyjna jest równoważona przez siły ciśnienia, czyli spełnione jest równanie

warunki równowagi gwiazdy

$$dp/dr = -[Gm(r)/r^2]\varrho,$$

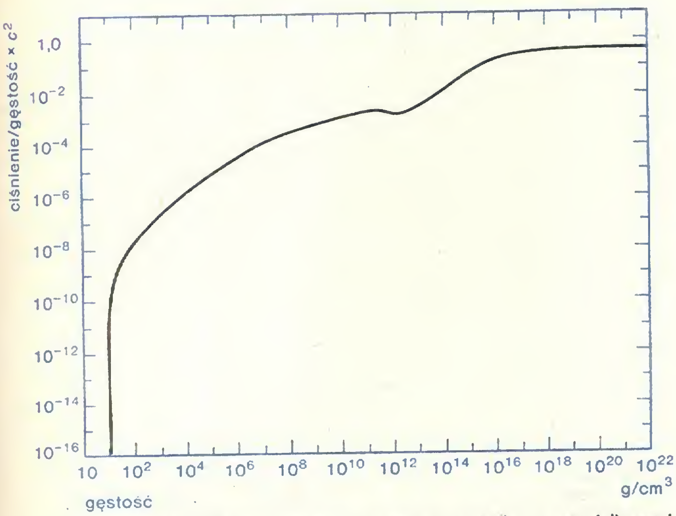
gdzie $m(r) = 4\pi \int_0^r \varrho(r) r^2 dr$. Relatywistyczne równanie równowagi, które ma postać

$$\frac{dp}{dr} = -\frac{G[m(r)+4\pi(p/c^2)r^3](\varrho+p/c^2)}{r^2\left(1-\frac{2Gm(r)}{c^2 r}\right)},$$

zawiera poprawki związane ze zmianą własności geometrycznych przestrzeni w obszarach, w których gęstość materii jest bardzo duża, oraz z wkładem do gęstości energii pochodzącym od energii wzajemnego oddziaływania między elementarnymi składnikami materii (jeśli przyjmiemy, że $c \to \infty$, to relatywistyczne równanie równowagi przechodzi w równanie nierelatywistyczne). Porównując te dwa równania widzimy, że w równaniu relatywistycznym ciśnienie występuje również po prawej stronie, a w mianowniku pojawił się dodatkowy czynnik $1-(2Gm(r)/c^2r)$. Z relatywistycznego równania równowagi wynika, że ciśnienie odpowiadające stanowi równowagi gęstej gwiazdy jest w każdym punkcie wyższe niż powinno być wg teorii Newtona. Efekt ten nazywa się samowzmacnianiem się ciśnienia, bowiem im większe jest ciśnienie, tym większy powinien być gradient ciśnienia (dp/dr), aby utrzymać gwiazdę w równowadze.

samowzmacnianie się ciśnienia

Na to, aby znaleźć strukturę gwiazdy w równowadze posługując się czy to teorią względności, czy teorią Newtona, należy znać równanie stanu (rys. 2). Dla gęstości mniejszych od gęstości materii jądrowej ($2{,}45\cdot10^{14}$ g/cm³) równanie stanu znane jest dość dokładnie, natomiast przy wyższych gęstościach stosuje się różne przybliżenia, wynikające z konieczności opisania słabo jeszcze zbadanych oddziaływań między hadronami, tak że zależnie od stosowanej metody przybliżeń otrzymuje się różne wyniki. Badając konfiguracje gęstej materii trzeba się zatem ograniczyć do pewnych prostych równań stanu opisujących własności układów cząstek poruszających się z prędkościami zbliżonymi do prędkości światła. Pomocne są tu ograniczenia nakładane przez teorię względności; gęstość masy powinna być dodatnia, a prędkość dźwięku nie może być większa od prędkości światła. Przyjmując zależność między ciśnieniem i gęstością postaci:

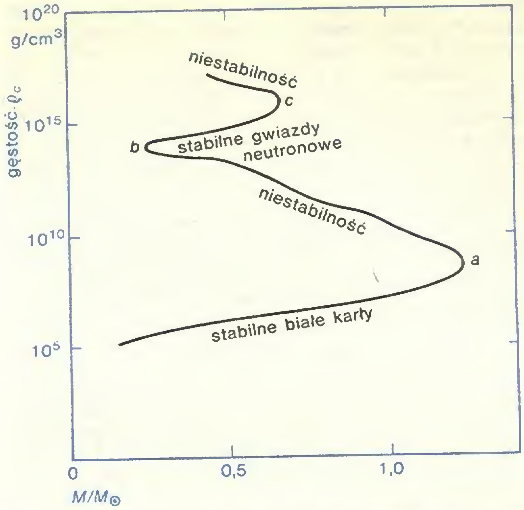


Rys. 2. Równanie stanu (zależność ciśnienia od gęstości) gęstej materii zaproponowane przez Harrisona i J. Wheelera

$$p = (\gamma-1)\varrho c^2$$

o γ należy założyć, iż $1 \leqslant \gamma \leqslant 2$. Dla układu nieoddziałujących cząstek poruszających się z prędkością światła $\gamma = {}^4/_3$.

Korzystając z tej ogólnej postaci równania stanu otrzymano wiele istotnych wniosków dotyczących stabilności bardzo gęstych konfiguracji. Okazało się, niezależnie od wartości γ, że jeżeli gęstość masy w centrum gwiazdy przewyższa 10^{17} g/cm³, to układ nie może się znajdować w stanie stabilnej równowagi (rys. 3). Taka gwiazda kurczy się nieograniczenie pod wpływem sił grawitacyjnych i tego procesu nie może zahamować nawet bardzo szybki wzrost ciśnienia w centrum. Proces katastroficznego kurczenia się pod wpływem niezrównoważonych sił grawitacyjnych nazywamy grawitacyjnym zapadaniem.



Rys. 3. Zależność między masą układu M mierzoną w jednostkach masy Słońca $M_\odot$ a gęstością centralną ρ_c dla gęstej, chłodnej materii spełniającej równanie stanu Harrisona–Wheelera. Pierwsza część krzywej poczynając od początku układu współrzędnych aż do punktu a opisuje stabilne gwiazdy — białe karły i planety. Punkt a odpowiada maksymalnej masie białych karłów $1{,}4M_\odot$. Część krzywej między punktami a i b opisuje niestabilne konfiguracje. Między punktami b i c położone są stabilne gwiazdy neutronowe. Punkt c odpowiada maksymalnej masie gwiazdy neutronowej zbudowanej z materii spełniającej równanie stanu Harrisona–Wheelera. Powyżej punktu c nie ma już stabilnych konfiguracji

grawitacyjne zapadanie

Czarne dziury

Stosując prawa grawitacji Newtona i pamiętając o tym, że prędkość światła jest maksymalną prędkością, z jaką może się poruszać cząstka, łatwo można przewidzieć, że z nieograniczenie kurczącą się gwiazdą dzieją się dziwne rzeczy. Z zasady zachowania energii wynika, że aby cząstkę oderwać od powierzchni ciała o masie M i promieniu R i oddalić do nieskończoności, trzeba jej nadać prędkość nie mniejszą niż

$$v = \sqrt{\frac{2GM}{R}},$$

prędkość ucieczki

zwaną prędkością ucieczki. Jeżeli nie zmieniając masy ciała będziemy je ściskali zmniejszając jego promień, to wówczas gdy osiągnie on wartość

$$R = r_g = \frac{2GM}{c^2},$$

prędkość ucieczki staje się równa prędkości światła. Z powierzchni sferycznego ciała, którego masa i promień spełniają ten związek, nie można wysłać do nieskończoności żadnej cząstki. Charakterystyczny promień r_g nazywamy promieniem Schwarzschilda lub promieniem grawitacyjnym ciała. Dla zwykłych ciał, cząstek elementarnych, planet i gwiazd ich promienie są znacznie większe od promienia Schwarzschilda i tak np. promień grawitacyjny Ziemi wynosi 1 cm, Słońca 2,95 km, natomiast promień Schwarzschilda naszej Galaktyki — ok. $5\cdot10^{10}$ km, czyli jest zaledwie 350 razy większy od średniej odległości Ziemi od Słońca.

promień Schwarzschilda

W ogólnej teorii względności na promień Schwarzschilda otrzymujemy taki sam wzór. W tym wypadku jednak proces kurczenia się gwiazdy z punktu widzenia dalekiego spoczywającego obserwatora, którego dalej będziemy nazywać obserwatorem w nieskończo-

obserwator w nieskończoności

ności, wygląda inaczej niż w teorii Newtona. Aby wyjaśnić te różnice, wyobraźmy sobie następujący eksperyment: z bardzo daleka wysyłamy radialnie w kierunku gwiazdy cząstkę próbną. Z jego punktu widzenia w ciągu czasu Δt cząstka przemieści się o Δr. Inny spoczywający obserwator, którego cząstka właśnie mija powie, że na jego zegarze odstępowi czasu Δt odpowiada okres czasu $\sqrt{1-r_g/r}\,\Delta t$ (co wynika z własności sferycznie symetrycznego pola grawitacyjnego w ogólnej teorii względności) i w tym czasie cząstka przebywa odległość $\Delta r/\sqrt{1-r_g/r}$. W obu tych wzorach r_g to promień grawitacyjny gwiazdy. Zatem spoczywający obserwator, którego cząstka właśnie mija, przypisze jej prędkość

$$v_r = \frac{1}{1-r_g/r}\, v_\infty,$$

gdzie v_∞ — prędkość cząstki mierzona przez obserwatora w nieskończoności. Prędkość v_r można wyznaczyć z zasady zachowania energii, która dla radialnego ruchu cząstki w sferycznym polu grawitacyjnym przyjmuje postać

$$E = \frac{mc^2}{\sqrt{1-v^2/c^2}}\sqrt{1-r_g/r} = \text{const}.$$

Jeżeli w chwili początkowej cząstka spoczywa w odległości r_0 od centrum, to

$$\sqrt{1-r_g/r}\,/\sqrt{1-v_r^2/c^2} = \sqrt{1-r_g/r_0},$$

skąd ostatecznie mamy

$$v_r = c\sqrt{1-\frac{1-r_g/r}{1-r_g/r_0}}.$$

Cząstka spadająca radialnie na zwykłą gwiazdę zostanie przyspieszona do bardzo dużych prędkości, ale mniejszych od prędkości światła, gdyż otrzymany wyżej związek obowiązuje tylko dla $r > R$ (R — powierzchnia gwiazdy), przy tym $R > r_g$. Natomiast gdy gwiazda kurczy się i jej powierzchnia zbliża się do powierzchni Schwarzschilda, to w granicy dla cząstki znajdującej się na powierzchni $r \to r_g$, $v_r \to c$.

Daleki obserwator przypisze tej samej cząstce prędkość v_∞ równą

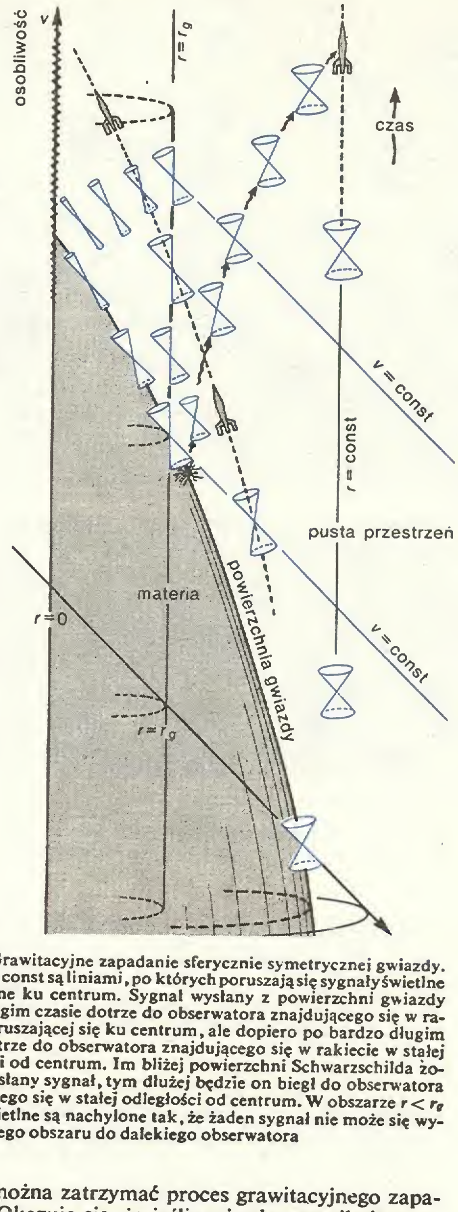
$$v_\infty = (1-r_g/r)v_r = (1-r_g/r)\sqrt{1-\frac{1-r_g/r}{1-r_g/r_0}}\cdot c,$$

a więc z jego punktu widzenia prędkość cząstki zbliżającej się do powierzchni Schwarzschilda maleje i gdy $r \to r_g$, to $v \to 0$. Co więcej, jeżeli dokładnie zbadać ruch cząstki z punktu widzenia obserwatora w nieskończoności, to okazuje się, że cząstka zbliży się do powierzchni Schwarzschilda dopiero po nieskończonym czasie. Efekt ten jest związany ze zwolnieniem tempa biegu zegarów w polu grawitacyjnym.

obserwator na powierzchni

Jak przebiega sferyczne zapadanie gwiazdy z punktu widzenia obserwatora znajdującego się na jej powierzchni? Moment przenikania przez powierzchnię Schwarzschilda nie jest dla niego niczym wyróżniony (rys. 4). Po przekroczeniu powierzchni Schwarzschilda gwiazda nadal się kurczy, aż do chwili, kiedy skurczy się do punktu, przy tym gęstość materii rośnie do nieskończoności. Taki punkt nazywamy osobliwością.

Wspominaliśmy już o tym, że dla gwiazd, planet i innych ciał niebieskich promień Schwarzschilda jest bardzo mały i gdyby je skurczyć do rozmiarów porównywalnych z ich promieniem Schwarzschilda, to gęstość materii wzrosłaby znacznie powyżej gęstości materii jądrowej ($\sim 10^{14}$ g/cm^3). Istotnie, gdyby Ziemię skurczyć do rozmiarów promienia Schwarzschilda, to średnia gęstość wynosiłaby ok. 10^{26} g/cm^3, a dla Słońca byłaby równa 10^{22} g/cm^3. Własności materii o takiej dużej gęstości nie są znane. Nie jest to jednak ogólna prawidłowość. Ciała o bardzo dużej masie mogą osiągać rozmiary porównywalne z promieniem Schwarzschilda, przy średniej gęstości bardzo małej, np. dla naszej Galaktyki wynosi ona zaledwie 10^{-4} g/cm^3, czyli jest mniejsza od gęstości powietrza. Z oszacowań tych wynika, że ciała o rozmiarach zbliżonych do swego promienia Schwarzschilda mogą być zbudowane z materii, której własności znamy bardzo dobrze.

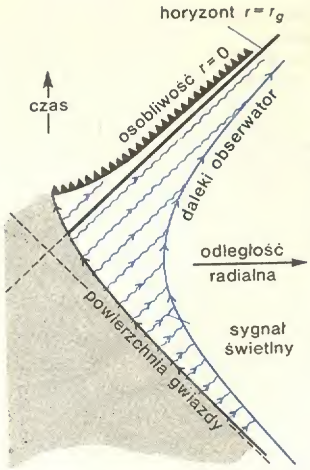


Rys. 4. Grawitacyjne zapadanie sferycznie symetrycznej gwiazdy. Linie v = const są liniami, po których poruszają się sygnały świetlne skierowane ku centrum. Sygnał wysłany z powierzchni gwiazdy po niedługim czasie dotrze do obserwatora znajdującego się w rakiecie poruszającej się ku centrum, ale dopiero po bardzo długim czasie dotrze do obserwatora znajdującego się w rakiecie w stałej odległości od centrum. Im bliżej powierzchni Schwarzschilda zostanie wysłany sygnał, tym dłużej będzie on biegł do obserwatora znajdującego się w stałej odległości od centrum. W obszarze $r < r_g$ stożki świetlne są nachylone tak, że żaden sygnał nie może się wydostać z tego obszaru do dalekiego obserwatora

Czy można zatrzymać proces grawitacyjnego zapadania? Okazuje się, że jeśli gwiazda przeniknie poza swój promień grawitacyjny, to procesu katastroficznego kurczenia zatrzymać już nie można. Ilustruje to rys. 5, na którym przedstawiono zapadającą się gwiazdę. Od momentu gdy powierzchnia gwiazdy przekroczy powierzchnię Schwarzschilda, wszystkie sygnały, nawet sygnały świetlne wysyłane radialnie na zewnątrz, są przyciągane przez silne pola grawitacyjne i zamiast oddalać się do nieskończoności, zbiegają ku centrum. Powierzchnia Schwarzschilda odgrywa teraz rolę półprzepuszczającej membrany, przez którą cząstki i sygnały niosące informację mogą przenikać do środka, ale nie mogą wydostawać się przez nią na zewnątrz. Taką powierzchnię nazywamy horyzontem.

powierzchnia horyzontu

Pole grawitacyjne pod horyzontem nie jest polem statycznym. W tym obszarze żaden obserwator nie może spoczywać i niezależnie od tego czy porusza się swobodnie, czy też znajduje się w rakiecie zaopatrzo-

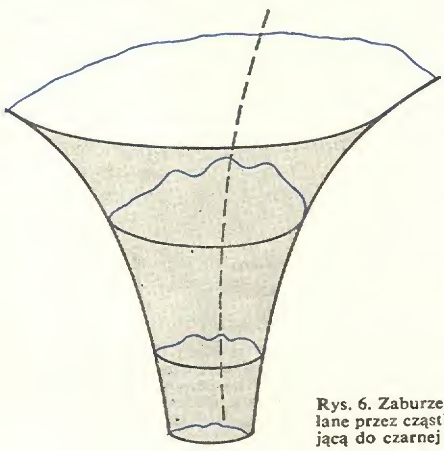


Rys. 5. Czasoprzestrzenny obraz grawitacyjnego zapadania gwiazdy. Sygnały świetlne rozchodzą się po prostych nachylonych pod kątem 45° do osi czasu. Linią falistą zaznaczono sygnały świetlne wysyłane z powierzchni gwiazdy radialnie na zewnątrz. Gdy promień gwiazdy jest większy od r_g, sygnały świetlne wysyłane z jej powierzchni docierają do dalekiego obserwatora, natomiast gdy promień gwiazdy jest mniejszy od r_g, wszystkie sygnały zbiegają do centralnej osobliwości

czarna dziura

nej w potężne silniki, po skończonym czasie zostanie przyciągnięty ku centrum, ku osobliwości. Taki niezwykły obiekt ograniczony zamkniętym, regularnym horyzontem nazywamy czarną dziurą.

Obraz, jaki przedstawiliśmy wyżej, dotyczył obiektu sferycznie symetrycznego. Obserwowane gwiazdy nie są dokładnie sferycznie symetryczne, przeważnie obracają się i mają, czasem nawet silne pola magnetyczne. Jak te dodatkowe czynniki wpływają na proces zapadania? Dokładnie można zbadać tylko sytuację, gdy odchylenia od sferycznej symetrii nie są duże (rys. 6). Wtedy, jak to pierwsi pokazali J. Zeldowicz, I. Nowikow i A. Doroszkiewicz, małe początkowe zaburzenia narastają w czasie zapadania, ale są jeszcze skończone w momencie przekraczania powierzchni Schwarzschilda. Dla dalekiego obserwatora asymetryczne składowe pola grawitacyjnego wytwarzanego przez zapadające się ciało znikają w czasie i pole grawitacyjne asymptotycznie nie zależy od czasu. Stwierdzono też, że gwiazda obracająca się niezbyt szybko, wytwarzająca słabe pole magnetyczne, też może ulegać procesowi zapadania.



Rys. 6. Zaburzenie wywołane przez cząstkę wpadającą do czarnej dziury

Interesujące jest pytanie, czy każdy proces grawitacyjnego zapadania prowadzi do powstania horyzontu, a więc czy z każdej zapadającej się gwiazdy powstanie czarna dziura. Niestety, nie ma na nie zadowalającej odpowiedzi. Wszystkie dotychczasowe próby podania przykładu takiej realistycznej sytuacji, w której w procesie zapadania nie powstaje horyzont, skończyły się niepowodzeniem. Na tej podstawie R. Penrose wprowadził „hipotezę kosmicznego cenzora", według której każdy realistyczny proces zapadania prowadzi do powstania horyzontu.

„hipoteza kosmicznego cenzora"

Ostatnie etapy ewolucji gwiazd o dostatecznie dużej masie (przypuszczalnie większej od 1,4 masy Słońca) powinny przebiegać w następujący sposób: gwiazda, która wyczerpała zapas paliwa jądrowego, zaczyna kurczyć się, co powoduje ściskanie w jej centrum elektronów i fotonów. Przy dużych gęstościach elektrony i fotony zaczynają intensywnie oddziaływać z jądrami powodując ich rozpad. Maleje przy tym liczba elektronów i fotonów, które są głównym źródłem ciśnienia, co prowadzi do powstania niestabilności. W ciągu ułamka sekundy rozpoczyna się proces gwałtownego zapadania. Centralne części gwiazdy zapadają się pociągając za sobą zewnętrzne warstwy. Z punktu widzenia dalekiego obserwatora powierzchnia gwiazdy zbliża się do powierzchni Schwarzschilda, przy czym w miarę przybliżania się do niej prędkość zapadania maleje i gwiazda jak gdyby zastyga, gdy osiągnie rozmiary swego promienia grawitacyjnego. Współporuszający się obserwator powie natomiast, że zapadająca się gwiazda przenika przez powierzchnię Schwarzschilda i następnie bardzo szybko powstaje osobliwość — obszar, w którym gęstość materii i siły grawitacyjne są nieskończenie wielkie. Proces zapadania można podzielić na cztery etapy: powstanie niestabilności, zapadanie się, powstanie horyzontu i powstanie osobliwości, której jednak nie może widzieć daleki obserwator.

Zapadanie grawitacyjne jest końcowym etapem ewolucji gwiazd o masach większych od kilku mas Słońca. Sądzi się, że pewne etapy ewolucji galaktyk zachodzą też przez grawitacyjne zapadanie.

możliwość obserwacji zapadania

Czy można obserwować proces grawitacyjnego zapadania? Aby odpowiedzieć na to pytanie, trzeba znaleźć asymptotyczne wyrażenie na zależność promienia oraz całkowitej jasności gwiazdy od czasu obserwacji t_0. Różnica między czasem obserwacji t_0 a czasem wysłania sygnału t będzie równa czasowi potrzebnemu na przebycie sygnału świetlnego od punktu emisji $r(t)$ (powierzchnia gwiazdy) do punktu obserwacji. Biorąc to pod uwagę otrzymujemy

$$r(t_0) = r_g + (r_1 - r_g) e^{-c(t_0 - t_0')/2r_g},$$

gdzie $r_1 = r(t_0')$, a t_0' — czas odpowiadający momentowi rozpoczęcia się procesu zapadania. Związek ten można uważać za równanie ruchu powierzchni gwiazdy widzialnej przez dalekiego obserwatora. Wynika z niego, że powierzchnia gwiazdy z punktu widzenia dalekiego obserwatora dopiero po nieskończonym czasie $t_0 \to \infty$ osiągnie powierzchnię Schwarzschilda i powstanie wówczas horyzont. Czas, po jakim odległość r zmniejszy się o e $\approx 2{,}7$, równy odwrotności współczynnika przy $2r_g/c$, jest bardzo krótki i rozmiary gwiazdy stają się porównywalne z jej promieniem grawitacyjnym już po upływie setnych części sekundy od momentu rozpoczęcia procesu zapadania. Równie szybko zanika całkowita jasność gwiazdy, spełniona jest bowiem zależność

$$L(t_0) = L(t_0') e^{-2c(t_0 - t_0')/3\sqrt{3}\, r_g}.$$

Charakterystyczny czas zaniku jasności dla gwiazd o masach zbliżonych do masy Słońca wynosi 10^{-4} s. Wydawać by się mogło, że bardzo łatwo jest stwierdzić, czy jakaś wybrana gwiazda uległa grawitacyjnemu zapadaniu czy nie. Należałoby w tym celu co pewien czas fotografować te same fragmenty nieba i sprawdzać, czy przypadkiem z pola widzenia nie zniknęła wybrana gwiazda. Tak jednak nie jest. Zapadająca się gwiazda jest otoczona świecącą atmosferą, której jasność może jeszcze przez długi czas nie ulegać zmianie. Zanik jasności, o którym mowa wyżej, dotyczy tylko jasności powierzchniowej tej części gwiazdy, która się zapada.

Pole grawitacyjne czarnych dziur

Z definicji czarnej dziury wynika, że powierzchnia horyzontu jest powierzchnią zerową, tzn. że jest ona „utkana" z granicznych promieni świetlnych,

które nie zbiegają się do centrum. Wynika stąd ważny wniosek, mianowicie że pole powierzchni horyzontu może rosnąć, ale nigdy nie maleje. Wykazano też, że czarne dziury mogą łączyć się ze sobą, natomiast nie mogą się rozpadać.

stan stacjonarny

Z analizy zaburzeń sferycznych czarnych dziur wynika, że z punktu widzenia dalekiego obserwatora zaburzenia bardzo szybko gasną i czarna dziura asymptotycznie dąży do stanu niezależnego od czasu, czyli do stanu stacjonarnego.

pole grawitacyjne obracającej się czarnej dziury

Początkowo przypuszczano, że podobnie jak dla zwykłych gwiazd pole grawitacyjne wytwarzane przez obracającą się (stacjonarną) czarną dziurę będzie bardzo złożone i aby go opisać, trzeba podać wiele parametrów, takich jak masa, moment pędu i wszystkie momenty charakteryzujące rozkład masy i momentu pędu. Okazało się, że pole grawitacyjne obracającej się czarnej dziury jest bardzo proste i zależy tylko od dwóch parametrów: od masy czarnej dziury i jej momentu pędu. Jeżeli moment pędu czarnej dziury jest równy zeru, to jest ona sferycznie symetryczna i wytwarza sferycznie symetryczne pole grawitacyjne. W ogólnej teorii względności takie pole zostało po raz pierwszy opisane przez Schwarzschilda. Natomiast pole grawitacyjne obracającej się czarnej dziury jest opisywane za pomocą podanego przez Kerra rozwiązania równań ogólnej teorii względności. Promień powierzchni horyzontu obracającej się czarnej dziury o masie m i momencie pędu na jednostkę masy $a = J/mc$ wynosi

$$r = Gm/c^2 + \sqrt{(Gm/c^2)^2 - a^2},$$

z czego wynika, że czarna dziura nie może obracać się zbyt szybko, gdyż wówczas wyrażenie pod pierwiastkiem mogłoby być ujemne.

zachowanie się zaburzeń

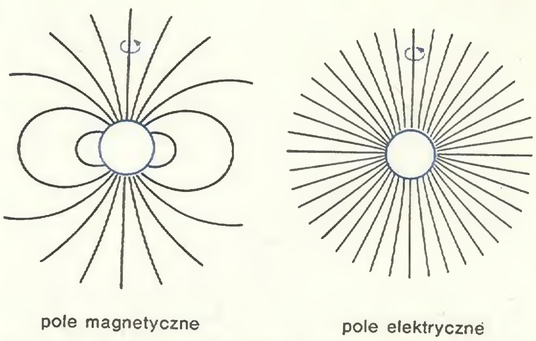
Bardzo ważnym zagadnieniem jest zbadanie zachowania się małych zaburzeń sferycznie symetrycznej (statycznej) i obracającej się (stacjonarnej) czarnej dziury. Jeżeli zaburzenia te nie narastają z biegiem czasu, to czarne dziury są stabilne. Tylko wówczas można przypuszczać, że czarne dziury występują w przyrodzie. Stabilność sferycznie symetrycznego pola grawitacyjnego zbadano już niemal dwadzieścia lat temu. Analizując zaburzenia w liniowym przybliżeniu wykazano, że te zaburzenia, które opisują rozchodzące się fale grawitacyjne, gasną bardzo szybko w czasie, a spośród zaburzeń stacjonarnych jedynie zaburzenia związane z obrotem spełniają warunki regularności w nieskończoności i w otoczeniu horyzontu. Sferycznie symetryczne czarne dziury są zatem stabilne ze względu na małe zaburzenia.

Znacznie trudniej było zbadać problem stabilności obracających się czarnych dziur. Z powodu złożoności równań opisujących małe zaburzenia trzeba się było ograniczyć do analizy numerycznej, którą przeprowadzono już dla bardzo szerokiej klasy zaburzeń. Dotychczasowe wyniki świadczą o tym, że obracające się czarne dziury są stabilne.

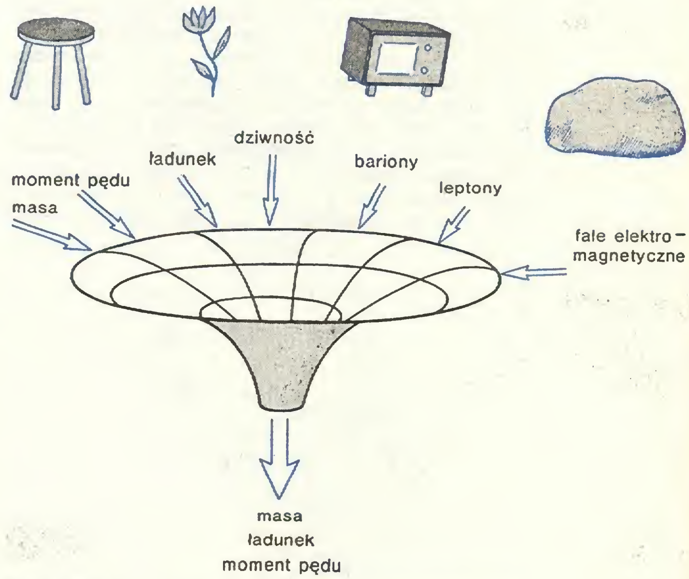
Wszystko wskazuje na to, że czarne dziury powinny istnieć w przyrodzie, ale oczywiście w pełni przekonywającego argumentu mogą dostarczyć tylko obserwacje. Zanim odwołamy się do danych obserwacyjnych zajmijmy się ogólnymi własnościami czarnych dziur.

Obracające się czarne dziury

Początkowo przypuszczano, że pole grawitacyjne wokół czarnej dziury jest podobne do pola grawitacyjnego na zewnątrz zwykłej gęstej obracającej się gwiazdy. Okazało się jednak, że pole grawitacylne na zewnątrz czarnej dziury ma bardzo specyficzne własności. Jest ono całkowicie scharakteryzowane przez dwa parametry: masę i moment pędu. Z ogólniejszych rozważań wynika, że czarna dziura może być nośnikiem ładunku elektrycznego i ładunku magnetycznego, gdyby się okazało, że taki istnieje. Linie sił pola elektrycznego i pola magnetycznego wokół takiej czarnej dziury przedstawione są na rys. 7. Zatem masa, moment pędu oraz ładunek elektryczny i magnetyczny są to jedyne parametry charakteryzujące czarną dziurę. John Wheeler powiada lapidarnie, że czarna dziura nie ma włosów (rys. 8).



Rys. 7. Linie sił pola magnetycznego i elektrycznego wokół czarnej dziury z ładunkiem elektrycznym i magnetycznym



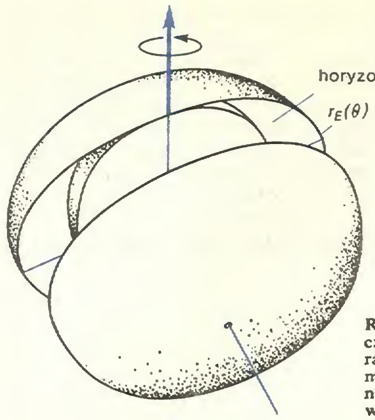
Rys. 8. Wyidealizowany obraz czarnej dziury. Stacjonarną czarną dziurę charakteryzują jedynie trzy parametry: masa, ładunek elektryczny i moment pędu. Innych parametrów opisujących materię, z której powstała czarna dziura, nie jesteśmy w stanie wyznaczyć

Omówimy teraz ogólne własności stacjonarnej czarnej dziury. Jak już stwierdziliśmy promień horyzontu jest dany równaniem $r = Gm/c^2 + \sqrt{(Gm/c^2)^2 - a^2}$, a więc przy takiej samej masie m promień powierzchni horyzontu obracającej się czarnej dziury jest mniejszy od promienia horyzontu czarnej dziury Schwarzschilda. Nasza intuicja tutaj zawodzi, wydawać by się mogło, że obrót spowoduje spłaszczenie czarnej dziury i powiększenie promienia horyzontu.

Wokół obracającej się czarnej dziury istnieje obszar o promieniu $r_E = Gm/c^2 + \sqrt{(Gm/c^2)^2 - a^2\cos^2\theta}$, w którym żadna cząstka ani żaden obserwator nie może spoczywać.

ergosfera

Obszar między powierzchnią horyzontu a powierzchnią $r = r_E\ (\theta)$ nazwano ergosferą. Zauważmy, że ergosfera w dwóch punktach wzdłuż osi obrotu jest styczna do powierzchni horyzontu (rys. 9). Istnienie ergosfery odgrywa ważną rolę w procesach zachodzących w otoczeniu czarnej dziury. Na przykład dzięki

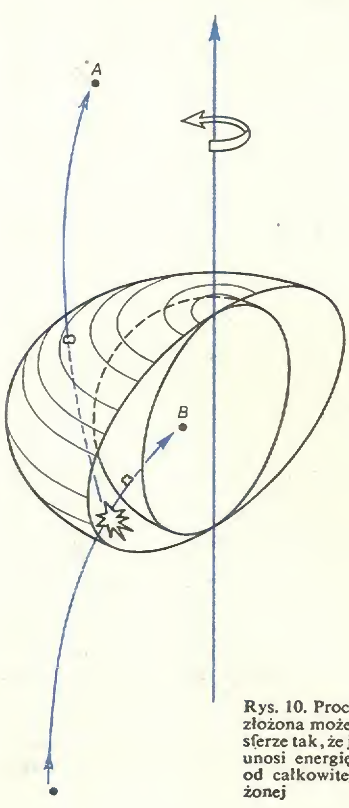


Rys. 9. Ergosfera czarnej dziury Kerra, to obszar pomiędzy powierzchnią horyzontu a powierzchnią $r = r_E(\theta)$

ergosferze można wykorzystywać energię rotacyjną obracającej się czarnej dziury. Metoda polega na tym, że daleki obserwator wyrzuca w kierunku czarnej dziury cząstkę złożoną z dwóch części A i B, która po przeniknięciu do ergosfery rozpada się (rys. 10). Cząstka A wraca do obserwatora, natomiast cząstka B, mająca z punktu widzenia dalekiego obserwatora energię ujemną, wpada pod horyzont. Okazuje przy tym, że energia cząstki A jest większa od energii cząstki złożonej. Oczywiście, nie ma tu żadnej sprzeczności i zasada zachowania energii i momentu pędu jest spełniona. Cząstka B wpadając do czarnej dziury zmniejsza jej moment pędu i jej energię ruchu obrotowego. Tego procesu nie można powtarzać nieskończenie wiele razy, kiedy bowiem czarna dziura przestaje się obracać, nie można już z niej czerpać energii.

własności czarnych dziur

Badając własności czarnych dziur sformułowano ogólne prawa dynamiki tych obiektów. Po pierwsze okazuje się, że siła grawitacyjna działająca na ciało znajdujące się na horyzoncie jest taka sama w każdym punkcie horyzontu. Po drugie pole powierzchni horyzontu nie może maleć. Po trzecie przy zaburzeniach czarnej dziury jej zmiany energii będą takie same, jak dla zwykłej obracającej się gwiazdy, inaczej mówiąc zmiana całkowitej energii czarnej dziury = = zmiana energii powierzchniowej + zmiana energii wewnętrznej + zmiana energii rotacyjnej.



Rys. 10. Proces Penrose'a. Cząstka złożona może rozpadać się w ergosferze tak, że jeden z jej składników unosi energię kinetyczną większą od całkowitej energii cząstki złożonej

Omówimy wnioski wynikające z praw dynamiki stacjonarnych czarnych dziur. Powierzchnia horyzontu obracającej się czarnej dziury wynosi:

$$A = 8\pi\left[(Gm/c^2)^2 + \sqrt{(Gm/c^2)^4 - (GJ/c^3)^2}\right].$$

nieredukowalna masa czarnej dziury

Korzystając z tego wzoru wprowadzono pojęcie nieredukowalnej masy czarnej dziury m_n, czyli masy jaką miałaby nie obracająca się czarna dziura o takiej samej powierzchni horyzontu, zatem

$$A = 16\pi\left(\frac{Gm_n}{c^2}\right)^2,$$

skąd

$$m_n^2 = \frac{1}{2}\left(m^2 + \sqrt{m^4 - \frac{c^2J^2}{G^2}}\right).$$

procesy odwracalne i nieodwracalne

Procesy, w których może brać udział czarna dziura, można podzielić na dwie klasy: procesy odwracalne i nieodwracalne. W procesie odwracalnym może ulegać zmianie masa i moment pędu czarnej dziury, ale tylko w taki sposób, aby masa nieredukowalna pozostała nie zmieniona. Według praw dynamiki czarnych dziur ich masa nieredukowalna nie może maleć. Procesy, w których masa nieredukowalna wzrasta, nazywamy procesami nieodwracalnymi. Część energii która została zużyta na zmianę wartości masy nieredukowalnej, jest energią straconą i nie można jej odzyskać. Warto tu zauważyć, że w przypadku statycznej, nie obracającej się czarnej dziury wszystkie procesy, które przeprowadzają ją znowu w inny stan statyczny, są procesami niodwracalnymi.

źródło energii grawitacyjnej

Kiedy istnieje więcej niż jedna czarna dziura można by z takiego układu czerpać nie tylko energię rotacyjną, ale również energię grawitacyjną. Rozpatrzmy dwie czarne dziury odpowiednio o masach m_1 i m_2 oraz o momentach pędu J_1 i J_2. Przypuśćmy, że następuje zderzenie pomiędzy nimi i w efekcie powstaje czarna dziura o masie m_3 i momencie pędu J_3. W czasie zderzenia część energii grawitacyjnej, a mianowicie $(m_1+m_2-m_3)c^2$, zostanie wypromieniowana w postaci fal grawitacyjnych. Pole powierzchni powstałej czarnej dziury nie może być mniejsze od sumy pól powierzchni zderzających się czarnych dziur, więc stosunek energii wypromieniowanej do energii grawitacyjnej przed zderzeniem $E = (m_1+ +m_2-m_3)/(m_1+m_2)$ nie może być większy od $^1/_2$, czyli w trakcie zderzenia można uzyskać co najwyżej 50% energii grawitacyjnej. Jeżeli zderzają się nie obracające się czarne dziury, to stosunek ten jest mniejszy i wynosi $1-2^{-1/2}$. Wtedy wyzwala się jedynie 29% energii. Są to jednak tylko maksymalne wartości i w realnym procesie zderzenia uzyskane energie będą mniejsze.

cząstka w pobliżu horyzontu

Stacjonarna czarna dziura jest obiektem bardzo trudnym do wykrycia. Daleko od niej pole grawitacyjne jest takie samo jak pole grawitacyjne zwykłej gwiazdy o tej samej masie. Różnice występują w pobliżu horyzontu. Na przykład, na cząstkę poruszającą się w pobliżu obracającej się czarnej dziury działa dodatkowo odśrodkowa siła bezwładności, którą można opisać podając prędkość kątową obrotu, wynosi ona

$$\omega = \frac{2Gmar/c^2}{(r^2+a^2)^2 - a^2\left(r^2 - \frac{2Gm}{c^2}r + a^2\right)\sin^2\theta}.$$

Z taką prędkością kątową obraca się względem dalekiego obserwatora oś spadającego swobodnie giroskopu. Dla zwykłych gwiazd efekt ten jest pomijalnie mały.

Z własności symetrii pola grawitacyjnego wytwarzanego przez obracającą się czarną dziurę wynika, że energia cząstki próbnej oraz rzut momentu pędu na oś obrotu są zachowane. Zupełnie niespodziewanie okazało się, że zachowana jest też wielkość mająca pewne własności całkowitego momentu pędu, dzięki czemu udaje się rozwiązać równania ruchu cząstek próbnych w polu grawitacyjnym obracającej się czarnej dziury.

Czarne dziury jako obiekty astronomiczne

W jaki sposób można zaobserwować czarną dziurę? Na pozór wydaje się, że jest to w ogóle niemożliwe. Czarna dziura nie promieniuje w pustej przestrzeni, a jej pole grawitacyjne dla dalekiego obserwatora nie różni się od pola grawitacyjnego zwykłej gwiazdy. Obserwacyjnie czarna dziura może się ujawnić tylko wówczas, gdy jest otoczona przez materię. Nawet jednak w takich sytuacjach, kiedy mamy czarną dziurę zanurzoną w obłoku pyłu i następuje sferyczny proces akrecji — spadania materii na czarną dziurę — wydzielona energia nie jest zbyt wielka. Nie wiadomo też dokładnie, jakie jest widmo powstającego wówczas promieniowania. Wielkość wyzwolonej energii niewiele się zmieni, jeżeli czarna dziura porusza się względem obłoku materii.

sferyczna akrecja

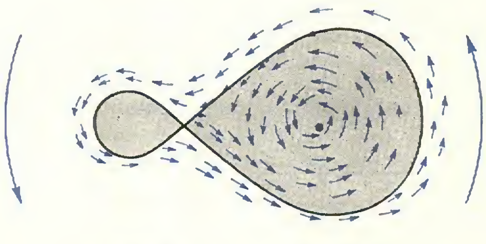
Inną sytuację mamy wówczas, gdy cząstki materii spadają na czarną dziurę z różnym od zera momentem pędu. Wtedy cząstki spadają tylko do odległości, na której siła odśrodkowa jest równoważona przez siły grawitacyjne. Ze względu na osiową symetrię pola grawitacyjnego spadająca materia tworzy dysk wokół czarnej dziury. Dzięki lepkości cząstki tracą moment pędu i promienie ich orbit powoli maleją. Zatem lepkość odgrywa podwójną rolę; wpływa na zmniejszenie momentu pędu i powoduje ogrzanie się dysku. Cząstka spadając powoli w kierunku czarnej dziury ma teraz wystarczająco dużo czasu, aby przekazać otoczeniu znaczną część swojej energii. Dzięki temu energia wydzielona w dysku jest znacznie większa niż przy sferycznej akrecji. Obliczając ilość wydzielonej energii w dysku można oszacować jasność obiektu. Dla szybko obracającej się czarnej dziury, gdy $Jc/Gm^2 = 0{,}998$, efektywność zamiany energii masy spoczynkowej na energię cieplną dochodzi do 35%. Znaczna część tej energii jest wypromieniowana w rentgenowskim obszarze widma elektromagnetycznego. Należy podkreślić, że promieniowanie jest emitowane nie przez czarną dziurę, a przez gorący gaz, który ją otacza.

dysk materii

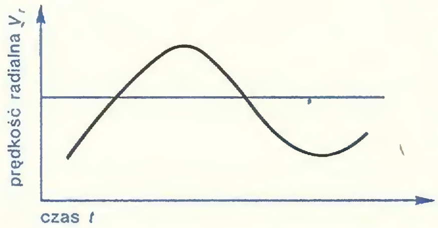
Dyski materii powstają wokół czarnej dziury wówczas, gdy istnieje źródło zapewniające ciągły dopływ materii. Występuje to wtedy, gdy czarna dziura wchodzi w skład układu podwójnego gwiazd i dzięki obecności towarzysza ma zapewniony stały dopływ materii (rys. 11). Układ taki powinien być silnym periodycznie zmiennym źródłem promieniowania *X* (→ Astronomia promieni *X* i *γ*).

Kiedy słynny już satelita Uhuru zaczął przekazywać dane o źródłach promieniowania *X*, największe zainteresowanie wzbudziło źródło Cyg X-1 znajdujące się w konstelacji Łabędzia. Jest to silne źródło promieniowania *X*, którego jasność jest ok. 10 tys. razy większa od jasności Słońca. Dochodzące do nas z Cyg X-1 sygnały fluktuują z okresem mniejszym od 0,1 s, co oznacza, że rozmiary źródła są małe i nie przekraczają $3 \cdot 10^4$ km. Udało się dokonać identyfikacji Cyg X-1 ze źródłem promieniowania widzialnego i okazało się, że istotnie jest to układ podwójny, przy czym okres obiegu wynosi 5,6 dnia (rys. 12). Zmiany jasności tego układu pozwalają oszacować

źródło Cyg X-1



Rys. 11. Czarna dziura w układzie podwójnym. Materia z gwiazdy towarzyszącej powoli spada do czarnej dziury wydzielając znaczne ilości energii

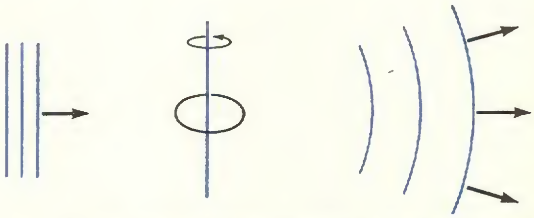


Rys. 12. Zależność prędkości radialnej V_r od czasu t dla układu podwójnego Cyg X-1

masę nie świecącego składnika — powinna ona być większa od czterech mas Słońca. Ponieważ nie może to być gwiazda neutronowa ze względu na zbyt dużą masę wielu astronomów sądzi, że jest to właśnie czarna dziura.

Jeżeli czarne dziury istnieją, to powinny mieć jeszcze jedną ciekawą własność. Podstawowe prawo dynamiki czarnych dziur powiada, że pole powierzchni horyzontu nie może maleć. Inną wielkością fizyczną, która nie może maleć, jest entropia. Biorąc pod uwagę tę bardzo luźną analogię pomiędzy polem powierzchni czarnej dziury a entropią wywnioskowano, że powinna istnieć wielkość, którą można interpretować jako temperaturę. Jeżeli jednak czarnej dziurze można przypisać temperaturę, to powinna ona promieniować, tak jak każde inne nagrzane ciało. Do tego samego wniosku doprowadziła analiza rozpraszania fal grawitacyjnych i fal elektromagnetycznych na obracającej się czarnej dziurze. Obliczono, że przy pewnych częstościach rozpraszane fale unoszą więcej

promieniowanie czarnej dziury



Rys. 13. Fale elektromagnetyczne rozpraszane na obracającej się naładowanej czarnej dziurze mogą unosić więcej energii niż energia jaką miała fala padająca. Ta nadwyżka energii jest uzyskiwana kosztem energii ruchu obrotowego czarnej dziury

energii niż miała jej fala padająca (rys. 13). Stąd już tylko krok do przypuszczenia, że czarna dziura może spontanicznie emitować cząstki. Przy czym im większa jest masa czarnej dziury, tym mniejsza jest jej temperatura i tym mniejsze prawdopodobieństwo emisji. W miarę jak masa promieniującej czarnej dziury maleje, jej temperatura rośnie, więc emituje ona coraz więcej energii. Proces ten jest coraz to szybszy i w końcu, po wypromieniowaniu całej swojej energii czarna dziura przestaje istnieć. Proces ten nie jest sprzeczny z twierdzeniem o niezmniejszaniu się pola powierzchni horyzontu. Przy wyprowadzeniu tego twierdzenia pominięto bowiem efekty kwantowe, które w procesie emisji czarnej dziury odgrywają istotną rolę.

Tempo emisji zależy od masy czarnej dziury i proces emisji nie odgrywa żadnej roli w ewolucji czarnych dziur o masach zbliżonych i większych od masy Słońca.

Czarne dziury powinny też odgrywać pewną rolę w procesach ewolucji galaktyk, kwazarów i innych gęstych układów gwiezdnych. Teoria ewolucji takich układów jest jednak bardzo daleka od doskonałości i wiele punktów jest w niej jeszcze niejasnych

M. Demiański *Astrofizyka relatywistyczna*, Warszawa 1978; C. De Witt, B. S. De Witt *Black Holes*, New York 1973; C. Misner, K. Thorne, J. A. Wheeler *Gravitation*, San Francisko 1973; M. Rees, R. Ruffini, J. A. Wheeler Black Holes, *Gravitational Waves and Cosmology*, New York 1974; J. B. Zeldowicz, J. D. Nowikow *Tieorija tiagotienija i ewolucyja zwiozd*. Moskwa 1971.

Fale grawitacyjne

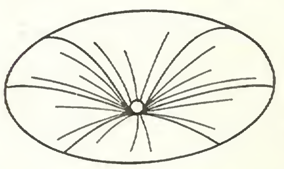
Marek Demiański

Zaburzenia pola elektromagnetycznego — fale elektromagnetyczne — rozchodzące się z prędkością zwaną prędkością światła są nam dobrze znane z codziennego życia. Radio, telewizja, światło — to tylko kilka przykładów praktycznego wykorzystania tych fal. Fale elektromagnetyczne mogą się rozchodzić w próżni, ponieważ ich propagacja nie wymaga istnienia ośrodka materialnego. Falować może samo pole elektromagnetyczne, a w każdym ustalonym punkcie będzie ulegać zmianom wektor natężenia pola elektrycznego i magnetycznego.

prędkość rozchodzenia się sygnałów

Podstawowym warunkiem występowania ruchu falowego jest skończona prędkość rozchodzenia się sygnałów. Wyjaśnijmy to dokładniej. Z równań opisujących własności pola elektromagnetycznego wynika np., że jeżeli raptownie zmienimy położenie ładunku w punkcie A, to informacja o tej zmianie stanu pola dotrze do obserwatora oddalonego od punktu A o r po upływie czasu $t = r/c$, gdzie c jest prędkością światła. Powiadamy, że oddziaływania elektromagnetyczne w próżni rozchodzą się ze skończoną prędkością, równą prędkości światła.

Według teorii grawitacji Newtona, która tak dobrze opisuje ruchy pod wpływem siły grawitacyjnej na powierzchni Ziemi i w Układzie Słonecznym, informacje o zmianie stanu pola rozchodzą się z nieskończoną prędkością. Przemieszczanie jakiejś masy powoduje natychmiastową zmianę pola grawitacyjnego w całej przestrzeni. Fakt ten jest oczywiście sprzeczny z postulatami szczególnej teorii względności, według której prędkość światła jest maksymalną możliwą prędkością przekazywania informacji. W 1916 r. A. Einstein sformułował relatywistyczną teorię grawitacji — ogólną teorię względności. Wiąże ona rozkład materii z geometrycznymi własnościami przestrzeni. Można powiedzieć, że im większą ilość energii zawiera pewna objętość, tym bardziej jest zakrzywiona przestrzeń w tym obszarze (rys. 1). Natężenie pola grawitacyjnego zależy od krzywizny przestrzeni. Wiadomo, że lokalnie, w małych obszarach przestrzeni, nie można za pomocą doświadczeń mechanicznych odróżnić siły grawitacyjnej od sił bezwładności działających w nieinercjalnych układach odniesienia. W geometrycznym języku ogólnej teorii względności oznacza to, że w małych obszarach przestrzeni można wprowadzić prostokątny kartezjański układ współrzędnych.



Rys. 1. Kulka ugina elastyczną membranę. W teorii grawitacji Einsteina masy powodują zakrzywienie przestrzeni

Własności fal grawitacyjnych

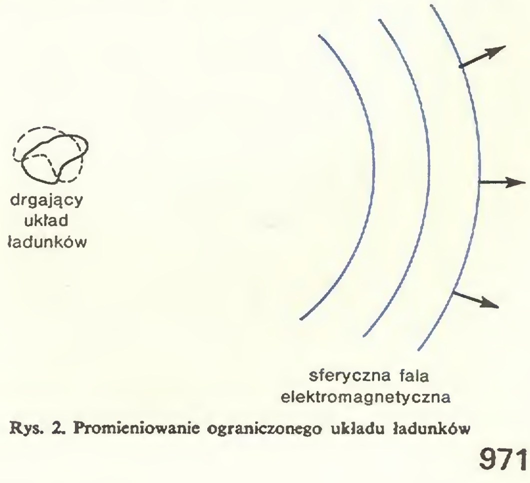
Według ogólnej teorii względności informacje o zmianach stanu pola grawitacyjnego rozchodzą się ze skończoną prędkością. Zatem powinny istnieć pola grawitacyjne o właściwościach podobnych do fal elektromagnetycznych. A. Einstein wykazał, że jeżeli pole jest bardzo słabe, to istnieją rozwiązania równań pola w ogólnej teorii względności. Tym samym wykazał, że można mówić o falach grawitacyjnych, jeżeli pola są słabe. Nie oznacza to jednak, że dowolne pole grawitacyjne dopuszcza możliwość występowania fal grawitacyjnych. Wyjaśnienie, czy tak jest istotnie, nie jest proste. Jedną z trudności jest nieliniowość równań opisujących własności pola grawitacyjnego. Jeżeli np. mamy dwa źródła, A i B wytwarzające pola grawitacyjne, to pole wytworzone jednocześnie przez oba źródła nie jest po prostu sumą pól wytwarzanych przez źródła A i B z osobna. Powiadamy, że relatywistyczne pole grawitacyjne nie spełnia zasady superpozycji. Równania opisujące pole elektromagnetyczne są liniowe, a pole spełnia zasadę superpozycji. Zatem fale grawitacyjne nie będą miały wszystkich własności fal elektromagnetycznych.

warunki występowania fal grawitacyjnych

Jakie więc pola grawitacyjne można nazywać falami grawitacyjnymi? Przy rozstrzyganiu tej kwestii skorzystajmy z analogii do własności fal elektromagnetycznych. Okazało się, że pole grawitacyjne wytwarzane przez ograniczony układ ciał daleko od tego obszaru też można przybliżyć przez falę kulistą o amplitudzie malejącej odwrotnie proporcjonalnie do odległości (rys. 2). Układ taki promieniuje fale grawitacyjne. Jeszcze raz skorzystajmy z analogii dotyczącej własności fal elektromagnetycznych. Podobnie jak układ ładunków promieniujący fale elektromagnetyczne traci energię, a jej straty są równe energii unoszonej przez fale, tak i ograniczony układ ciał oddziałujących grawitacyjnie traci energię, jeżeli promieniuje fale grawitacyjne.

analogie z falami elektromagnetycznymi

Z analogii tych wynika, że o tym, czy dane pole grawitacyjne można nazwać falą grawitacyjną czy nie, decydują własności pola bardzo daleko od źródeł. Jest to istotne ograniczenie, dotyczy bowiem tylko takich sytuacji, w których materia występuje w obszarach o skończonych rozmiarach. O falach grawitacyjnych można mówić i w ogólniejszym wypadku. Jak już wspominaliśmy, zakrzywienie przestrzeni w jakimś obszarze zależy od ilości zawartej w nim energii.



Rys. 2. Promieniowanie ograniczonego układu ładunków

Zakrzywienie przestrzeni, podobnie jak zakrzywienie dwuwymiarowych powierzchni, można opisywać za pomocą wielkości zwanej krzywizną. Skalarną wielkością określającą zakrzywienie przestrzeni jest promień krzywizny. Rozpatrzmy obszar przestrzeni, w którym promień krzywizny jest nie mniejszy od R. Zaburzenia pola grawitacyjnego w obszarach znacznie mniejszych od R można traktować jako fale grawitacyjne rozchodzące się w krzywej przestrzeni. Ten drugi punkt widzenia umożliwia rozważanie występowania i analizowanie fal grawitacyjnych w przestrzeniach z nieograniczonym rozkładem materii, które są wykorzystywane jako proste modele Wszechświata.

pola stacjonalne i promieniste

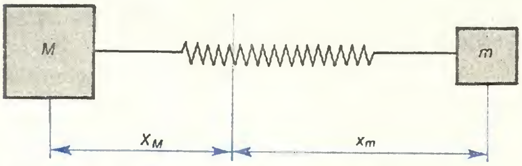
Za pomocą tych dwóch sposobów opisu fal grawitacyjnych można klasyfikować pola grawitacyjne na pola stacjonarne (bez promieniowania) i pola promieniste. Ścisłe rozwiązania równań ogólnej teorii względności opisujące pola promieniste znane są w najprostszych wypadkach, opisują one płaską falę grawitacyjną i rozchodzącą się falę kulistą. W elektrodynamice każde pole promieniowania można dzięki liniowości równań — rozłożyć na sumę fal płaskich lub kulistych, przy czym różne składniki tej sumy nie oddziałują ze sobą. Inaczej jest w ogólnej teorii względności: jeżeli się nawet rozłoży grawitacyjne pole promieniowania na sumę fal płaskich lub kulistych, to różne składniki tej sumy oddziałują ze sobą. Innymi słowy, współczynniki rozkładu pola promieniowania grawitacyjnego, np. na fale kuliste, zależą od czasu. Aby opisać pole promieniowania grawitacyjnego, trzeba podać, w jaki sposób te współczynniki zależą od czasu. Jest to problem bardzo trudny i do tej pory nikomu się nie udało znaleźć ścisłych rozwiązań opisujących pole promieniowania nawet bardzo prostych układów fizycznych, np. gwiazd podwójnych.

moc promieniowania elektromagnetycznego

Z elektrodynamiki wiadomo, że układ ładunków promieniuje wówczas, gdy moment dipolowy zmienia się w czasie (jeżeli tylko stosunek e/m dla różnych ładunków jest różny). Średnia moc promieniowania w najprostszym wypadku (kiedy moment dipolowy d zmienia się periodycznie z częstością kołową ω, a więc gdy $d = d_0 \sin \omega t$) jest dana wzorem:

$$\frac{dE}{dt} = -\frac{2}{3c^3}\left\langle\frac{d^2d}{dt^2}\right\rangle = -\frac{1}{3}\frac{\omega^4}{c^3}d_0^2,$$

gdzie $\langle\ \rangle$ oznacza wartość średnią w ciągu okresu $T = 2\pi/\omega$. Jak widać, moc promieniowania zależy od drugiej pochodnej po czasie momentu dipolowego. Promieniowanie dipolowe jest dominującym rodzajem promieniowania układu ładunków. Inaczej jest w wypadku układu mas. Rozpatrzmy prosty układ dwóch mas, M i m, połączonych sprężyną (rys. 3),



Rys. 3. Oscylator z dipolowym rozkładem mas

który może drgać tylko wzdłuż jednej osi. Z zasady zachowania pędu wynika, że $M\frac{dx_M}{dt} + m\frac{dx_m}{d} = \frac{dd}{dt} =$ $=$ const, zatem druga pochodna po czasie momentu dipolowego jest równa zeru. Wniosek ten jest bardzo ważny, oznacza bowiem, że grawitacyjne promieniowanie dipolowe nie występuje. Moc promieniowania grawitacyjnego układu ciał zależy od trzech pochodnych po czasie kolejnego wyższego momentu multipolowego. Kolejny moment multipolowy jest macierzą, której składowymi są sumy iloczynów mas przez iloczyny dwóch współrzędnych określających ich położenie, np. $Q_{xy} = \sum_i m_i x_i y_i$. Nazywamy go momentem kwadrupolowym. Układ ciał promieniuje grawitacyjnie wówczas, gdy moment kwadrupolowy zmienia się w czasie. Średnia moc promieniowania takiego układu, którego moment kwadrupolowy zmienia się w czasie z częstością kołową ω, wynosi

moc promieniowania grawitacyjnego

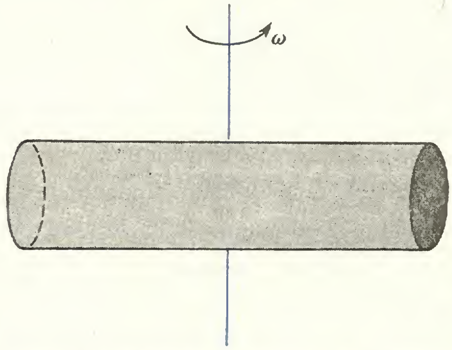
$$\frac{dE}{dt} = -\frac{G\omega^6}{45c^5}Q^2,$$

gdzie G — stała grawitacyjna, c — prędkość światła, Q — średni w czasie moment kwadrupolowy. Czynnik G/c^5 ma wartość $2{,}75\cdot 10^{-60}$, zatem w ogólnym wypadku moc promieniowania grawitacyjnego będzie bardzo mała.

Źródła fal grawitacyjnych

Fale grawitacyjne może emitować nie tylko układ ciał, ale i pojedyncze ciało, nawet traktowane jako ciało punktowe, jeżeli się porusza z bardzo dużym przyspieszeniem. Wyobraźmy sobie, że chcemy zbudować generator fal grawitacyjnych. W tym celu należy zmieniać w czasie moment kwadrupolowy układu mas. Można to osiągnąć np. obracając jakieś ciało względem osi, która nie jest jego osią symetrii. Przypuśćmy, że mamy do dyspozycji stalowy walec o promieniu 1 m i długości 20 m, ważący ok. 500 t. Obracajmy go wokół osi prostopadłej do osi walca i dzielącej ją na połowę (rys. 4). Maksymalną pręd-

laboratoryjne generatory fal grawitacyjnych



Rys. 4. Laboratoryjny generator fal grawitacyjnych — na przykład stalowy walec obracający się ze stałą prędkością kątową względem osi prostopadłej do osi symetrii walca

kość obrotu wyznaczamy z warunku równości odśrodkowej siły bezwładności i siły sprężystej. Stąd otrzymujemy, że maksymalna prędkość kątowa obrotu wynosi 28 s^{-1}. Obracający się z maksymalną możliwą prędkością kątową walec emituje fale grawitacyjne o mocy $dE/dt = 2\cdot 10^{-29}$ J/s. Jest to więc bardzo słabe źródło. Żaden inny laboratoryjny generator nie jest bardziej efektywny.

astrofizyczne źródła fal grawitacyjnych

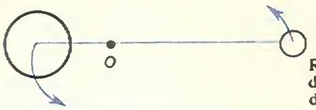
Oczywiste jest zatem, że silnymi źródłami promieniowania grawitacyjnego mogą być tylko obiekty astronomiczne, które się poruszają z bardzo dużą prędkością i mają bardzo silne pola grawitacyjne. Obiektami takimi mogą np. być białe karły, gwiazdy neutronowe i czarne dziury.

Każda wybuchająca gwiazda supernowa czy też gwiazda zapadająca się grawitacyjnie powinna być silnym źródłem fal grawitacyjnych. Ostatnie fazy grawitacyjnego zapadania różnych gwiazd powinny przebiegać podobnie. Można więc przypuszczać, że zapadająca się gwiazda emituje fale grawitacyjne o kształcie charakterystycznym dla tego procesu. Niestety nie znamy dokładnie dynamiki procesu zapadania gwiazd i trudno jest ocenić moc promieniowania, a tym bardziej kształt wysyłanych impulsów.

Innym silnym źródłem fal grawitacyjnych są ukła-

podwójny układ gwiazd

dy podwójne gwiazd (rys. 5). W miarę jak układ podwójny traci energię, gwiazdy zbliżają się do siebie. Energia układu gwiazd poruszających się po orbitach



Rys. 5. Układ gwiazd podwójnych. *O* oznacza środek masy

kołowych o promieniu a wynosi $E = -\frac{1}{2}Gm_1m_2/a$, natomiast straty energii wynoszą

$$\frac{dE}{dt} = -\frac{32}{5}\frac{G}{c^5}m_1^2m_2^2(m_1+m_2)a^{-5},$$

zatem odległość między gwiazdami ciągle się zmniejsza zgodnie z wzorem

$$a = a_0(1-t/\tau_0)^{1/4},$$

gdzie $a_0 = a(0)$, a

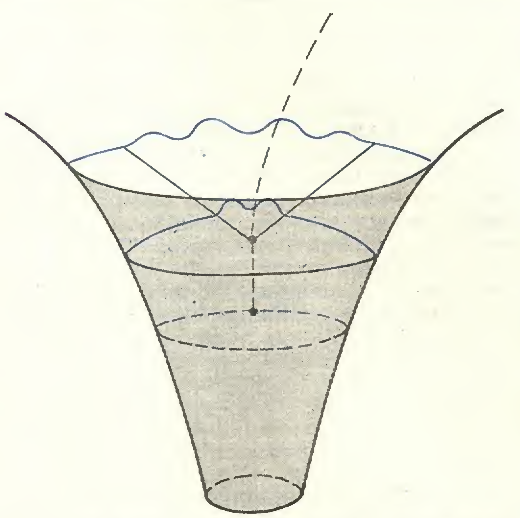
$$\tau_0 = \frac{5}{256}\frac{c^5a_0^4}{G^3m_1m_2(m_1+m_2)}.$$

Jeżeli tylko siły niegrawitacyjne nie zaburzą ruchu układu, to po czasie $t = \tau_0$ gwiazdy spadną na siebie. Czas τ_0 typowych układów podwójnych jest porównywalny z wiekiem Wszechświata i wynosi ok. $(1-1{,}5)\cdot10^{10}$ lat. Moc promieniowania takich układów wynosi 10^{23} J/s, co daje strumień na powierzchni Ziemi ok. 10^{-20} J/cm^2 s.

pulsar

Innym źródłem fal grawitacyjnych może być obracająca się gwiazda neutronowa — pulsar. Strumień promieniowania zależy od momentu kwadrupolowego, który w wypadku pulsarów można ocenić tylko co do rzędu wielkości. Prócz fal grawitacyjnych, wysyłanych z częstością równą okresowi obrotu, obserwowane nagłe zmiany okresu obrotu pulsara, związane ze zmianą kształtu powierzchni, mogą też powodować powstanie silnych impulsów promieniowania grawitacyjnego. Oszacowanie ilości wypromieniowywanej energii jest bardzo trudne ze względu na fragmentaryczne wiadomości o budowie gwiazd neutronowych.

Pole grawitacyjne w pobliżu czarnej dziury jest bardzo silne i rośnie nieograniczenie w miarę zbliżania się do horyzontu (→ Czarne dziury i zapadanie grawitacyjne). Każda cząstka, która zostanie złapana przez czarną dziurę, będzie — zanim przeniknie przez



Rys. 6. Schematyczny dwuwymiarowy obraz zmian krzywizny przestrzeni wywołany przez cząstkę wpadającą do czarnej dziury. Zakrzywienie przestrzeni w otoczeniu czarnej dziury można wyobrażać sobie jako rodzaj lejka, który rozszerza się i staje się płaski bardzo daleko od czarnej dziury. Tor cząstki zaznaczono linią przerywaną, linie faliste reprezentują rozchodzącą się falę grawitacyjną

horyzont — przyspieszona do bardzo dużych prędkości i w momencie przenikania przez horyzont osiągnie prędkość światła. Taka cząstka wyśle silny impuls fal grawitacyjnych. Szacuje się, że mały obiekt o masie m, wpadając do czarnej dziury (rys. 6) o masie M, wypromieniuje ok. 0,01 mc^2 (m/M) energii w postaci impulsu fal grawitacyjnych w czasie $\tau\sim10^{-4}$ (m/M) s.

Cząstka poruszająca się po zamkniętej orbicie kołowej bardzo blisko czarnej dziury też doznaje dużych przyspieszeń. Wysyła ona grawitacyjne promieniowanie synchrotronowe. Aby się cząstka mogła po takiej orbicie poruszać, musi mieć bardzo dużą energię kinetyczną, porównywalną z energią spoczynkową mc^2. Nie znamy jednak naturalnych procesów, które by mogły dostarczyć cząstkom takich dużych energii.

zderzenie dwóch czarnych dziur

Spośród wszystkich rozpatrywanych dotychczas hipotetycznych źródeł fal grawitacyjnych najbardziej interesujące jest zderzenie dwóch czarnych dziur. Z oszacowań wynika, że jeżeli się zderzają dwie nie obracające się czarne dziury o jednakowych masach M, to energia wypromieniowana w postaci fal grawitacyjnych może sięgać $(^1/_3Mc^2$, a jeśli się te czarne dziury obracają — aż $^1/_2mc^2$.

Nagłe wybuchy, które obserwujemy w innych galaktykach, oraz procesy zderzeń w centrum naszej Galaktyki są zapewne źródłami fal grawitacyjnych. Natura tych źródeł jest prawie nieznana. Niektóre własności astrofizycznych źródeł fal grawitacyjnych zebrano w tabeli.

Astrofizyczne źródła promieniowania grawitacyjnego

Źródło	Widmo	Strumień J/m^2·s	Amplituda
Gwiazdy podwójne w Galaktyce są monochromatycznym źródłem promieniowania grawitacyjnego			
Nasilniejsze źródło (i Boo)	$P\approx1$ h	10^{-13}	10^{-20}
Strumień od wszystkich gwiazd podwójnych	$P_{max}\approx8$ h	10^{-10}	10^{-20}
Pulsar w Krabie (PSR 0532)			
Promieniowanie ciągłe	$\nu=60$ Hz	$3\cdot10^{-16}$	10^{-27}
Gwałtowne zmiany prędkości obrotu	$\nu\approx10^3$ Hz	10^{-3}	10^{-21}
Supernowa i zapadanie			
w Galaktyce raz na 100 lat wybuchy trwające od 10^{-3} do 1 s	$\nu\approx10^3$–10^4 Hz	10^7	10^{-17}
we wszystkich galaktykach leżących bliżej niż gromada galaktyki Virgo		10	10^{-20}
Szybko obracająca się gwiazda neutronowa	$\nu\approx10^3$ Hz	10^{-3}	10^{-22}
Wybuchy w jądrach galaktyk i kwazarów	$P\approx100$ d	10^{-16}	10^{-21}
Czarna dziura o masie 10^5–$10^8M_\odot$ w centrum Galaktyki, do której wpada gwiazda o masie $1M_\odot$			
Krótki impuls	$P\approx10$–10^4 s	10^{-6}	10^{-19}

Ostatnio rozważa się inną możliwość generacji fal grawitacyjnych, a mianowicie wytwarzanie fal grawitacyjnych przez szybkozmienne i bardzo silne pola elektromagnetyczne. W tym celu należałoby zbudować potężny kondensator o polu powierzchni płytki ok. 100 m^2 i między okładkami wytwarzać szybkozmienne pole o natężeniu ok. 10^5 V/m. Niestety w warunkach laboratoryjnych nie umiemy jeszcze otrzymywać pól o takich natężeniach.

Wykrywanie fal grawitacyjnych

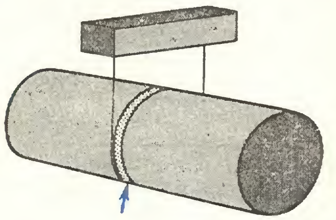
Oddziaływanie grawitacyjne jest bardzo słabe i z zasady zachowania pędu wynika, że nie występuje promieniowanie dipolowe. Generowanie i detekcja fal grawitacyjnych związana jest więc ze zmianami mo-

mentu kwadrupolowego. Zatem generacja fal grawitacyjnych o mierzalnym natężeniu nie jest w laboratoryjnych warunkach praktycznie możliwa. Aby rejestrować fale grawitacyjne, trzeba by umieć mierzyć bardzo małe zmiany momentu kwadrupolowego anteny. Równanie opisujące zmiany w czasie momentu kwadrupolowego anteny przypomina równanie oscylatora tłumionego, na który działa siła wymuszająca, przy czym człon z siłą wymuszającą zawiera informacje o zmianach krzywizny przestrzeni.

Joseph Weber nie zraził się wspomnianymi trudnościami i od kilkunastu lat prowadzi badania w celu wykrycia fal grawitacyjnych. Aby rejestrować fale grawitacyjne i wyznaczać ich natężenie, trzeba umieć mierzyć bardzo słabe zmiany krzywizny przestrzeni. Mierzymy krzywiznę, np. śledząc zmiany odległości dwóch położonych blisko siebie i spadających swobodnie cząstek. Oczywiście taka metoda jest mało efektywna i nikt jej w praktyce nie stosuje.

antena Webera

J. Weber zaproponował, aby jako antenę wykorzystać walec aluminiowy długości 1,5 m, o średnicy 0,6 m i całkowitej masie 1400 kg. Walec zawieszono za pomocą lin stalowych na metalowej ramie odpowiednio izolowanej od wszelkich zewnętrznych wstrząsów (rys. 7). Środkowa część walca została



Rys. 7. Walec aluminiowy używany przez J. Webera jako antena fal grawitacyjnych. Strzałka wskazuje piezoelektryczne kryształy

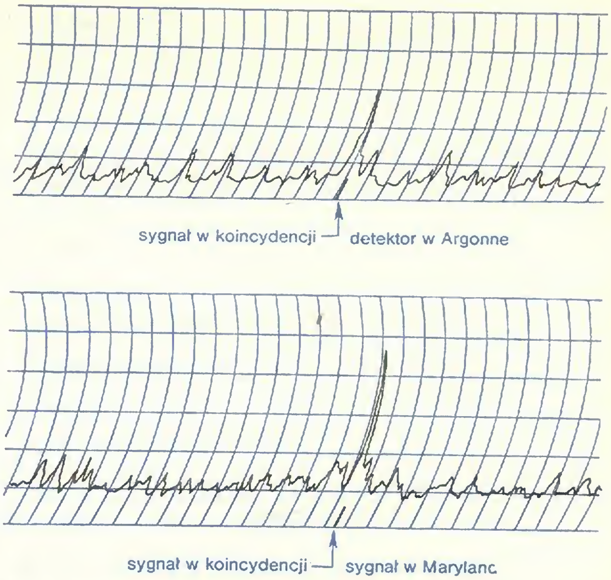
oblepiona piezoelektrycznymi kryształami. Kiedy prostopadle do osi walca pada na niego fala grawitacyjna, zmienia się moment kwadrupolowy, co wywołuje drgania mechaniczne.

Drgania mechaniczne są za pomocą kryształów piezoelektrycznych przetwarzane na drgania elektryczne, które się następnie bardzo wzmacnia i w końcu rejestruje. Podstawowa częstość tego detektora wynosi 1660 Hz. Dzięki temu pomysłowemu urządzeniu Weber mógł mierzyć względne naprężenia rzędu 10^{-16}, czemu odpowiada przesunięcie końców walca o $2 \cdot 10^{-14}$ cm, a więc mniej niż o $^{1}/_{10}$ klasycznego promienia elektronu. To jest fantastyczna dokładność!

Stosując dwa takie detektory, odległe od siebie o 1000 km, Weber badał koincydencje pomiędzy sygnałami dochodzącymi z obu anten. Sygnały z jednej anteny, umieszczonej w pobliżu Chicago, były przekazywane za pomocą linii telefonicznej do Maryland, gdzie pracowała druga antena. Sygnały, które dochodziły jednocześnie z dokładnością do 0,4 s, były liczone jako koincydencyjne. Weber twierdzi, że od 1969 r. obserwuje dziennie ok. 4 koincydencji. Zapis sygnałów pochodzących z obu anten przedstawia rys. 8. Odnosząc swoje dane do czasu gwiazdowego Weber w 1970 r. stwierdził, że obserwuje więcej sygnałów w koincydencji, gdy anteny skierowane są ku centrum Galaktyki. Jeśliby się przypuściło, że źródło promieniowania znajduje się w centrum Galaktyki, to obserwowaną liczbę koincydencji mogłoby dawać źródło przetwarzające na fale grawitacyjne energię masy spoczynkowej ok. $1000 M_{\odot}$ w ciągu roku.

Liczba ta wydawała się astronomom zbyt duża. Przeprowadzili oni oszacowania energii, która może być wypromieniowana z centrum Galaktyki nie powodując zauważalnych ruchów gwiazd. Na górną wartość tej energii otrzymano $70 M_{\odot}$ w ciągu roku, a więc wynik Webera jest o czynnik czternaście za duży.

Od 1970 r. pracuje kilka innych anten. Zbudowali je, wzorując się na modelu Webera, R. Drever



Rys. 8. Fragment zapisu wzbudzeń obu anten Webera

w Glasgow, J. A. Tyson w Holmdel, W. Braginski w Moskwie i in. (il. 219, tabl. 58). Choć wszystkie układy działały na bardzo podobnej zasadzie jak antena Webera, żadnej innej grupie nie udało się potwierdzić obserwacji Webera. Trzeba przyznać, że nikt nie zbudował dokładnie takiej samej pary anten, ale sądzi się powszechnie, że anteny Webera są wzbudzane przez jakieś inne czynniki, nie przez fale grawitacyjne.

W tym czasie powstały nowe projekty anten grawitacyjnych. W. Fairbank buduje antenę, która będzie pracowała w bardzo niskiej temperaturze, ok. $2 \cdot 10^{-3}$ K. Już samo obniżenie temperatury pozwoli na zwiększenie czułości anteny o dwa rzędy wielkości. Znacznemu zmniejszeniu ulegną bowiem szumy anteny związane z termicznym ruchem atomów. Fairbank chce też zastosować nową metodę przetwarzania drgań mechanicznych cylindra na drgania elektryczne, wykorzystując obwody nadprzewodzące; przypuszcza, że w ten sposób podniesie czułość anteny o dalszy rząd wielkości. Jego antena powinna zacząć działać w końcu lat osiemdziesiątych. Fairbank ma nadzieję, że za pomocą swojej anteny będzie mógł obserwować wybuchy supernowych w otaczających nas bliskich galaktykach.

metody detekcji fal grawitacyjnych

Już w 1960 r. Weber wystąpił z projektem użycia Ziemi jako detektora fal grawitacyjnych. W tym celu należałoby dokładnie mierzyć zmiany natężenia pola grawitacyjnego w kilku punktach na powierzchni Ziemi. Pomiary takie są bardzo trudne, gdyż poziom szumów sejsmicznych jest na Ziemi niezwykle wysoki. Postanowiono więc potraktować Księżyc jako detektor. Załoga statku Apollo 17 umieściła na Księżycu grawimetr (niestety układ nie działa).

Od pewnego czasu trwają też dokładne pomiary odległości pomiędzy Ziemią i Księżycem. Mogą one być wykorzystane do wykrywania fal grawitacyjnych o bardzo dużej długości fali.

Rozważa się też kilka innych wariantów anten grawitacyjnych, w których by się wykorzystywało interferencje dwóch strumieni laserowych do badania zmian odległości między dwoma zwierciadłami. Braginski zamierza zbudować antenę, w której podstawowym elementem będzie olbrzymi, ważący ok. 5 kg sztuczny diament.

Bardzo trudno jest przewidzieć, które z tych eksperymentów zakończą się powodzeniem. Być może jesteśmy na progu nowych możliwości i za kilka lat dysponować będziemy nowymi danymi obserwacyjnymi o otaczającym nas Wszechświecie. Może isę

jednak okazać, że podniesienie czułości anten — nawet o cztery rzędy wielkości — jeszcze nie wystarczy a wówczas trzeba będzie czekać na kolejną rewolucję technologiczną, aby eksperymentalnie potwierdzić istnienie fal grawitacyjnych.

Pośrednio na istnienie fal grawitacyjnych wskazują obserwacje pulsara PSR 1913+16. Jest to pulsar radiowy wchodzący w skład układu podwójnego, przy czym drugim składnikiem jest najprawdopodobniej też gwiazda neutronowa. Obserwowane zmiany okresu orbitalnego tego układu dobrze zgadzają się z przypuszczeniem, że wywołuje je emisja fal grawitacyjnych.

W. B. Braginskij, A. B. Manukin *Izmierienije małych sił w fiziczeskich ekspierimientach*, Moskwa 1975; M. Demiański *Astrofizyka relatywistyczna*, Warszawa 1978; M. Rees, R. Ruffini, J. A. Wheeler *Black Holes, Gravitational Waves and Cosmology*, New York 1974; J. Weber *General Relativity and Gravitational Waves*, New York 1961; W. D. Zacharow *Grawitacjonnyje wolny w tieorii tiagotienija Einsztejna*, Moskwa 1972; J. B. Zeldowicz, I. D. Nowikow *Tieorija tiagotienija i ewolucyja zwiozd*, Moskwa, 1971.

Promieniowanie kosmiczne

Marcin Kubiak

Promieniowaniem kosmicznym nazywamy strumień cząstek o bardzo dużych energiach, 10^7–10^{20} eV, dobiegający do Układu Słonecznego z przestrzeni międzygwiazdowej. Cząstki promieniowania kosmicznego wbiegające w atmosferę ziemską z prędkością bliską prędkości światła są źródłem cząstek wtórnych, których rodzaj i energia zależą od wysokości w atmosferze. Promieniowanie kosmiczne nie dociera do powierzchni Ziemi i w swej pierwotnej postaci może być obserwowane tylko na dużych wysokościach (w idealnym wypadku — poza Ziemią, a nawet poza granicami Układu Słonecznego). Aby podkreślić, o który rodzaj cząstek chodzi, używa się niekiedy pojęć: pierwotne i wtórne promieniowanie kosmiczne.

promieniowanie kosmiczne pierwotne i wtórne

Wynikiem oddziaływania promieniowania kosmicznego z atmosferą jest wzrost jej przewodnictwa elektrycznego; w nieobecności jakichkolwiek czynników jonizujących gaz tworzący atmosferę powinien mieć właściwości dielektryka. Fakt, że atmosfera ziemska jest ośrodkiem przewodzącym, znany był od 1900 r., jednak dopiero w 1912 r. V. F. Hess wykazał obserwacyjnie, że przewodnictwo atmosfery wzrasta z wysokością i wysunął potwierdzoną później hipotezę, że zjawisko to jest następstwem jonizacji wywołanej przez przenikliwe promieniowanie pochodzenia pozaziemskiego. Korpuskularny charakter tego promieniowania stwierdzono w 1929 r. Dalsze obserwacje wykazały, że promieniowanie kosmiczne zawiera trzy wyraźnie różne składowe: stały w czasie i niezależny od kierunku strumień cząstek pochodzący spoza Układu Słonecznego, zmienny w czasie strumień cząstek pochodzący ze Słońca, wykazujący wyraźny związek z rozbłyskami na jego powierzchni, oraz mniej intensywny strumień fotonów γ pochodzenia pozasłonecznego. Do czasu wprowadzenia laboratoryjnych akceleratorów cząstek, promieniowanie kosmiczne było dla fizyków jądrowych naturalnym źródłem cząstek i kwantów γ o bardzo dużych energiach. W ostatnich latach promieniowanie kosmiczne stało się przedmiotem szczególnego zainteresowania astrofizyków zajmujących się procesami wielkich energii. Jednocześnie pojęcie promieniowania kosmicznego uległo zawężeniu; promieniowanie korpuskularne, którego źródłem są rozbłyski słoneczne, stało się przedmiotem zainteresowania heliofizyki, natomiast obserwacje kwantów γ rozwinęły się w nową gałąź dzisiejszej astrofizyki obserwacyjnej (→ Astronomia promieni X i γ). Całkowicie nowe możliwości obserwacji promieniowania kosmicznego pojawiły się wraz z rozwojem techniki rakietowej i satelitarnej.

składowe promieniowania kosmicznego

Metody obserwacji promieniowania kosmicznego

Najważniejsze sposoby obserwacji cząstek promieniowania kosmicznego w różnych zakresach ich energii kinetycznej podane są w poniższej tabeli. Celem tych obserwacji jest dokonanie jednoznacznej identyfikacji cząstek, określenie ich widma energetycznego (tzn. względnej liczby cząstek danego rodzaju, o energii zawartej w jednostkowym przedziale wokół danej energii E) oraz kierunku przylotu.

Metody obserwacji promieniowania kosmicznego w różnych zakresach energii

Zakres energii	Metoda obserwacji
$2 \cdot 10^7$–$3 \cdot 10^8$ eV	detektory umieszczone na pokładach sztucznych satelitów o orbitach ekscentrycznych, wybiegających poza magnetosferę
$2 \cdot 10^8$–$5 \cdot 10^{10}$ eV	bloki emulsji i liczniki Czerenkowa wynoszone przez balony stratosferyczne
10^{10}–10^{14} eV	spektrometry jonizacyjne i bloki emulsji rejestrujące reakcje jądrowe, umieszczone na pokładach satelitów i balonów stratosferycznych
10^{12}–10^{16} eV	podziemne pomiary strumienia mionów
10^{13}–10^{20} eV	obserwacje wielkich pęków atmosferycznych

Jednoznaczna identyfikacja cząstki wymaga określenia jej ładunku elektrycznego oraz masy. Wszystkie cząstki promieniowania kosmicznego są jądrami atomowymi pozbawionymi całkowicie powłok elektronowych; określenie ładunku jest więc równoznaczne z określeniem liczby atomowej jądra (czyli liczby znajdujących się w nim ładunków dodatnich). Określenie znaku ładunku pozwala na odróżnienie materii od antymaterii. Do identyfikacji cząstek można wykorzystać pewne właściwości ich oddziaływania z materią. Cząstka promieniowania kosmicznego przechodząca przez gęsty ośrodek traci część swej energii na jonizację lub wzbudzenie tworzących go atomów. Straty te nazywamy stratami jonizacyjnymi. W ogólnym wypadku zależą one od ładunku oraz prędkości cząstki; tak więc oprócz określenia strat jonizacyjnych (czyli ubytku energii na jednostkowej drodze) do identyfikacji cząstki potrzebne są również niezależne informacje o jej prędkości. W przypadku cząstek relatywistycznych (poruszających się z prędkością bliską prędkości światła) straty jonizacyjne zależą praktycznie tylko od ładunku.

identyfikacja cząstek

straty jonizacyjne

Innej możliwości identyfikacji cząstek dostarcza analiza tzw. promieniowania Czerenkowa. Promieniowanie to pojawia się zawsze, gdy prędkość cząstki obdarzonej ładunkiem elektrycznym, poruszającej się w przezroczystym ośrodku o współczynniku załamania n, przekracza wartość c/n, czyli jest większa od prędkości światła w tym ośrodku (pozostając jednak mniejsza od prędkości światła w próżni). Natężenie oraz barwa promieniowania Czerenkowa emitowanego na jednostkowej drodze zależy od ładunku cząstki i od jej prędkości. Gdy cząstki poruszają się z prędkością dużo większą od prędkości granicznej c/n, właściwości promieniowania Czerenkowa zależą tylko od ładunku cząstki.

promieniowanie Czerenkowa

Wymienione wyżej zjawiska, tzn. straty jonizacyjne oraz promieniowanie Czerenkowa, są — w takiej

czy innej odmianie — zasadą działania wszystkich urządzeń stosowanych do obserwacji promieniowania kosmicznego.

pomiar energii cząstek

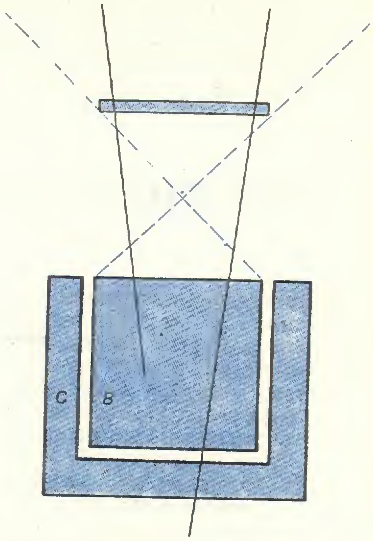
W zakresie energii mniejszych od około $3 \cdot 10^8$ eV straty jonizacyjne mogą doprowadzić do całkowitego zatrzymania cząstki zanim oddziaływania jądrowe zamienią ją w cząstkę lub cząstki innego rodzaju. Całkowita energia cząstki może być wówczas określona np. na podstawie analizy jej śladu w bloku emulsji jądrowej (rys. 1). Innym rodzajem analizatora energii jest urządzenie składające się z dwóch scyntylatorów, którego zasada działania jest wyjaśniona na rys. 2.

Do określenia energii cząstek o większych energiach wykorzystuje się następujące urządzenia, metody oraz zjawiska.

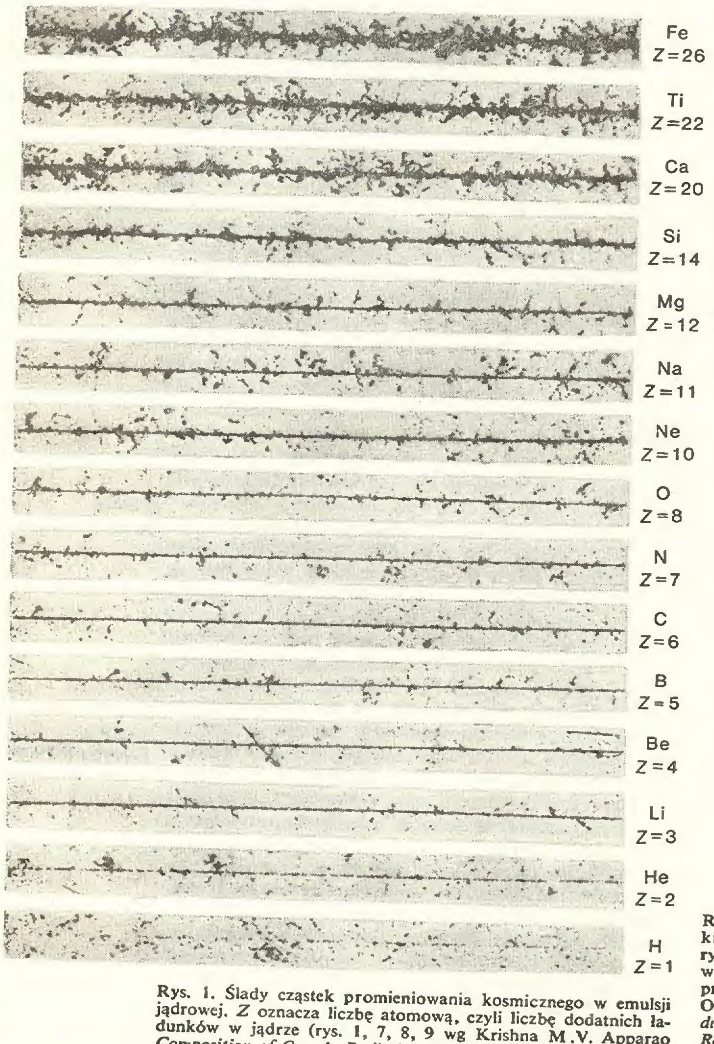
Liczniki Czerenkowa — zależnie od współczynnika załamania użytej substancji rejestrują tylko cząstki o energiach większych od określonej energii progowej.

Spektrometry magnetyczne — urządzenia, w których wykorzystuje się fakt, że cząstka o pędzie p i ładunku Z, poruszająca się w poprzecznym polu magnetycznym o natężeniu B, zakreśla koło o promieniu $r = p/ZB$. Urządzenia tego rodzaju są zazwyczaj używane w połączeniu z układami komór iskrowych lub z blokami emulsji, pozwalającymi dokładnie określić tory cząstek.

Spektrometry jonizacyjne — urządzenia, w których pierwotna cząstka traci swą energię w grubej warstwie absorbującej wskutek oddziaływań jądrowych, prowadzących do powstania kaskad mezonowych. Ostatecznie, niemal cała energia kinetyczna cząstki zamienia się w energię kaskady elektronów i może być zmierzona.



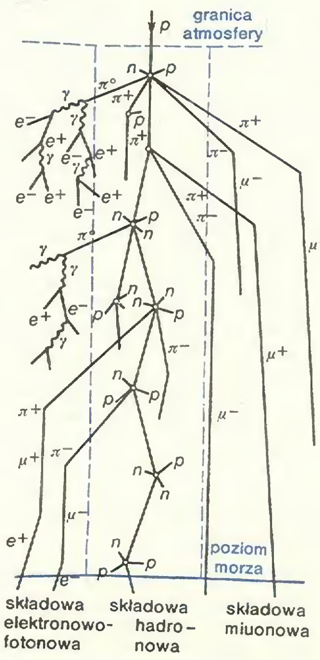
Rys. 2. Zasada działania najprostszego „teleskopu" do obserwacji promieniowania kosmicznego, będącego zarazem analizatorem energii. Rejestrowane są tylko cząstki, które przejdą prawie jednocześnie przez scyntylator *A* i *B*, cząstki przechodzące przez wszystkie trzy scyntylatory nie są rejestrowane. W układzie takim są więc rejestrowane tylko takie cząstki, których energia jest całkowicie tracona w scyntylatorze *B*. Jednocześnie stosunki geometryczne układu określają „pole widzenia" teleskopu, zaznaczone na rysunku niebieską linią przerywaną



Rys. 1. Ślady cząstek promieniowania kosmicznego w emulsji jądrowej. *Z* oznacza liczbę atomową, czyli liczbę dodatnich ładunków w jądrze (rys. 1, 7, 8, 9 wg Krishna M.V. Apparao *Composition of Cosmic Radiation*, London 1975)

wielkie pęki atmosferyczne

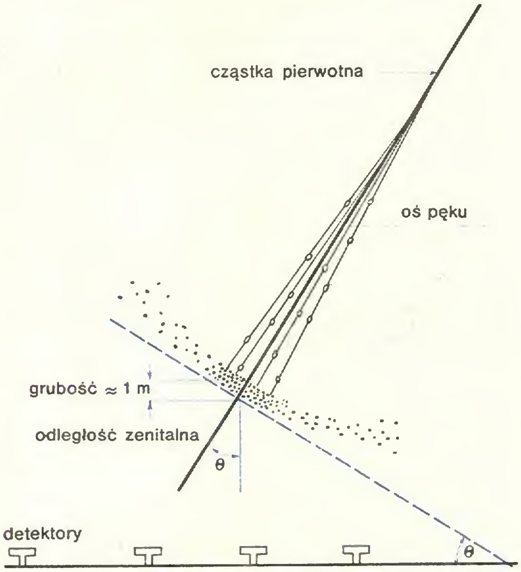
Wielkie pęki atmosferyczne są wykorzystywane do obserwacji cząstek o energiach większych od ok. 10^{15} eV; cząstek o takich energiach nie potrafimy jeszcze rejestrować w sposób bezpośredni. Wielkie pęki atmosferyczne powstają wskutek zderzeń cząstek promieniowania kosmicznego o dużych energiach z jądrami atomów tworzących atmosferę ziemską. Cząstki wtórne produkowane w pęku dobiegają do powierzchni Ziemi, a nawet wnikają pod jej powierzchnię. Układy detektorów, takich jak liczniki



Rys. 3. Schemat wielkiego pęku atmosferycznego zapoczątkowanego przez szybki proton (rys. 3 i 5 wg O. C. Allkofer, *Introduction to Cosmic Radiation*, München 1975)

liczba i rodzaj cząstek pierwotnych

scyntylacyjne lub liczniki Geigera-Müllera, rozmieszczone na powierzchni od kilkuset metrów do kilkunastu kilometrów kwadratowych (pod ziemią, na poziomie morza lub na szczytach gór) pozwalają określić zarówno rodzaj cząstek pierwotnych, jak i ich energię i kierunek przylotu (rys. 3 i 4). Rozwojowi wielkiego pęku atmosferycznego towarzyszy krótkotrwały błysk promieniowania Czerenkowa w dziedzinie widzialnej, który może być zarejestrowany przez fotomnożnik, kliszę fotograficzną lub inny element światłoczuły umieszczony w ognisku niewielkiego nawet teleskopu. Analiza błysków Czerenkowa pozwala na określenie energii cząstki pierwotnej oraz jej kierunku, nie daje jednak informacji na temat rodzaju



Rys. 4. Układ detektorów rozłożonych na powierzchni Ziemi pozwalający na określenie kierunku przylotu cząstki powodującej powstanie wielkiego pęku atmosferycznego

cząstki będącej źródłem pęku (wyjątek stanowią tylko błyski pochodzące od pęków zapoczątkowanych przez kwanty γ; można je stosunkowo łatwo odróżnić od błysków pochodzących od pęków zapoczątkowanych przez cząstki; → Astronomia promieni X i γ). Zastosowanie układów teleskopów optycznych rejestrujących atmosferyczne promieniowanie Czerenkowa rokuje nadzieje na podniesienie dokładności pomiarów strumienia cząstek o dużych energiach.

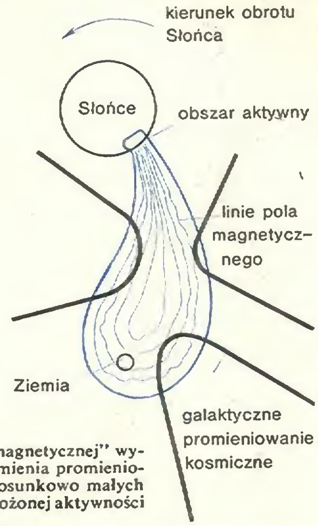
Oddziaływanie promieniowania kosmicznego z magnetosferą i materią międzyplanetarną

Pole magnetyczne, a dokładniej jego składowa poprzeczna, wywiera istotny wpływ na ruch cząstek obdarzonych ładunkiem elektrycznym. Tory cząstek promieniowania kosmicznego są więc określone zarówno przez rozkład pola magnetycznego w Galaktyce (natężenie pola rzędu 10^{-10} T), jak i przez znacznie silniejsze pole magnetyczne otaczające Ziemię w postaci tzw. magnetosfery lub geokorony (→ Fizyka przestrzeni okołoziemskiej). Gdy obserwacje wykonywane są w pobliżu Ziemi kierunek przylotu cząstki promieniowania kosmicznego o energii mniejszej od ok. 10^{12} eV nie daje żadnych informacji na temat kierunku, z którego dana cząstka wbiegła do układu planetarnego. Chwilowy wzrost natężenia pola magnetycznego otaczającego Ziemię może spowodować, że promieniowanie kosmiczne o energii mniejszej od pewnej energii granicznej w ogóle nie dobiegnie do Ziemi. Od 1937 r. znany jest tzw. efekt Forbusha, polegający na tym, że ogólny strumień promieniowania kosmicznego ulega znacznemu zmniejszeniu

efekt Forbusha

w okresach burz geomagnetycznych. Efekt Forbusha tłumaczy się działaniem tzw. butelki magnetycznej (rys. 5): obszar, w którym pole magnetyczne jest wzmocnione, odbija dobiegające z zewnątrz cząstki promieniowania kosmicznego. Efekt Forbusha jest

modulacja promieniowania kosmicznego



Rys. 5. Model „butelki magnetycznej" wyjaśniający osłabienie strumienia promieniowania kosmicznego (o stosunkowo małych energiach) w okresie wzmożonej aktywności słonecznej

jednym z przejawów znacznie ogólniejszego zjawiska modulacji promieniowania kosmicznego w 11-letnim cyklu aktywności słonecznej. Modulacja ta polega na okresowych zmianach natężenia promieniowania kosmicznego, przy czym wielkość tych zmian zależy od energii promieniowania kosmicznego. Zjawisko modulacji słonecznej jest wynikiem oddziaływania cząstek promieniowania kosmicznego z materią i polem magnetycznym, unoszonymi ze Słońca w postaci tzw. wiatru słonecznego. Cząstki promieniowania kosmicznego, dyfundujące poprzez nieregularności pola magnetycznego w wietrze słonecznym, są wynoszone wraz z tymi nieregularnościami z układu planetarnego. Istniejące teorie tego zjawiska pozwalają przewidzieć w sposób ilościowy osłabienie strumienia promieniowania kosmicznego (zależnie od energii) oraz zmianę widma energetycznego cząstek w wyniku oddziaływania z wiatrem słonecznym.

modulacja słoneczna

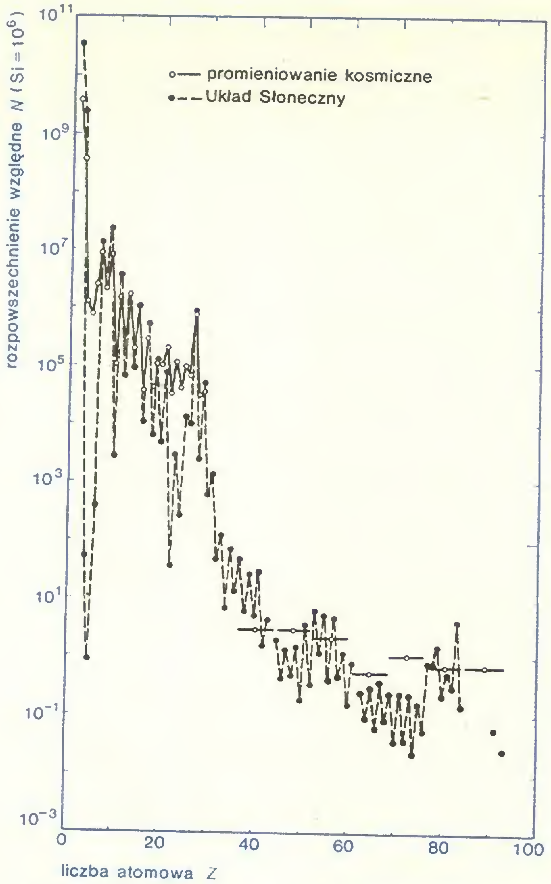
Skład i widmo energii promieniowania kosmicznego

Promieniowanie kosmiczne w pobliżu Ziemi (ponad atmosferą) składa się w 86% z protonów, w 13% z cząstek alfa (jąder helu); około 1% stanowią elektrony i jądra pierwiastków cięższych (o liczbie atomowej $Z \geqslant 3$).

skład promieniowania kosmicznego

Na rys. 6 porównano skład promieniowania kosmicznego ze składem chemicznym materii w Układzie Słonecznym. Wodór i hel występują w promieniowaniu kosmicznym mniej obficie, natomiast pierwiastki o liczbie atomowej $18 < Z < 24$ znacznie obficiej niż w materii Układu Słonecznego. Najbardziej uderzający jest jednak fakt, że pierwiastki lit, beryl i bor mają w promieniowaniu kosmicznym rozpowszechnienie pięć rzędów wielkości większe. Przypuszcza się, że nadobfitość tych jąder jest wynikiem reakcji jądrowych zachodzących między szybkimi cząstkami promieniowania kosmicznego i materią międzygwiazdową; w wyniku tych reakcji jądra ciężkie ulegają rozbiciu na jądra lżejsze.

Jak dotychczas nie udało się wykryć jąder antymaterii w promieniowaniu kosmicznym. Istniejące dane obserwacyjne pozwalają na określenie górnej grasicy rozpowszechnienia antyjąder na 10^{-3}–10^{-4}, zależnie od zakresu energii. Pozytony (antyelektrony obserwowane w promieniowaniu kosmicznym są



Rys. 6. Porównanie rozpowszechnienia *N* pierwiastków w Układzie Słonecznym ze składem promieniowania kosmicznego w pobliżu Ziemi (rys. 6 i 10 wg F. B. McDonald and C. E. Fichtel (ed.) *High Energy Particles and Quanta in Astrophysics*, Cambridge 1974)

cząstkami wtórnymi, powstającymi w wyniku reakcji jądrowych promieniowania kosmicznego z materią międzygwiazdową.

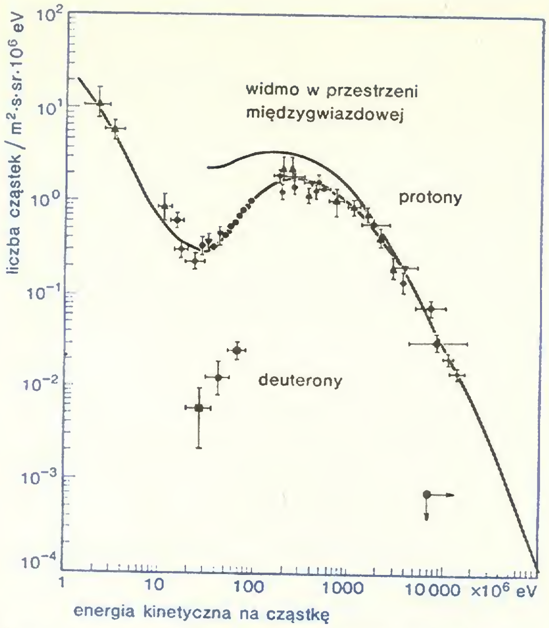
protony

Protony są w zasadzie jedynymi cząstkami w promieniowaniu kosmicznym niosącymi pojedynczy ładunek elementarny. W liczbach bezwzględnych, całkowity strumień protonów o energiach większych od 10^6 eV jest równy ok. 2500 cząstek na m^2 steradian i sekundę. Obok protonów obserwuje się również trwały izotop wodoru — deuter. Drugi izotop wodoru — tryt — najprawdopodobniej nie występuje w promieniowaniu kosmicznym. Tryt jest izotopem nietrwałym i ulega przemianie w He^3 z okresem połowicznego zaniku 12,5 lat. W pobliżu Ziemi można by więc zaobserwować tylko te jądra trytu, które powstały nie dawniej niż kilka razy 12,5 lat. Eksperymentalna górna granica stosunku trytu do protonów jest rzędu 10^{-3}. Różniczkowe widmo energetyczne protonów i deutronów w promieniowaniu kosmicznym jest przedstawione na rys. 7. Protony o energiach mniejszych od 50 MeV (1 MeV = 10^6 eV) były obserwowane z satelitów i sond kosmicznych wybiegających poza magnetosferę Ziemi. Obecnie uważa się że protony te nie są pochodzenia galaktycznego lecz słonecznego, albo też są w jakiś sposób przyspieszane w pobliżu Słońca.

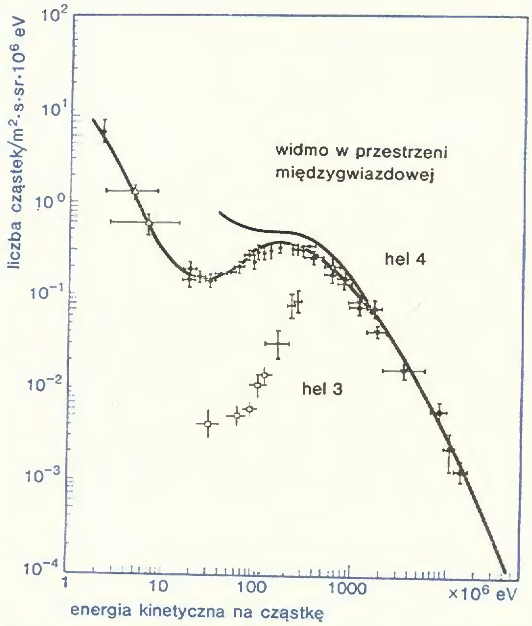
hel

Widmo energetyczne jąder He^4 i He^3 występujących w promieniowaniu kosmicznym jest przedstawione na rys. 8. Widmo He^3 jest określone tylko dla energii mniejszych od 300 MeV; jak dotychczas nie potrafimy jeszcze rozdzielać izotopów o energiach większych od ok. 300 MeV na nukleon.

Strumień jąder pierwiastków ciężkich jest średnio



Rys. 7. Energetyczne widmo protonów i deuteronów. Linią ciągłą przedstawiony jest przebieg widma teoretycznego, uzyskanego po uwzględnieniu efektów modulacji słonecznej



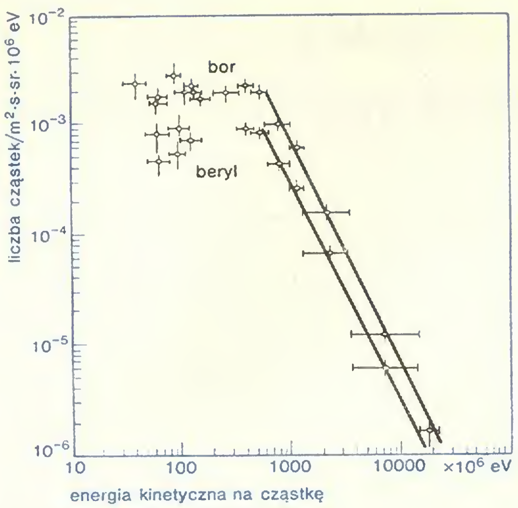
Rys. 8. Energetyczne widmo jąder He^4 i He^3. Linią ciągłą przedstawiony jest przebieg widma teoretycznego, uzyskanego po uwzględnieniu efektów modulacji słonecznej

kilka lub kilkanaście tysięcy razy mniejszy od strumienia protonów. Na rys. 9 przedstawione jest widmo energetyczne jąder boru i berylu. Widma takie są typowe dla widm wszystkich jąder cięższych od helu. Szczególnie interesująca byłaby ewentualna obecność w promieniowaniu jąder pierwiastków transuranowych.

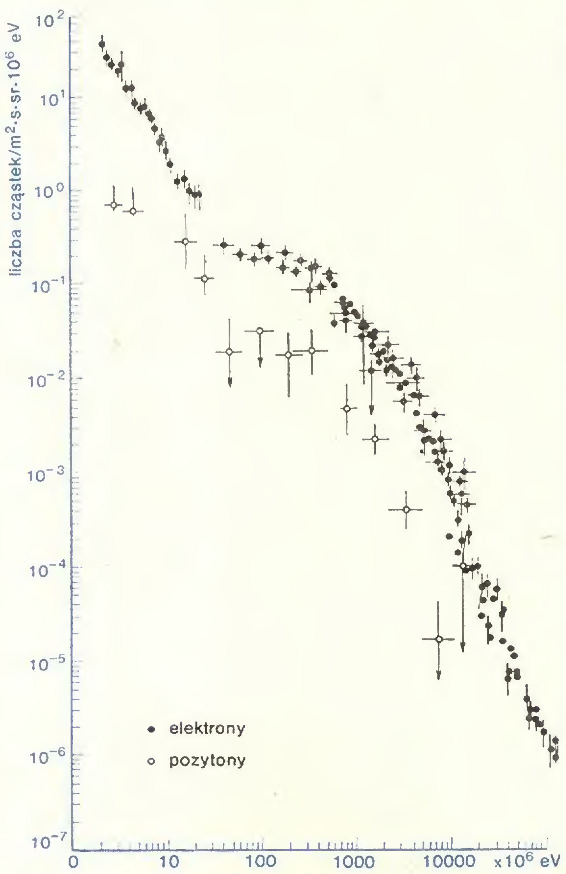
pierwiastki ciężkie

elektrony i pozytony

Widmo energetyczne elektronów i pozytonów w pobliżu Ziemi jest przedstawione na rys. 10. Uwagę zwraca wzrost strumienia elektronów o energiach mniejszych od ok. 20 MeV. Nie zostało jeszcze wyjaśnione, czy elektrony o tych energiach są pochodzenia kosmicznego czy słonecznego. Badania czasowych zmian strumienia elektronów sugerują jednak, że elektrony te są pochodzenia galaktycznego. Pocho-



Rys. 9. Energetyczne widma berylu i boru



Rys. 10. Obserwowane widmo energetyczne elektronów i pozytonów w pobliżu Ziemi

dzenie pozytonów wydaje się obecnie dobrze znane: pozytony o energiach powyżej kilku MeV powstają z rozpadu mezonów π^+ produkowanych w reakcjach jądrowych zachodzących między promieniowaniem kosmicznym i jądrami materii międzygwiazdowej. Pozytony o energiach mniejszych od 1 MeV mogą być pochodzenia zarówno kosmicznego, jak i słonecznego. Badanie widma pozytonów w tym zakresie energii może dostarczyć ciekawych informacji na temat innych (niż rozpad mezonów) źródeł pozytonów.

Pochodzenie promieniowania kosmicznego

Każda teoria mająca na celu wyjaśnienie pochodzenia promieniowania kosmicznego musi uwzględniać trzy podstawowe fakty: 1) Gęstość energii promieniowania kosmicznego jest rzędu 10^{-19} J·cm^{-3}, czyli jest tego samego rzędu co inne postacie energii występujące w przestrzeni międzygwiazdowej, takie jak energia związana z promieniowaniem reliktowym, energia pola magnetycznego, czy energia kinetyczna obłoków materii międzygwiazdowej. 2) Widmo energetyczne promieniowania kosmicznego można przybliżyć wyrażeniem $E^{-\gamma}$, gdzie E oznacza energię, a wykładnik γ jest równy około 2,5. 3) Średni czas życia cząstek promieniowania kosmicznego jest rzędu 10^8 lat, po upływie tego czasu cząstki tracą znaczną część swej energii wskutek oddziaływań z rozproszoną materią, polem magnetycznym i rozrzedzonym promieniowaniem, występującymi w przestrzeniach międzygwiazdowych, a być może również — międzygalaktycznych (→ Radioastronomia). Tempo utraty energii jest tym większe, im większa jest energia cząstki. Jeżeli źródłem promieniowania kosmicznego są procesy przebiegające w obrębie Galaktyki, wówczas należy jeszcze uwzględnić straty cząstek o bardzo dużych energiach wynikające z ich „wyciekania" z obszaru Galaktyki.

Obserwowane obecnie promieniowanie kosmiczne należy więc uważać za wynik ustalenia się stanu stacjonarnego między procesami tworzenia nowych cząstek o dużych energiach i procesami, które prowadzą do zmniejszenia ich energii; jeżeli promieniowanie kosmiczne nie jest zjawiskiem chwilowym, to jego źródła muszą istnieć również obecnie. Podstawowy problem, czy źródłami tymi są obiekty galaktyczne czy pozagalaktyczne, nie został jeszcze ostatecznie rozwiązany. Za galaktycznym pochodzeniem promieniowania kosmicznego przemawia fakt, że energia konieczna do podtrzymania obserwowanego strumienia promieniowania kosmicznego może być dostarczona przez znane nam obecnie obiekty galaktyczne, w których przebiegają procesy gwałtownego wydzielania ogromnych ilości energii. Biorąc pod uwagę podaną wyżej gęstość energii promieniowania kosmicznego, jego średni czas życia, oraz objętość Galaktyki łącznie z halo galaktycznym (10^{68} cm^3), możemy ocenić energię konieczną do zachowania strumienia promieniowania kosmicznego na jego obecnym poziomie. Jest on rzędu $10^{-19}\cdot 10^{68}/3\cdot 10^{15} = 3\cdot 10^{33}$ J·s^{-1}. Energia ta może być wyzwalana podczas wybuchów supernowych (jeden wybuch supernowej może prowadzić do wydzielenia energii rzędu 10^{43} J, przy czym wybuchy następują średnio co 50 lat) oraz może być dostarczana w sposób ciągły kosztem energii rotacyjnej → Pulsarów (energie rotacji pulsarów ocenia się na 10^{45} J.

pochodzenie galaktyczne

Z drugiej jednak strony wiadomo, że niektóre obiekty pozagalaktyczne (np. radiogalaktyki lub kwazary) emitują ogromne ilości energii, której część może mieć postać promieniowania kosmicznego. Izotropowość strumienia cząstek o energiach większych od 10^{17} eV sugeruje, że cząstki te, a przynajmniej ich część, mogą być pochodzenia pozagalaktycznego. Strumień takich cząstek w przestrzeni międzygalaktycznej nie może być jednak zbyt duży, ponieważ ich oddziaływanie z protonami materii międzygalaktycznej, nawet przy bardzo małej jej gęstości, byłoby źródłem silniejszego niż się obserwuje izotropowego strumienia kwantów gamma o energiach większych od 10^8 eV. Ponadto, droga wysokoenergetycznych cząstek w przestrzeni międzygalaktycznej nie może być zbyt długa, ze względu na straty energii wynikające z oddziaływania z fotonami promieniowania reliktowego. Obecnie uważa się za bardziej prawdopodobne, że promieniowanie kosmiczne jest pochodzenia galaktycznego.

pochodzenie pozagalaktyczne

O. C. Allkofer *Introduction to Cosmic Radiation*, München 1975; K. M. V. Apparao *Composition of Cosmic Radiation*, London 1975; A. M. Hillas *Cosmic Rays*, New York 1972.

Rozpowszechnienie pierwiastków chemicznych i molekuł we Wszechświecie

Bronisław Kuchowicz

W połowie ubiegłego stulecia twórca filozofii pozytywnej A. Comte głosił, iż nauki powinny ograniczyć się do badania zjawisk bezpośrednio dostępnych obserwacji, a zatem skład chemiczny gwiazd pozostać musi poza zasięgiem ludzkiego poznania. W kilka lat po tym stwierdzeniu powstała analiza widmowa (R. W. Bunsen i G. R. Kirchhoff — 1859 r.), która jest podstawą wyznaczania składu chemicznego atmosfer gwiezdnych. W 1868 r., podczas całkowitego zaćmienia Słońca, J. Janssen i N. Lockyer zaobserwowali w widmie protuberancji słonecznych jasną linię żółtopomarańczową, niepodobną do żadnej linii znanej z laboratorium. Tak odkryto hel; pierwiastek ten dopiero po dwudziestu kilku latach wykryto w laboratorium ziemskim.

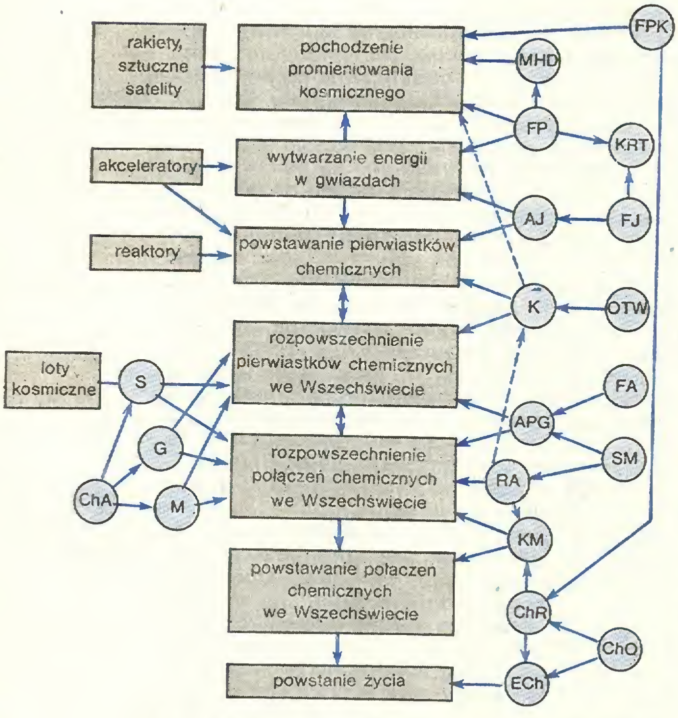
odkrycie helu

Rozwój analizy widmowej spowodował, że obecnie znacznie dokładniej znamy skład chemiczny atmosfer gwiazd oddalonych o milion lat świetlnych niż skład chemiczny wnętrza Ziemi. Metodami radioastronomii rozszyfrowano w ostatnim dziesięcioleciu skład chemiczny obłoków gazowo-pyłowych w przestrzeni kosmicznej (→ Radioastronomia). Pojęcie składu chemicznego nie jest jednoznaczne, można mieć bowiem na myśli zarówno skład pierwiastkowy jak i skład molekularny. W związku z tym w kosmochemii można wyróżnić: kosmochemię molekularną oraz kosmochemię pierwiastków i związaną z nią kosmochemię ich izotopów. Skład molekularny materii ziemskiej, meteorytowej, księżycowej wyznacza się stosując różne metody laboratoryjne chemii, natomiast obecność połączeń chemicznych w atmosferach gwiazd — wykorzystując widma molekularne znane z laboratoriów ziemskich. Identyfikację połączeń chemicznych obecnych w obłokach materii kosmicznej prowadzi się na podstawie detekcji ich linii widmowych w zakresie radiowym i mikrofalowym. Podkreślmy w tym miejscu, że kosmochemia molekularna zajmuje się kosmicznym rozpowszechnieniem wszelkich połączeń chemicznych, nie tylko molekuł, ale także rodników i jonów molekularnych. Przy znikomych gęstościach materii w przestrzeni kosmicznej (obłoki gazowo-pyłowe uważane są za bardzo gęste, jeśli w objętości 1 cm^3 znajduje się 10^5–10^7 atomów!) rodniki i jony nie różnią się w zasadzie trwałością od normalnych molekuł.

Kosmochemia zawiera także istotny element historyczny: opis procesów powstawania zasadniczych struktur chemicznych we Wszechświecie. Procesy powstawania pierwiastków chemicznych (ściślej zaś: procesy powstawania rozmaitych odmian jąder atomowych) wyjaśnia się dziś reakcjami jądrowymi przebiegającymi w sposób naturalny we Wszechświecie (czy to w rozszerzającym się Wszechświecie w pierwszych fazach jego rozwoju, czy też w gwiazdach). Analogicznie odwołujemy się do różnych reakcji chemicznych (w znacznej części do tych, którymi zajmuje się w laboratoriach chemia radiacyjna) w celu wyjaśnienia sposobu powstawania połączeń chemicznych obserwowanych w Układzie Słonecznym, w atmosferach gwiazd i obłokach gazowo-pyłowych.

rozpowszechnienie a powstawanie pierwiastków

Powiązania między problematyką rozpowszechnienia pierwiastków chemicznych i ich połączeń a powstaniem ich we Wszechświecie oraz różnymi kierunkami badań laboratoryjnych ukazuje rysunek 1.



Rys. 1. Powiązanie badań kosmochemicznych z zagadnieniami wytwarzania energii w gwiazdach i nukleosyntezy pierwiastków. W kółkach umieszczono skróty nazw dyscyplin naukowych: FPK Fizyka promieniowania kosmicznego, MHD magnetohydrodynamika, KRT kontrolowane reakcje termojądrowe (raczej krąg zagadnień, z którego może rozwinąć się odpowiednia dyscyplina techniczna), FP fizyka plazmy, FJ fizyka jądrowa, AJ astrofizyka jądrowa, K kosmologia, OTW ogólna teoria względności, APG astrofizyka powłok gwiezdnych, FA fizyka atomowa, SM spektroskopia molekularna, RA radioastronomia, KM kosmochemia molekularna, ChR chemia radiacyjna, ChO chemia organiczna, ECh ewolucja chemiczna (kompleks zagadnień z pogranicza chemii, biochemii i biologii), S selenochemia (oraz chemia innych planet, do których docierają ludzie lub próbniki, np. chemia Marsa itp.), G geochemia, M meteorytyka, ChA chemia analityczna. Z lewej strony wymieniono niektóre nowoczesne urządzenia doświadczalne i metody badań, związane z przedstawioną tematyką

Najogólniejsze dane o pierwiastkach chemicznych i ich izotopach

Zestawienie liczby znanych pierwiastków i ich izotopów wskazuje na znaczenie mechanizmów jądrowych w powstawaniu pierwiastków chemicznych (tabela 1). Pierwiastek nazywa się trwałym, gdy przynajmniej jeden z jego izotopów jest trwały; w przeciwnym razie pierwiastek nazywamy promieniotwórczym. Trwałymi są kolejne pierwiastki w układzie okresowym od wodoru do bizmutu, z wyjątkiem technetu i prometu. Istnieje 81 pierwiastków trwałych, na które przypadają 272 trwałe nuklidy (→ Jądra atomowe i ich wzbudzenia). Każdy pierwiastek trwały ma więc średnio po 3,36 trwałych izotopów. Jeśli podzielimy pierwiastki na nieparzyste (tzn. mające nieparzystą liczbę porządkową Z) i parzyste, natychmiast uwidoczni się charakterystyczne uprzywilejowanie pierwiastków parzystych. Jest ich 41, mają zaś łącznie 216 izotopów, tak więc na każdy pierwiastek parzysty przypada średnio po 5,27 trwałych izotopów. Jedynie trzy pierwiastki parzyste mają mniej niż po trzy izotopy trwałe. Grupa 40 pierwiastków nieparzystych składa się z 24 pierwiastków mających tylko jeden izotop trwały i z 16 pierwiastków mających po dwa

pierwiastki parzyste i nieparzyste

Tabela 1. Znane pierwiastki i ich izotopy

	Liczba	
	pierwiastków	nuklidów
Trwałe (od $Z = 1$ do 83, bez 43 i 61)	81	272
Promieniotwórcze występujące na Ziemi		
a) $Z \leqslant 83$ i $A < 206$	—	13*
b) $Z > 83$ i $A \geqslant 206$ (od $Z = 84$ (Po) do $Z = 92$ (U) i $Z = 94$ (Pu))	10	46
Promieniotwórcze obserwowane poza Ziemią ($Z = 43$ (Tc), 61 (Pm), 95 (Am) i 96 (Cm))	4	4**
Łącznie występujące w przyrodzie (trwałe i promieniotwórcze)	95	335
Sztuczne wytworzone (wyłącznie promieniotwórcze) — nuklidy wg zestawienia z 1968 r.	12	1192
Ogółem	107	1527

* Należą tu nuklidy długożyjące, które do dziś przetrwały w skorupie ziemskiej, np. $^{144}_{60}Nd$ o okresie połowicznego zaniku ok. $2{,}1 \cdot 10^{15}$ lat, jak również krótkożyiący tryt $^{3}_{1}H$ ($T_{1/2} = 12{,}346$ lat) i radiowęgiel $^{14}_{6}C$ ($T_{1/2} = 5730$ lat), wytwarzane stale w górnych warstwach atmosfery przez promieniowanie kosmiczne.
** Podajemy minimalną liczbę nuklidów, ponieważ nie jest dokładnie znany skład izotopowy czterech wymienionych pierwiastków w warunkach pozaziemskich.

izotopy trwałe. Żaden pierwiastek nieparzysty nie ma większej liczby trwałych izotopów; średnia w tej grupie wynosi 1,40 trwałego izotopu na pierwiastek.

uprzywilejowanie liczb parzystych

Uprzywilejowanie liczb parzystych staje się jeszcze bardziej widoczne, jeśli grupę 272 trwałych nuklidów podzielimy na cztery podgrupy zależnie od kombinacji parzystej lub nieparzystej liczby protonów Z z parzystą lub nieparzystą liczbą neutronów N. Jest aż 161 nuklidów parzysto-parzystych (tj. o parzystych liczbach Z i N), 55 parzysto-nieparzystych, 50 nieparzysto-parzystych i zaledwie 6 nieparzysto-nieparzystych. Znaczną przewagę nuklidów parzysto-parzystych (prawie 60%) powiązać można ze znanym z fizyki jądrowej faktem: najtrwalej związane są jądra o parzystej liczbie protonów i parzystej liczbie neutronów, słabiej — jądra, w których tylko jedna z tych liczb jest parzysta, najmniejszą wreszcie trwałością odznaczają się jądra nieparzysto-nieparzyste.

Skład pierwiastkowy i izotopowy obiektów we Wszechświecie

Najstarszym działem kosmochemii jest geochemia; obecnie znamy dokładnie średni skład chemiczny warstw powierzchniowych Ziemi. Analiza widmowa dostarczyła danych o obecności różnych pierwiastków chemicznych w atmosferach gwiazd i galaktyk. Detektory promieniowania jonizującego umieszczane na pokładach rakiet i sztucznych satelitów umożliwiają badanie składu chemicznego i izotopowego pierwotnego promieniowania kosmicznego i wiatru słonecznego.

Z astrofizyki obserwacyjnej wiadomo, że wszędzie, gdzie materia występuje w postaci atomów, są to atomy pierwiastków, które znamy z laboratorium na Ziemi. (Zastrzeżenie „w postaci atomów" jest konieczne z uwagi na możliwość występowania tak niezwykłych form materii, jak materia nadgęsta (jądrowa) w gwiazdach neutronowych (→ Czarne dziury i zapadanie grawitacyjne, a także czarne dziury — produkt końcowy zapadania grawitacyjnego). Wszechświat jest właściwie jednolity, a zarazem niezwykle różnorodny. Nie ma dwóch gwiazd o identycznych widmach. Przy wszystkich jednak różnicach składu chemicznego między indywidualnymi ciałami niebieskimi udało się znaleźć pewne wspólne dla nich prawidłowości rozpowszechnienia pierwiastków. Są one widoczne dopiero po nagromadzeniu dużego materiału obserwacyjnego oraz po zrozumieniu mechanizmów fizycznych związanych z powstaniem widm liniowych w atmosferach gwiazd. Występowanie linii (kształt i natężenie) widmowych określonych pierwiastków w obserwowanym przez nas widmie gwiazdy wywołane jest w pierwszym rzędzie takimi czynnikami jak temperatura i ciśnienie w obszarze, z którego promieniowanie to pochodzi. Stwierdzenie linii widmowych określonego pierwiastka nie wystarcza do oceny ilości jego atomów w atmosferze gwiazdy, ze stosunku zaś natężeń linii w widmach dwóch gwiazd o odmiennych temperaturach czy też różniących się ciśnieniem na ich powierzchni (np. biały karzeł i olbrzym) nie można bezpośrednio wnioskować o stosunku rozpowszechnienia tegoż pierwiastka w obu rozpatrywanych atmosferach.

prawidłowości rozpowszechnienia pierwiastków

skład pierwiastkowy Układu Słonecznego

Najlepiej znamy skład pierwiastkowy Układu Słonecznego: dotyczy to zwłaszcza pierwiastków o małym rozpowszechnieniu, których linii w odległych gwiazdach nie da się zauważyć. Skład pierwiastkowy materii Układu Słonecznego zwykło się uważać za typowy dla materii kosmicznej z tej części Wszechświata, która dostępna jest obserwacjom. Nasza obecna wiedza kosmochemiczna o składzie izotopowym pierwiastków chemicznych w przeważającej części jest oparta na wyznaczeniu składu izotopowego tychże pierwiastków w materii ziemskiej — metodami spektroskopii masowej. Skład izotopowy próbek materii meteorytowej, księżycowej, a w przyszłości chyba i próbek pobranych z różnych planet Układu Słonecznego nieznacznie tylko różni się od składu izotopowego próbek ziemskich. Wreszcie skład izotopowy pierwiastków chemicznych na Ziemi nie różni się w sposób istotny od składu tychże pierwiastków w atmosferach Słońca i większości gwiazd, a także w obłokach gazowo-pyłowych — przynajmniej w tych nielicznych wypadkach, w których analiza widmowa w optycznym, a nawet radiowym zakresie długości fal, pozwala na oznaczenie takiego składu.

Analiza widm atomowych jedynie w nielicznych wypadkach, np. dla pierwiastków najlżejszych, mających po dwa izotopy trwałe znacznie różniące się masą (wodór lub hel), pozwala ustalić skład izotopowy. Przesunięcie izotopowe odpowiednich linii widmowych cięższej odmiany takiego pierwiastka w stosunku do lżejszej można jednak zaobserwować w widmie gwiazdy tylko wtedy, gdy zawartości obu odmian w atmosferze gwiazdy niewiele się różnią. Weźmy np. hel; jeśli skład izotopowy tego pierwiastka w gwiazdach nie odbiega znacznie od składu jego na Ziemi, tzn. na milion atomów ^{4}He przypada mniej więcej jeden atom ^{3}He, nie ma praktycznie możliwości wykrycia tak znikomych ilości lekkiego izotopu helu ^{3}He w gwiazdach. Tak też jest istotnie dla większości gwiazd. Możliwość obserwacji ^{3}He istnieje natomiast dla pewnych gwiazd z anomaliami składu, jak gwiazda magnetyczna 21 Aquilae, którą charakteryzuje względny nadmiar (w porównaniu z helem ziemskim i słonecznym) lekkiego izotopu helu ^{3}He.

Oprócz pierwiastków najlżejszych, mierzalne wartości przesunięć izotopowych w widmach atomowych wykazują także pierwiastki ciężkie, poczynając od lantanowców. W środkowej natomiast części układu okresowego, dla pierwiastków między wapniem a lantanowcami ($20 \leqslant Z \leqslant 56$), całkowite przesunięcie izotopowe (stanowiące wypadkową zwykłego i szczególnego efektu masowego oraz efektu objętościowego, dających wkłady o różnych znakach) praktycznie równe jest zeru. Składu izotopowego powyższych pierwiastków nie dałoby się więc wyznaczyć na gruncie analizy widmowej nawet wtedy, gdyby izotopy jakiegoś pierwiastka odznaczały się zbliżonymi wartościami rozpowszechnienia.

Nie ma powodu aby przypuszczać, iż materia ziemska odznacza się jakimś specyficznym składem izotopowym. Przyjęcie, że Ziemia nie znajduje się w centralnej, specjalnie wyróżnionej pozycji, nazywa się

zasada Kopernika

w kosmologii zasadą Kopernika. Tylko mały krok dzieli tę powszechnie zaakceptowaną zasadę od stwierdzenia, że pozycja i skład chemiczny Układu Słonecznego są typowe. Skład izotopowy niektórych pierwiastków w skorupie ziemskiej z umieszczonej niżej tabeli 2 uważać więc można, zgodnie z ową zasadą, za typowy dla pierwiastków w materii kosmicznej.

Tabela 2. Skład izotopowy niektórych pierwiastków w skorupie ziemskiej

Pierwiastek	Zawartość poszczególnych izotopów (w %)
Pierwiastki lekkie	
Wodór $_1H$	$^1_1H:^2_1H$ = 99,985:0,015
Hel $_2He$	He:^4He = 0,00013:99,99987
Węgiel $_6C$	$^{12}_6C:^{13}_6C$ = 98,893:1,107
Krzem $_{14}Si$	$^{28}_{14}Si:^{29}_{14}Si:^{30}_{14}Si$ = 92,21:4,70:3,09
Siarka $_{16}S$	$^{32}_{16}S:^{33}_{16}S:^{34}_{16}S:^{36}_{16}S$ = 95,0:0,76:4,22:0,014
Pierwiastki ciężkie	
Tellur $_{52}Te$	$^{120}_{52}Te:^{122}_{52}Te:^{123}_{52}Te:^{124}_{52}Te:^{125}_{52}Te:^{126}_{52}Te:^{128}_{52}Te:$ $:^{130}_{52}Te$ = 0,089:2,46:0,87:4,61:6,99:18,71: :31,79:34,48
Antymon $_{51}Sb$	$^{121}_{51}Sb:^{123}_{51}Sb$ = 57,25:42,75
Dysproz $_{66}Dy$	$^{156}_{66}Dy:^{158}_{66}Dy:^{160}_{66}Dy:^{161}_{66}Dy:^{162}_{66}Dy:^{163}_{66}Dy:$ $:^{164}_{66}Dy$ = 0,052:0,090:2,29:18,88:25,53: :24,97,:28,18
Rtęć $_{80}Hg$	$^{196}_{80}Hg:^{198}_{80}Hg:^{199}_{80}Hg:^{200}_{80}Hg:^{201}_{80}Hg:^{202}_{80}Hg:$ $:^{204}_{80}Hg$ = 0,146:10,02:16,84:23,13:13,22: 29,80:6,85

Istnieją pewne grupy obiektów kosmicznych, w których skład pierwiastkowy i skład izotopowy pierwiastków różnią się istotnie od odpowiednich wartości dla materii ziemskiej; różnice te można wyjaśnić na gruncie teorii ewolucji gwiazd i związanych z nią procesów nukleosyntezy pierwiastków chemicznych. Nie wnikając głębiej w to zagadnienie podajmy dla przykładu dane odnoszące się do składu izotopowego rtęci w dwu gwiazdach osobliwych typu *A*.

skład izotopowy rtęci

Z analizy widma gwiazdy o numerze katalogowym HR 5883 wynika, że obecne są w niej tylko dwa najcięższe izotopy rtęci: ^{202}Hg (3%) i ^{204}Hg (97%). Pięć lżejszych izotopów rtęci, stanowiących ilościowo prawie $^2/_3$ rtęci ziemskiej (patrz najniższy wiersz w tabeli 2), w gwieździe owej praktycznie całkiem nie ma. A oto oceniony przez G.W. Prestona (1971 r.) skład izotopowy rtęci w innej gwieździe osobliwej o numerze katalogowym HR 4072: ^{200}Hg i ^{201}Hg — łącznie ok. 4%, ^{202}Hg — 37%, ^{204}Hg — 59%. I tu praktycznie nie ma wcale trzech najlżejszych izotopów rtęci.

W ostatnim dziesięcioleciu wyznaczono metodami radioastronomii skład izotopowy wielu pierwiastków dość rozpowszechnionych i wchodzących w skład obserwowanych molekuł. Przesunięcie izotopowe w widmach molekularnych (w zakresie radiowym i mikrofalowym) jest czasami znacznie łatwiej stwierdzić niż w widmach atomowych. Szczególnie dobrze nadają się do tego celu proste molekuły dwuatomowe i rodniki (jak OH, CO, CS). Tą drogą udało się oznaczyć skład izotopowy węgla, tlenu, siarki i wodoru w obłokach materii międzygwiazdowej; skład ten niewiele różni się od składu ziemskiego. W analogiczny sposób wykorzystuje się pasma molekularne w widmach chłodniejszych gwiazd (typy widmowe *G8* i późniejsze), w których oznaczać można skład izotopowy węgla lub tlenu.

skład izotopowy C, O, S, H w materii międzygwiazdowej

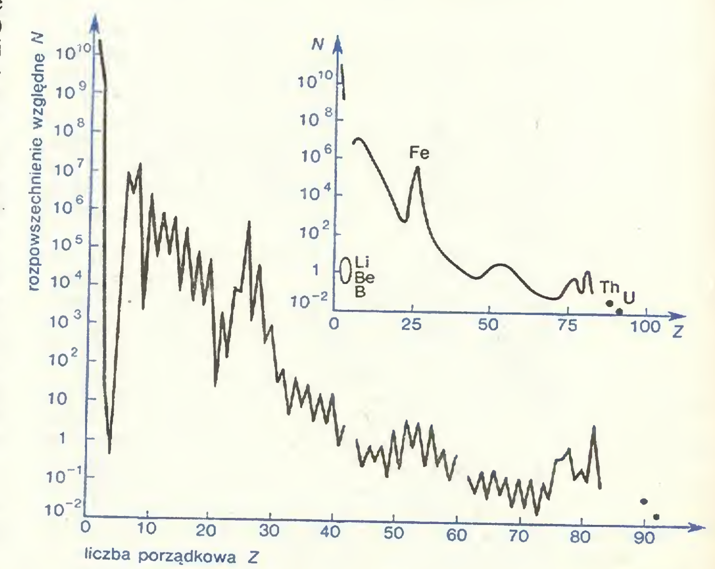
Skład izotopowy poszczególnych pierwiastków w materii ziemskiej niewiele różni się od składu materii Układu Słonecznego, natomiast nie można tego powiedzieć o składzie pierwiastkowym. Rozkład temperatur i ciśnień w obłoku gazowo-pyłowym, z którego powstały (drogą kondensacji, nukleacji i wzrostu przez akrecję) ciała Układu Słonecznego nie był równomierny. Spadek temperatur i ciśnień w miarę oddalenia się od ciała centralnego wywołał zróżnicowanie materii mgławicy pierwotnej, spowodowane różną lotnością pierwiastków, co szczególnie wpłynęło na lokalne przebiegi procesu kondensacji. W wyniku działania tych czynników planety bliższe Słońca (planety grupy ziemskiej) — Merkury, Wenus, Ziemia i Mars — są ubogie w pierwiastki najbardziej lotne — wodór i hel, które dominują w materii Jowisza, bardziej oddalonego od Słońca. Nierównomierność rozpowszechnienia pierwiastków w Układzie Słonecznym odnosi się nie tylko do pierwiastków o różnej lotności, ale i frakcjonowania metali i krzemianów, które m.in. doprowadziło do utworzenia jąder metalicznych planet grupy ziemskiej.

Uniwersalne krzywe rozpowszechnienia i pewne ogólne prawidłowości

Od kilkudziesięciu lat trwają próby uogólnienia danych o rozpowszechnieniu pierwiastków i ich izotopów w materii ziemskiej, meteorytowej, słonecznej, atmosferach gwiazd i promieniowaniu kosmicznym. Informacje uzyskane z różnych źródeł okazały się dość zbliżone i pozwoliły na ustalenie średniego rozpowszechnienia pierwiastków (i ich izotopów) w materii naszej Galaktyki. Największa liczba najdokładniejszych danych odnosi się do materii Układu Słonecznego, stąd też otrzymane krzywe średniego rozpowszechnienia: pierwiastków (rys. 2) i izobarów (rys. 3) nazywa się niekiedy krzywymi rozpowszechnienia dla Układu Słonecznego, choć stosowana jest i ogólna nazwa uniwersalnych krzywych rozpowszechnienia. Tę drugą nazwę będziemy dalej stosować w celu podkreślenia zasadniczej różnicy między tymi uśrednionymi krzywymi a konkretną krzywą rozpowszechnienia odnoszącą się do określonego ciała niebieskiego.

rozpowszechnienie $N(X)$ pierwiastka

Obie krzywe empiryczne przedstawione są przy użyciu skali logarytmicznej dla osi rzędnych, różnica bowiem między najwyższą i najniższą wartością rozpowszechnienia sięga niemal dwunastu rzędów wielkości. Rozpowszechnieniem $N(X)$ jakiegoś pierwiastka X nazywa się liczbę atomów tego pierwiastka przypadającą na milion atomów krzemu w danym ośrodku. Tak więc u podstaw skali rozpowszechnienia

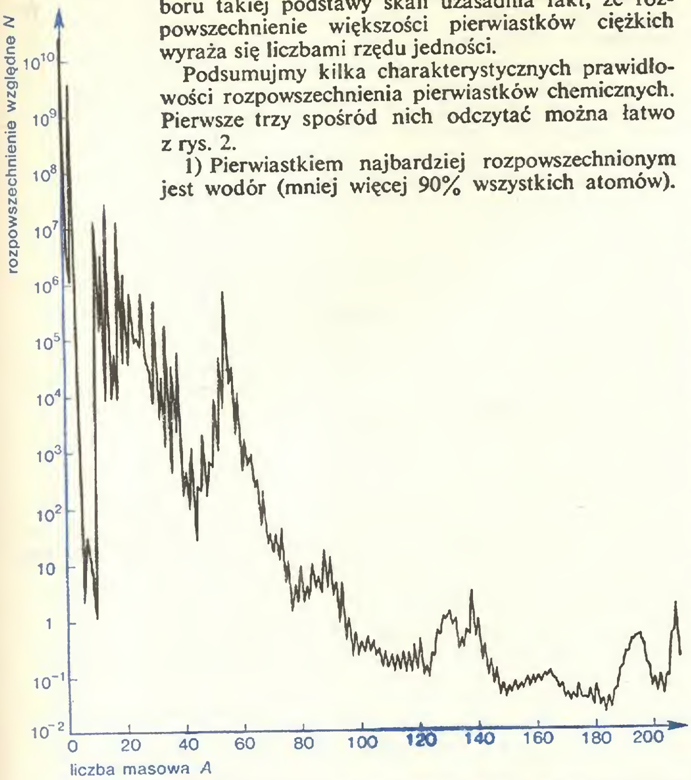


Rys. 2. Krzywa rozpowszechnienia pierwiastków chemicznych wg Camerona (1973); w górze na prawo — schematyczny kontur krzywej z zaznaczeniem lokalnych maksimów i minimum Li–Be–B

tkwi umowne przyjęcie $N(\mathrm{Si}) = 10^6$; dogodność wyboru takiej podstawy skali uzasadnia fakt, że rozpowszechnienie większości pierwiastków ciężkich wyraża się liczbami rzędu jedności.

Podsumujmy kilka charakterystycznych prawidłowości rozpowszechnienia pierwiastków chemicznych. Pierwsze trzy spośród nich odczytać można łatwo z rys. 2.

1) Pierwiastkiem najbardziej rozpowszechnionym jest wodór (mniej więcej 90% wszystkich atomów).



Rys. 3. Krzywa rozpowszechnienia izobarów wg Camerona (1973)

hel i wodór

Drugie miejsce zajmuje hel (o rząd mniej niż wodoru). Reszta pierwiastków stanowi łącznie mniej niż 1% całkowitej liczby atomów, a prawie 90% tej reszty, to atomy czterech lekkich pierwiastków: węgla C ($Z = 6$), tlenu O ($Z = 8$), azotu N ($Z = 7$) i neonu Ne ($Z = 10$).

pierwiastki parzyste

2) Począwszy od węgla, pierwiastki parzyste mają na ogół większe rozpowszechnienie niż sąsiadujące z nimi pierwiastki nieparzyste. Różnica ta stopniowo maleje ze wzrostem liczby atomowej Z.

3) Rozpowszechnienie pierwiastków (jeśli pominąć wspomniany przed chwilą efekt parzystości Z) maleje na ogół regularnie ze wzrostem liczby atomowej Z.

4) Lokalne nieregularności na krzywej rozpowszechnienia pojawiają się w kilku miejscach: a) Charakterystyczna nieciągłość występuje przy gwałtownym spadku rozpowszechnienia (o ok. 7 rzędów wielkości) dla trzech pierwiastków lekkich: litu, berylu i boru. Po przejściu przez tę nieciągłość krzywa rozpowszechnienia biegnie tak, jak gdyby była przedłużeniem odcinka łączącego wartości rozpowszechnienia dla wodoru i helu. b) Dalej na opadającej krzywej rozpowszechnienia pojawiają się maksima lokalne: pewne pierwiastki czy też ich grupy odznaczają się rozpowszechnieniem znacznie większym niż sąsiednie.

maksima krzywej rozpowszechnienia

Pierwsze z tych maksimów oznaczone symbolem Fe związane jest z grupą żelazowców. Występowanie jego wiąże się ze znanym z fizyki jądrowej faktem, iż jądra pierwiastków z otoczenia żelaza mają największą wartość energii wiązania przypadającą na pojedynczy nukleon, są więc najtrwalsze. Dalsze maksima lokalne wiążą się z magicznymi liczbami neutronów i odczytać je można wyraźniej z krzywej rozpowszechnienia izobarów, która zawiera więcej szczegółów niż krzywa rozpowszechniania pierwiastków.

izobary o parzystym *A*

Dalsze prawidłowości empiryczne odczytać można z krzywej rozpowszechnienia izobarów.

5) Izobary o parzystym A są na ogół bardziej rozpowszechnione niż sąsiednie izobary o nieparzystym A (analogicznie jak w punkcie 2).

jądra o *A*:4

6) Wśród jąder lekkich szczególnie dużym rozpowszechnieniem odznaczają się te jądra, których liczba masowa A jest wielokrotnością czwórki (np. ^{4}He, ^{12}C, ^{16}O, ^{20}Ne, ^{24}Mg, ^{28}Si, ^{32}S). Ostatnia grupa prawidłowości empirycznych jest związana ze składem izotopowym pierwiastków mających więcej niż jeden izotop trwały. Wymienimy tu jedynie dwie najważniejsze prawidłowości.

pierwiastki o $Z \leqslant 30$

7) Wśród pierwiastków lekkich (pierwiastki o $Z \leqslant 30$, tj. od wodoru do cynku włącznie) najbardziej rozpowszechnione są izotopy lżejsze; wyjątkami są tu hel He ($Z = 2$), lit Li ($Z = 3$), bor B ($Z = 5$), argon Ar ($Z = 18$) i wanad V ($Z = 23$).

pierwiastki o $Z > 30$

8) Pierwiastki ciężkie (o liczbie atomowej $Z > 30$, tj. $^2/_3$ układu okresowego, nie licząc odległych transuranowców, w przyrodzie nie obserwowanych) charakteryzuje przewaga izotopów cięższych, przy czym przewaga ta jest tym wyraźniejsza, im więcej trwałych izotopów ma dany pierwiastek. Dla pierwiastków o dwóch izotopach trwałych możliwe są niewielkie odstępstwa od tej reguły (patrz np. antymon Sb w tabeli 2).

Skład izotopowy przykładowo wybranych pierwiastków w tabeli 2 ilustruje dość dobrze dwie ostatnie prawidłowości. Można jeszcze wymienić jedną prawidłowość empiryczną, mniej wyraziście uwypukloną (istnieje bowiem aż 6 wyjątków).

9) Najlżejszy izotop parzystych pierwiastków ciężkich jest z reguły izotopem o najniższym rozpowszechnieniu, a zawartość jego w składzie danego pierwiastka jest zwykle rzędu 0,1%, niekiedy tylko dochodzi do ok. 2%. Wyjątkami są tu: german Ge ($Z = 32$), cyrkon Zr ($Z = 40$), molibden Mo ($Z = 42$), ruten Ru ($Z = 44$), neodym Nd ($Z = 60$) i samar Sm ($Z = 62$).

Prawidłowości rozpowszechnienia nuklidów a fizyka jądrowa

brak okresowych prawidłowości

Informacje podane w poprzednim punkcie oraz brak jakiejkolwiek okresowości w uniwersalnej krzywej rozpowszechnienia z rys. 2 (a więc brak związku między rozpowszechnieniem pierwiastków chemicznych we Wszechświecie a ich właściwościami chemicznymi) wskazują, że rozpowszechnienia pierwiastków niepodobna powiązać w sposób przyczynowy z chemią. Już od czasów Mendelejewa usiłowano bezskutecznie znaleźć okresowe prawidłowości w rozpowszechnieniu pierwiastków chemicznych. Pierwiastki o tak podobnych właściwościach chemicznych jak metale alkaliczne różnią się dość istotnie rozpowszechnieniem (stosunek rozpowszechnienia w skorupie ziemskiej jest: Li:Na:K:Rb:Cs = 470:125 000:67 100:137:3,8), natomiast pierwiastki o odmiennych właściwościach chemicznych mają zbliżone wartości rozpowszechnienia (np. litu jest mniej więcej tyle co chloru). Nie dziwmy się, że tak jest. Wszak chemia nie opisuje powstania pierwiastków chemicznych, jej przedmiotem badań są właściwości tych pierwiastków oraz powstawanie i właściwości ich związków. Powstanie pierwiastków — to już właściwie metachemia.

Natomiast fizyka jądrowa, dziedzina zajmująca się przemianami promieniotwórczymi jednych pierwiastków w inne oraz reakcjami jądrowymi prowadzącymi do przekształcania się jąder atomowych, wskazuje nam od ponad pół wieku na mechanizm tworzenia się pierwiastków chemicznych. To, co realizujemy dziś w laboratoriach jądrowych, co jest produktem ubocznym pracy reaktora jądrowego i eksplozji bomb atomowych i wodorowych, może równie dobrze przebiegać w warunkach naturalnych, w przyrodzie. Reakcje jądrowe zachodzą bowiem we wnętrzach gwiazd, dostarczając im nieustannie energii, a jednocześnie zmienia się skład chemiczny materii we wnętrzu gwiazdy, tworzą się nowe pierwiastki.

Krzywe z rys. 2 i 3 są pewnym średnim wzorcem, z którym porównywać można rozpowszechnienie

pierwiastków i izobarów w poszczególnych obiektach. Występujące indywidualne odchylenia można wytłumaczyć odwołując się do przebiegających w danych obiektach procesów jądrowych i do ewolucji tych obiektów. Uniwersalne krzywe rozpowszechnienia nazywać można bowiem krzywymi historycznymi — w tym znaczeniu, że zapisany jest w nich wynik procesów ewolucji materii w naszej Galaktyce. Można sobie wyobrazić, iż analogiczne krzywe dla materii z jakiegoś odległego, na razie bezpośrednio naszemu wglądowi niedostępnego zakątka Wszechświata przebiegać będą nieco odmiennie. Jeśli jednak przyjąć, że wszędzie obowiązują te same prawa fizyki jądrowej (co chyba nie budzi większych wątpliwości), to różnice przebiegu mogą być tylko niewielkie (szybszy lub wolniejszy spadek, wyższe lub niższe maksima), a maksima i minima lokalne wystąpić muszą w tych samych miejscach.

Kosmochemia molekularna

połączenia chemiczne w kosmosie

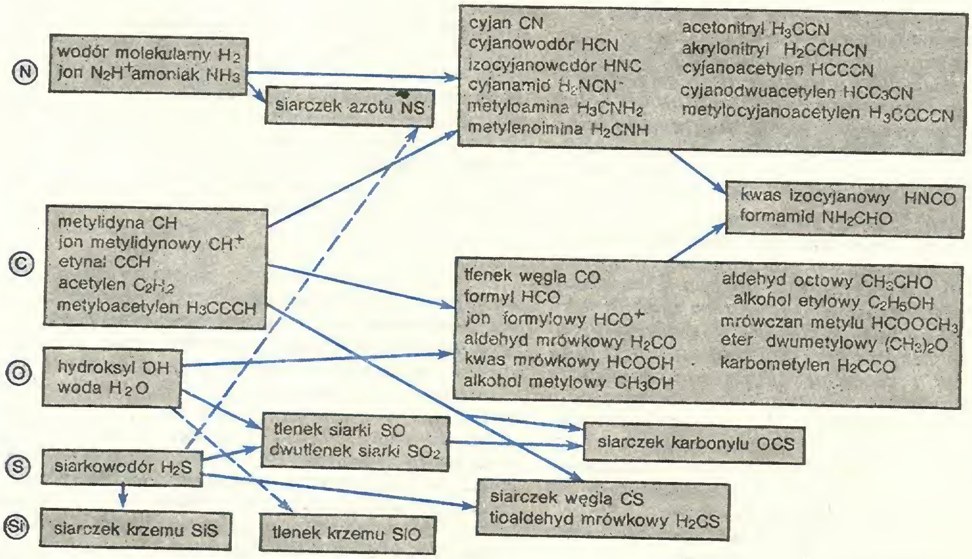
Do niedawna niewiele wiedziano o występowaniu połączeń chemicznych w kosmosie. Badania laboratoryjne nad składem molekularnym obejmowały meteoryty jako jedyne próbki materii pochodzenia pozaziemskiego. W atmosferach chłodniejszych gwiazd stwierdzono obecność prostych cząsteczek i rodników kilkoatomowych, zdolnych oprzeć się rozerwaniu w wysokiej temperaturze; były to tlenki metali oraz połączenia zawierające tak rozpowszechnione pierwiastki jak wodór czy węgiel (np. C_2, CN, CH, AlH itp.). Na podstawie widm optycznych gwiazd stwierdzono już przed kilkudziesięciu laty, że w przestrzeni międzygwiazdowej występują połączenia CH, CH^+ i CN. W 1963 r. wykryto obecność rodnika hydroksylowego OH w przestrzeni międzygwiazdowej. Wielka seria odkryć, które doprowadziły do gruntownego przekształcenia naszych wyobrażeń o chemii przestrzeni międzygwiazdowej, zaczęła się w 1968 r. odkryciem amoniaku i wody w obłokach materii gazowo-pyłowej. Odtąd, głównie dzięki wykorzystaniu radioteleskopów do obserwacji widmowych linii emisyjnych i absorpcyjnych pochodzących od molekuł, rodników i jonów, nieustannie wzrasta liczba odkrywanych połączeń. Krótkie ich zestawienie z początku 1977 r. zawiera rys. 4. Łatwo dostrzec, iż większość połączeń na nim występujących zaliczyć można do połączeń organicznych. Nie ma w tym nic dziwnego, jeśli spojrzeć na krzywe rozpowszechnienia z rys. 2.

połączenia organiczne

Obfitość występowania w przestrzeni kosmicznej zasadniczych substratów potrzebnych do tworzenia złożonych molekuł nie może się przecież różnić w sposób istotny od wartości rozpowszechnień z krzywej uniwersalnej. Najbardziej rozpowszechnionymi pierwiastkami są kolejno: wodór, hel, tlen i węgiel (niektórzy zmieniają kolejność tych dwóch pierwiastków), azot i neon. Wszystkie wymienione pierwiastki, z wyjątkiem gazów szlachetnych, występują w odkrytych molekułach. W dalszej kolejności według rozpowszechnienia wymienić należy: magnez, krzem, żelazo, siarkę, glin i wapń. Wśród połączeń chemicznych, odkrytych w przestrzeni międzygwiazdowej, występują proste molekuły zawierające w swym składzie atom krzemu lub siarki. Nieobecność połączeń zawierających atomy metali (magnezu, żelaza, glinu i wapnia) wydaje się tylko pozorna, spowodowana wchodzeniem tych pierwiastków do składowej pyłowej ośrodka międzygwiazdowego, do której badania metody radioastronomii niewiele się przydają. Wszystkie natomiast zaobserwowane w przestrzeni międzygwiazdowej połączenia chemiczne wchodzą do składowej gazowej.

połączenia 9-atomowe

Obecnie znamy już dwie molekuły dziewięcioatomowe. Przypuszcza się, że jest tylko kwestią czasu i konstrukcji odpowiednio czułej aparatury, by zaobserwować linie radiowe tak złożonych połączeń organicznych, jak np. aminokwasy. Przemawiają za tym wypadki znalezienia przeróżnych aminokwasów w materii meteorytowej. Do niedawna podejrzewano, że próbki meteorytowe były zanieczyszczone substancjami organicznymi pochodzenia ziemskiego. Zastrzeżenia te rozwiało chromatograficzne rozdzielenie form lewo- i prawoskrętnych poszczególnych aminokwasów. Okazało się, że stosunek ich zbliżony jest do jedności, co przemawia na rzecz syntezy abiogennej (nie przebiegającej w organizmach żywych). Powstawanie złożonych substancji w ośrodku międzygwiazdowym wyjaśnić można w sposób naturalny jako rezultat reakcji chemicznych typu jon-molekuła między produktami oddziaływania wysokoenergetycznego promieniowania kosmicznego z materią międzygwiazdową, mającą skład pierwiastkowy zgodny z krzywą uniwersalną. W lokalnych zagęszczeniach owej materii, zawierających masy rzędu kilku do kilkuset mas słonecznych, przebiega ciąg reakcji znanych z chemii radiacyjnej; nieustannie tworzą się i ulegają zniszczeniu złożone molekuły i rodniki. Przez dzie-



Rys. 4. Molekuły, rodniki i jony wykryte w ośrodku międzygwiazdowym — ułożone według pierwiastków ciężkich i wzrastającej złożoności. Najprostsze połączenia (z wodorem) podane są z lewej strony, obok zaś — związki zawierające co najmniej dwa atomy różnych pierwiastków cięższych od wodoru

sięciolecia panowało przekonanie o niemożności utworzenia się złożonych molekuł w przestrzeni kosmicznej. Ledwo molekuła taka zdążyłaby powstać, już jej dalszemu bytowi miałoby zagrażać wszechobecne promieniowanie ultrafioletowe. Gdy obecnie odkrycia wielu molekuł zadały kłam temu przekonaniu, sądzi się, iż istotną rolę odgrywa osłanianie wewnętrznej części gęstego obłoku przez warstwę powierzchniową, pochłaniającą promieniowanie ultrafioletowe. Jedynie wysokoenergetyczne promieniowanie kosmiczne wnikać może do środka. Dominująca część wodoru w centralnych obszarach gęstych obłoków występuje zatem jako wodór molekularny H_2, a nie atomowy, i dzięki temu wiele reakcji w fazie gazowej staje się możliwymi. Obłoki gazowo-pyłowe uważane są za gęste, gdy 1 cm^3 zawiera od 10^4 do 10^7 atomów. Kilka przybliżonych stężeń obserwowanych molekuł w różnych obłokach zawiera tabela 3.

Obserwacje radioastronomiczne pozwalają na otrzymanie informacji o składzie molekularnym materii obłoków gazowo-pyłowych także innych galaktyk. Łatwa obserwowalność przesunięć izotopowych w zakresie radiowym umożliwia wyznaczenie stosunków izotopowych. Tak więc np. dla węgla, który ma dwa izotopy trwałe, stosunek ich rozpowszechnienia w materii ziemskiej wynosi: $^{13}C/^{12}C = {}^{1}/_{89}$. Górna granica tego stosunku dla kilku radioźródeł wynosi: centrum galaktyczne $^{1}/_{25}$, obszar ξ Oph — $^{1}/_{80}$, obłok w Orionie — $^{1}/_{9}$. Podane wartości nie muszą wskazywać na sprzeczność z danymi ziemskimi; są to przecież tylko górne granice,

Tabela 3. Ocena stężenia niektórych molekuł w Galaktyce

Molekuła lub rodnik	Obszar	Gęstość liczba molekuł w cm^3
Wodór H_2	obłoki gęste	10^4
	obłoki bardzo gęste	10^6
Hydroksyl:		
odmiana pospolita ^{16}OH	Sgr *A*	10^{-4}
odmiana rzadsza ^{18}Oh	Sgr *A*	10^{-6}
Amoniak NH_3	Sgr *A*	10^{-8}
Tlenek węgla CO	Sgr *A*	10^{-1}
Alkohol metylowy CH_3OH	Orion	0,25

które dzięki dalszym pomiarom mogą ulec obniżeniu, zbliżając się do wartości znanej na Ziemi.

Zwróćmy tu uwagę na aspekt biologiczny odkryć radioastronomicznych. W przestrzeni kosmicznej powstają samorzutnie w dużych ilościach złożone substancje chemiczne, w tym przynajmniej niektóre spośród związków chemicznych podstawowych dla życia organicznego. Stanowi to poparcie hipotezy powstawania życia w samorzutny sposób w różnych miejscach Wszechświata.

B. Kuchowicz *Kosmochemia*, Warszawa 1979; B. Kuchowicz *Problemy i osiągnięcia astrofizyki jądrowej. Część I. Rozpowszechnienie nuklidów i ich kosmiczna synteza*, Post. Fiz. 22, 495 (1971); A. K. Ławruchina, G. M. Kolesow *Powstawanie pierwiastków chemicznych we Wszechświecie*, Warszawa 1965; Paul W. Merrill *Chemia Kosmosu*, Warszawa 1966.

Reakcje jądrowe w gwiazdach

Bronisław Kuchowicz

aspekt astrofizyczny i kosmochemiczny

Reakcje jądrowe w gwiazdach mają dwa aspekty: astrofizyczny i kosmochemiczny. Wytwarzanie energii w gwiazdach, czego przejawem jest ich świecenie, ma aspekt astrofizyczny, natomiast przekształcanie jednych pierwiastków w drugie, nukleosynteza, ma aspekt kosmochemiczny. Powyższe aspekty są od siebie nieodłączne, albowiem reakcje jądrowe wytwarzające energię prowadzą do nieustannej zmiany składu chemicznego materii wnętrza gwiazd; w wyniku tego w wytwarzaniu energii odgrywają rolę coraz to inne reakcje jądrowe, zmienia się przy tym szybkość wytwarzania energii, zatem i jasność gwiazd. A więc ewolucja gwiazdy jest jednocześnie ewolucją energetyczną i ewolucją składu chemicznego i jest uwarunkowana reakcjami jądrowymi (→ Ewolucja gwiazd, Reakcje jądrowe, Rozpady jąder atomowych).

Warunki realizacji reakcji jądrowych w gwiazdach

Zależność średniej energii wiązania nukleonu w jądrze od liczby masowej jądra wykazuje maksimum w otoczeniu $A = 60$ (w otoczeniu jądra żelaza). Możliwe są zatem dwa kierunki przemian jądrowych prowadzących od jąder słabiej do jąder silniej związanych — reakcje termojądrowe (strzałka *A* na rys. 1) oraz reakcje rozszczepienia jąder ciężkich (strzałka *B*). Oba rodzaje przemian jądrowych zachodzą z wydzieleniem energii (reakcje egzoenergetyczne), oba udaje się realizować w warunkach ziemskich (→ Energia termojądrowa, Energia jądrowa). Natomiast w warunkach, które panują w gwiazdach, występować mogą tylko reakcje termojądrowe i one właśnie są źródłem energii gwiazd.

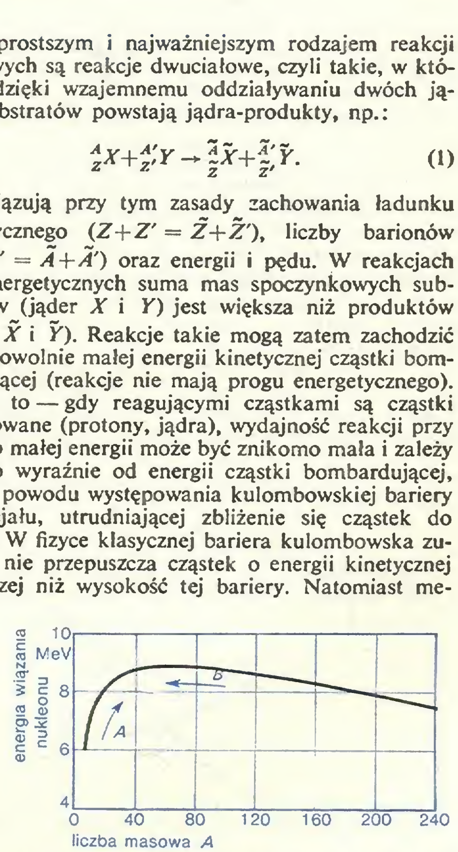
reakcje termojądrowe — źródło energii gwiazd

reakcje dwuciałowe

Najprostszym i najważniejszym rodzajem reakcji jądrowych są reakcje dwuciałowe, czyli takie, w których dzięki wzajemnemu oddziaływaniu dwóch jąder-substratów powstają jądra-produkty, np.:

$$^{A}_{Z}X + ^{A'}_{Z'}Y \to ^{\tilde{A}}_{\tilde{Z}}\tilde{X} + ^{\tilde{A}'}_{\tilde{Z}'}\tilde{Y}. \quad (1)$$

Obowiązują przy tym zasady zachowania ładunku elektrycznego $(Z+Z' = \tilde{Z}+\tilde{Z}')$, liczby barionów $(A+A' = \tilde{A}+\tilde{A}')$ oraz energii i pędu. W reakcjach egzoenergetycznych suma mas spoczynkowych substratów (jąder X i Y) jest większa niż produktów (jąder $\tilde{X}$ i $\tilde{Y}$). Reakcje takie mogą zatem zachodzić przy dowolnie małej energii kinetycznej cząstki bombardującej (reakcje nie mają progu energetycznego). Mimo to — gdy reagującymi cząstkami są cząstki naładowane (protony, jądra), wydajność reakcji przy bardzo małej energii może być znikomo mała i zależy bardzo wyraźnie od energii cząstki bombardującej, a to z powodu występowania kulombowskiej bariery potencjału, utrudniającej zbliżenie się cząstek do siebie. W fizyce klasycznej bariera kulombowska zupełnie nie przepuszcza cząstek o energii kinetycznej mniejszej niż wysokość tej bariery. Natomiast me-



Rys. 1. Schematyczny przebieg zależności średniej energii wiązania nukleonu w jądrze *E* od liczby masowej *A*. Strzałki wskazują dopuszczalne kierunki reakcji egzoenergetycznych

chanika kwantowa (tkwiąca u podstaw współczesnej fizyki jądrowej) przewiduje zjawisko tunelowe — możliwość przeniknięcia cząstki przez barierę potencjału nawet wtedy, gdy energia cząstki jest znacznie niższa od wysokości bariery. Prawdopodobieństwo pojedynczego aktu przeniknięcia cząstki przez barierę silnie maleje ze wzrostem ładunku oraz ze spadkiem energii kinetycznej lub prędkości tej cząstki. Przy prędkości bliskiej zera przenikalność bariery też spada do zera.

kolumbowska bariera potencjału

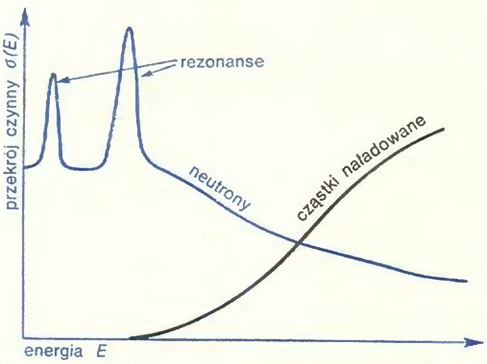
Choć prawdopodobieństwo przeniknięcia przez barierę — gdy energia cząstki padającej jest znacznie mniejsza niż wysokość bariery — jest znikome, nie jest ono nigdy dokładnie równe zeru. Jest to istotne dla procesów jądrowych odbywających się wewnątrz gwiazd. Wpływ występowania bariery kulombowskiej na czas trwania ewolucji gwiazd zilustrujemy następującym przykładem. Przeciętna gwiazda zawiera co najmniej 10^{50} jąder atomowych. Jeśli przyjąć ostrożnie, że w obszarze centralnym, w których zachodzi wydzielanie energii, znajduje się zaledwie 0,1% masy gwiazdy, a prawdopodobieństwo zajścia reakcji między dwoma jądrami wynosi zaledwie 10^{-10} na rok, to w ciągu roku zachodzi 10^{37} reakcji. Przy takim tempie „spalania" paliwo jądrowe może wystarczyć na miliardy lat!

Oczywiście dla cząstek pozbawionych ładunku, jak neutrony czy kwanty γ, bariera potencjału nie występuje. Neutrony mogą się przedostawać do jądra atomowego przy dowolnie małej energii prawie z takim samym prawdopodobieństwem jak neutrony o wielkiej energii. Dlatego wydajność reakcji wywoływanych przez neutrony przewyższa o kilka rzędów wielkości wydajność reakcji wywoływanych przez protony lub cząstki α o tej samej małej energii, nie mówiąc już o reakcjach powodowanych przez jądra cięższe, które mają wyższą barierę kulombowską.

Bariera potencjału jest barierą dwustronną, działa nie tylko na cząstki wnikające do jądra, ale i na te, które usiłują się z niego wydostać (np. w przemianie α). Jądro powstające w wyniku przyłączenia neutronu do jądra początkowego z większym prawdopodobieństwem pozbędzie się nadmiaru energii (równego energii wiązania owego neutronu) przez wysłanie kwantu γ niż cząstki naładowanej. A zatem spośród reakcji wywoływanych przez neutrony największą rolę w gwiazdach odgrywają reakcje (n, γ). Reakcje jądrowe mogą przebiegać także pod działaniem kwantów γ o energii dostatecznie dużej na to, by z jądra atomowego wybić jeden lub kilka składników. Dwustronność bariery kulombowskiej sprawia, że jeżeli energia kwantu γ jest mniejsza od wysokości bariery, to z największą wydajnością przebiega reakcja fotojądrowa typu (γ, n).

przekroje czynne

Przekroje czynne na różne reakcje jądrowe zależą od rodzaju i energii cząstek oddziałujących; w reakcjach endoenergetycznych równe są zawsze zeru, gdy energia jest mniejsza od progowej. Przekroje czynne na reakcje bombardowania jąder atomowych cząstkami naładowanymi, nawet o energii kinetycznej większej niż wysokość bariery kulombowskiej, wynoszą zazwyczaj najwyżej dziesiętne części barna (1 b = $= 10^{-28}$ m²). Natomiast największymi przekrojami czynnymi, niekiedy o wartościach rzędu kilkunastu i więcej barnów, odznaczają się reakcje wychwytu radiacyjnego neutronów powolnych (n, γ). Podczas gdy przekroje czynne na reakcje między cząstkami naładowanymi rosną zazwyczaj wyraźnie z energią, przekroje czynne na egzoenergetyczne reakcje (n, γ) na ogół maleją ze wzrostem energii (rys. 2). Neutron szybszy przebywa bowiem krócej w sąsiedztwie jądra niż neutron powolny, stąd mniejsze są szanse na jego wychwyt w wyniku działania sił jądrowych.



Rys. 2. Schematyczny przebieg przekroju czynnego przy małych energiach dla reakcji egzoenergetycznych wywołanych przez cząstki naładowane

występowanie rezonansu

Przekrój czynny na reakcję (1) może wzrosnąć o jeden rząd wielkości, a nawet i więcej (rezonans — rys. 2), jeśli energia cząstek oddziałujących jest tak dobrana, że w etapie pośrednim może powstać jądro złożone V w jednym ze swoich stanów wzbudzonych:

$$^{A}_{Z}X + ^{A'}_{Z'}Y \rightarrow ^{A+A'}_{Z+Z'}V^*. \quad (2)$$

W przebiegającym w gwiazdach procesie spalania helu na węgiel właśnie występowanie rezonansu przy energii 7,65 MeV jest przyczyną wydajnego tworzenia się węgla (→ Powstawanie pierwiastków chemicznych).

Zwiększenie energii wiązania w jądrze o magicznej liczbie neutronów N przejawia się zmniejszeniem przekroju czynnego na wychwyt neutronu. Przekrój czynny jąder o $N = 50, 82, 126$ na wychwyt neutronu o energii rzędu 1 MeV jest 10–100 razy mniejszy niż jąder sąsiednich. Natomiast jądra o liczbie neutronów większej o jedność od liczby magicznej mają szczególnie duży przekrój czynny na reakcję fotojądrową (γ, n).

Reakcje jądrowe między cząstkami naładowanymi

Pod względem wytwarzania energii najważniejszą rolę odgrywają w gwiazdach reakcje dwuciałowe, typu reakcji (1), między cząstkami naładowanymi — jądrami atomowymi. Liczba aktów reakcji zachodzących w jednostce objętości na jednostkę czasu jest tym większa, im większe są stężenia oddziałujących ze sobą jąder oraz im wyższa jest temperatura. Wyższa temperatura oznacza bowiem większą wartość średniej energii kinetycznej oddziałujących jąder, a zatem większe prawdopodobieństwo przeniknięcia przez barierę kulombowską.

prawdopodobieństwo reakcji

Uśrednione prawdopodobieństwo $\langle\sigma, v\rangle$ reakcji (1), odniesione do jednej pary cząstek reagujących na jednostkę czasu w zakresie energii, w którym nie występują rezonanse, można z dobrym przybliżeniem przedstawić w postaci całki iloczynu dwóch funkcji:

$$\langle\sigma, v\rangle = C_1 T^{-2/3} \int_0^\infty e^{-\sqrt{E_G/E}}\, e^{-E/kT}\, dE. \quad (3)$$

Funkcja $e^{-E/kT}$ przedstawia (z dokładnością do współczynnika) rozkład maxwellowski energii (prędkości) względnej, funkcja zaś $e^{-\sqrt{E_G/E}}$ — prawdopodobieństwo przeniknięcia przez barierę potencjału cząstki o energii E. Wartość energii E_G jest tzw. energią Gamowa; określa ją wzór

energia Gamowa

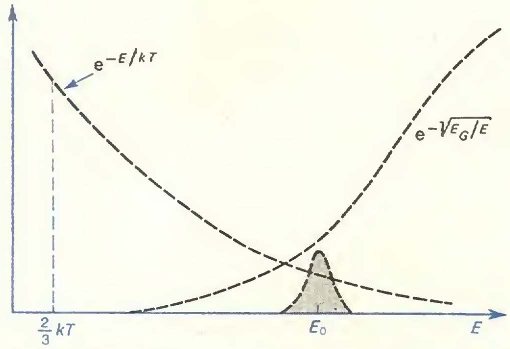
$$E_G = 2\mu(\pi ZZ'e^2/\hbar)^2, \quad (4)$$

gdzie μ — masa zredukowana oddziałujących jąder X i Y,

$$\mu = M_X \cdot M_Y/(M_X + M_Y). \quad (5)$$

(W powyższych wzorach C_1 — współczynnik niezależny od temperatury, T — temperatura bezwzględna, k — stała Boltzmanna, e — ładunek elementarny,

$\hbar$ — stała Plancka podzielona przez 2π). Przebieg obu czynników występujących pod znakiem całki przedstawiony jest na rys. 3. Iloczyn ich obu, tzn. funkcja podcałkowa, osiąga maksymalną wartość przy energii $E_0 = (^1/_2 kT\sqrt{E_G})^{2/3}$, która zazwyczaj przewyższa znacznie średnią energię kinetyczną $^3/_2\ kT$. Obszar zacieniowany na rysunku przedstawia iloczyn obu czynników pod znakiem całki (3). Widać że najwięcej jest cząstek o takich energiach, przy których przenikalność bariery kulombowskiej jest



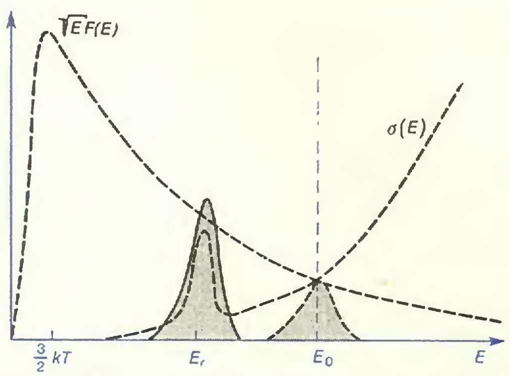
Rys. 3. Schematyczny przebieg zależności od energii dwóch czynników pod znakiem całki we wzorze (3) w wypadku, gdy w istotnym dla przebiegu reakcji termojądrowych zakresie energii nie ma żadnego rezonansu. Obszar zacieniowany odpowiada iloczynowi obu funkcji pod znakiem całki we wzorze (3)

niemal równa zeru. Liczba cząstek spada natomiast do zera, gdy wzrasta ich zdolność przechodzenia przez ową barierę. Podstawowy wkład do całki we wzorze (3) wnosi stosunkowo wąski przedział energii w pobliżu energii E_0; obszar ten nazywa się pikiem Gamowa. Gdy pik jest dość wąski, wtedy słuszne jest w przybliżeniu wyrażenie postaci

pik Gamowa

$$\langle\sigma, v\rangle = C_2 T^{-2/3} e^{-C_3 T^{-1/3}} \qquad (6)$$

(stałe C_2 i C_3 nie zależą od temperatury). Największy wkład do uśrednionego prawdopodobieństwa reakcji termojądrowej pochodzi od zderzeń przy energiach znacznie mniejszych od wysokości bariery, a jednocześnie znacznie większych od energii średniej ($E_0 \gg \gg kT$). Wynika to stąd, że bariera ma przenikalność różną od zera już przy stosunkowo małych energiach. Przybliżone wyrażenie (6) wskazuje na dość dużą zależność szybkości reakcji termojądrowych od temperatury. Dokładne obliczenie szybkości reakcji wymaga całkowania numerycznego z uwzględnieniem rezonansów możliwych w danej reakcji.



Rys. 4. Analogiczna zależność jak na rys. 3 w wypadku, gdy w obszarze bardzo małych energii występuje niewielki rezonans. Obszary zacieniowane wskazują na wartość iloczynu dwóch funkcji podcałkowych ($\sqrt{E}\cdot F(E)$ — związanej z rozkładem prędkości i energii, oraz $\sigma(E)$ — przekroju czynnego) we wzorze stanowiącym uogólnienie wyrażenia (3). (Wzoru (3) nie stosujemy bezpośrednio w przypadku rezonansu). Wkład pochodzący od niskoenergetycznego rezonansu większy jest od wkładu z otoczenia energii E_0

Jeśli w przekroju czynnym przy małych energiach w obszarze podbarierowym występuje rezonans (jeden lub kilka), to wkład do szybkości reakcji termonuklearnej z obszaru energii wokół tego rezonansu może się okazać większy od wkładu piku Gamowa. Sytuację taką przedstawia rys. 4.

Między reakcjami jądrowymi w laboratorium a reakcjami termojądrowymi w gwiazdach występuje pewna istotna różnica. W gwiazdach oprócz jąder atomowych jest gęsty gaz elektronowy. Jeśli nawet atomy są całkowicie zjonizowane, to obecność gazu elektronowego obniża w pewnym stopniu wzajemne odpychanie kulombowskie jąder i przekroczenie bariery kulombowskiej jest łatwiejsze.

faza kontrakcji grawitacyjnej

Wstępna faza ewolucji gwiazdy, zwana fazą kontrakcji grawitacyjnej, kończy się z chwilą, gdy w centrum gwiazdy osiągnięta zostaje temperatura wystarczająca na to, by możliwe stały się reakcje jądrowe z udziałem najobfitszego paliwa — wodoru, odznaczającego się zarazem (dzięki posiadaniu minimalnego — w porównaniu z innymi pierwiastkami — ładunku elektrycznego) największą łatwością przenikania przez barierę kulombowską. Wodór może się więc spalać w takiej temperaturze (i energii średniej jąder!), w jakiej reakcje między jądrami innych pierwiastków są jeszcze niemożliwe.

spalanie wodoru

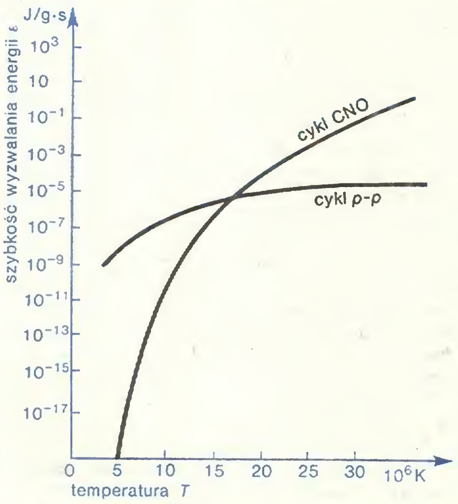
Dwa zasadniczo różne sposoby spalania wodoru to: cykl *p-p* oraz cykl katalityczny CNO, zwany cyklem węglowo-azotowym albo cyklem Bethego–Weizsäckera. W każdym z nich można rozróżnić tzw. gałąź główną i dwie gałęzie boczne (tabela 1). Względna częstość występowania poszczególnych cykli i gałęzi zależy od warunków fizycznych wewnątrz gwiazdy. Tak więc np. w gwiazdach pierwszego pokolenia, które się składały wyłącznie z helu i wodoru, bez domieszki węgla czy azotu, możliwy był jedynie cykl *p-p*. Obecnie w gwiazdach o niewielkiej masie (np. w Słońcu) za wydzielanie energii odpowiada wyłącznie cykl *p-p*, w gwiazdach zaś bardzo masywnych, o masie rzędu kilkudziesięciu mas słonecznych, wodór wypala się wyłącznie w cyklu CNO. Przyczyna tego jest prosta. W gwiazdach o większej masie, dysponujących większym zapasem energii grawitacyjnej, wstępna kontrakcja grawitacyjna może doprowadzić do znacznie silniejszego ogrzania wnętrza gwiazdy. W wyższej zaś temperaturze jądra wodoru przenikają skuteczniej przez bariery kulombowskie wokół jąder węgla i azotu, a procesy (10, 12, 13 i 15) z tabeli 1 przebiegają znacznie szybciej od pierwszej reakcji cyklu *p-p*, choć bariera kulombowska jest w niej nieporównywalnie niższa. Reakcja (1) jest bowiem jedyną (oprócz przemian β^+ i wychwytów elektronu) reakcją jądrową z tabeli 1, wywoływaną oddziaływaniami słabymi. Stąd też przekrój czynny na tę reakcję jest mniejszy o kilkanaście rzędów wielkości od przekrojów czynnych na inne reakcje jądrowe, wywoływane oddziaływaniami silnymi. Wartość jego przy tych energiach protonów, przy jakich one ze sobą w gwieździe reagują (tj. poniżej 1 MeV), wynosi ok. 10^{-23} b, podczas gdy przekroje na inne reakcje (przy porównywalnych energiach) są zazwyczaj rzędu mikrobarnów do milibarnów.

Aby scharakteryzować efektywność gałęzi głównych obu cykli spalania wodoru, podajemy w tabeli 1 czasy życia jąder ze względu na daną reakcję, oszacowane przy założeniu następujących warunków fizycznych w obszarze, w którym reakcje te zachodzą: gęstość wodoru równa 100 g/cm³, temperatura — ok. $1{,}3\cdot10^7$ K. Wypadkową szybkość przebiegu cyklu określa najdłuższy czas życia. Na rys. 5 przedstawiono zależność szybkości wyzwalania energii od temperatury przy ustalonym składzie chemicznym. Widać stąd, że cykl CNO dominuje w gwiazdach o dużej jasności, masie i temperaturze (znajdujących się po lewej górnej stronie ciągu głównego gwiazd).

Podobnie jak przewaga określonego cyklu spalania wodoru, tak też przewaga różnych gałęzi danego cy-

Tabela 1. Reakcje spalania wodoru

Cykl	Gałąź	Reakcja (w nawiasie czas życia)		Uwagi	
p-p	Główna	$^{1}H+^{1}H \to ^{2}H+e^{+}+\nu_e(1,4\cdot 10^{10}$ l)	(1)	konkuruje reakcja: $^{1}H+^{1}H+e^{-} \to ^{2}H+\nu_e$	(1a)
		$^{2}H+^{1}H \to ^{3}He+\gamma(5,7s)$	(2)		
		$^{3}He+^{3}He \to ^{4}He+2^{1}H(10^{6}$ l)	(3)		
	Boczna I	Najpierw (1) i (2), potem:			
		$^{3}He+^{4}He \to ^{7}Be+\gamma$	(4)		
		$^{7}Be+e^{-} \to ^{7}Li+\nu_e$	(5)		
		$^{7}Li+^{1}H \to 2^{4}He$	(6)		
	Boczna II	Najpierw (1), (2) i (4), potem:			
		$^{7}Be+^{1}H \to ^{8}B+\gamma$	(7)		
		$^{8}B \to ^{8}Be^{*}+e^{+}+\nu_e$	(8)	lub $^{8}B+e^{-} \to ^{8}Be+\nu_e$	(8a)
		$^{8}Be \to 2^{4}He$	(9)		
CNO	Główna	$^{12}C+^{1}H \to ^{13}N+\gamma(1,3\cdot 10^{7}$ l)	(10)		
		$^{13}H \to ^{13}C+e^{+}+\nu_e(7m)$	(11)	lub $^{13}N+e^{-} \to ^{13}C+\nu_e$	(11a)
		$^{13}C+^{1}H \to ^{14}N+\gamma(2,7\cdot 10^{6}$ l)	(12)		
		$^{14}N+^{1}H \to ^{15}O+\gamma(< 3,2\cdot 10^{8}$ l)	(13)		
		$^{15}O \to ^{15}N+e^{+}+\nu_e$ (82s)	(14)	lub $^{15}O+e^{-} \to ^{15}N+\nu_e$	(14a)
		$^{15}N+^{1}H \to ^{12}C+^{4}He$ (1,1·10[5] l)	(15)		
	Boczna I	↓ $^{15}N+^{1}H \to ^{16}O+\gamma$	(16)		
		$^{16}O+^{1}H \to ^{17}F+\gamma$	(17)		
		$^{17}F \to ^{17}O+e^{+}+\nu_e$	(18)	lub $^{17}F+e^{-} \to ^{17}O+\nu_e$	(18a)
		$^{17}O+^{1}H \to ^{14}N+^{4}He$	(19)		
		teraz powrót do gałęzi głównej (reakcja (13))			
	Boczna II	↓ $^{17}O+^{1}H \to ^{18}F+\gamma$	(20)		
		$^{18}F \to ^{18}O+e^{+}+\nu_e$	(21)	lub $^{18}F+e^{-} \to ^{18}O+\nu_e$	(21a)
		$^{18}O+^{1}H \to ^{15}N+^{4}He$	(22)		
		teraz powrót do gałęzi głównej (reakcja (15))			



Rys. 5. Szybkość wyzwalania energii ε cykli *p-p* i CNO w zależności od temperatury T przy gęstości 100 g/cm³ i zawartości wagowej: wodoru H — 80%, węgla C i azotu N (łącznie) — 0,6%

kłu zależy od warunków fizycznych wewnątrz gwiazdy. Tak więc np. w najniższych temperaturach (oczywiście powyżej temperatury „zapalania się" wodoru) przebiegają wyłącznie te reakcje cyklu *p-p*, które wchodzą w skład jego gałęzi głównej. Ze wzrostem temperatury włączają się kolejno obie gałęzie boczne. Taką sytuację ilustruje rys. 6.

reakcje jądrowe w Słońcu

Oto kilka danych orientacyjnych dotyczących Słońca. Wkład cyklu CNO do produkcji energii jest rzędu paru procent najwyżej. W cyklu *p-p* podstawową rolę odgrywa gałąź główna (względna częstość występowania ok. 91%). następnie gałąź boczna I (względna częstość ok. 9%). Temperatura centralna jest zbyt niska, by większą rolę mogła odgrywać gałąź boczna II (względna częstość poniżej 0,01%). Dodajmy, że choć energia wydzielona podczas przemiany wodoru w hel jest na każdej drodze taka sama, rozkłada się ona w rozmaity sposób między neutrina i promieniowanie elektromagnetyczne; fakt ten tkwi u podstaw astronomii neutrinowej, dzięki której można jednocześnie wyznaczać temperaturę wnętrza gwiazdy i sprawdzać, czy i jakie reakcje jądrowe tam zachodzą. Wszystkie reakcje spalania wodoru prowadzą zatem do jednego tylko produktu końcowego: helu ^{4}He; zapisać je można schematycznie:

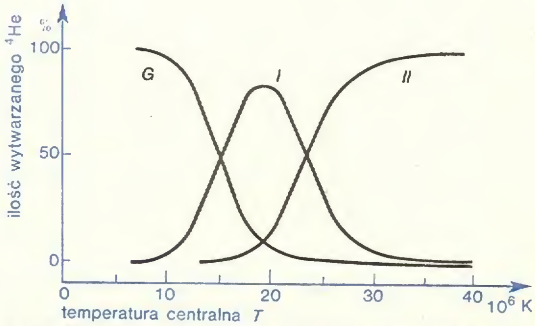
produkt reakcji spalania wodoru — ⁴He

$$4^{1}H \to ^{4}He+2e^{+}+2\nu_e+2\gamma. \qquad (7)$$

Gdy spalanie jest stacjonarne, wszystkie pozostałe jądra, występujące w kolejnych reakcjach (z wyjątkiem takich jąder katalizatorów jak np. ^{12}C), mają stężenie niewielkie. Tak więc procesy z tabeli 1 prawie nie zwiększają rozpowszechnienia litu, berylu czy boru w materii przechodzącej przez fazę spalania wodoru.

Gwiazdy większą część swego życia spędzają w fazie spalania wodoru. Zarówno z tego względu, jak i dlatego, że jest to proces największej wydajności energetycznej (największej wartości wydzielonej energii przypadającej na jednostkę masy paliwa), spalanie wodoru należy do najważniejszych reakcji jądrowych w przyrodzie.

Następnymi reakcjami jądrowymi — i pod względem ważności, i kolejności występowania w ewolucji



Rys. 6. Wytwarzanie helu ⁴He w warunkach stacjonarnych przy założeniu $X = Y$ (X — zawartość wagowa wodoru, Y — zawartość wagowa helu) w gałęzi głównej (G) i gałęziach bocznych (I i II) cyklu *p-p* w zależności od temperatury centralnej T. Podobne wykresy można przedstawić dla innych wartości stosunku $X:Y$ (wg Parker, Bahcall i Fowler, 1964)

spalanie helu

gwiazdy — są reakcje spalania helu. Gdy w centralnych obszarach gwiazdy spalającej wodór zostanie on zupełnie zużyty, wtedy zmaleje udział ciśnienia promieniowania (bo mniej promieniowania powstaje) w łącznym ciśnieniu przeciwstawiającym się kurczeniu grawitacyjnemu. Następuje grawitacyjna kontrakcja wnętrza gwiazdy, przy czym wzrasta gęstość (do ok. 10^5 g/cm^3) oraz — wskutek wydzielania się energii grawitacyjnej — temperatura (do ok. 10^8 K). W takich warunkach możliwe staje się zapalenie następnego paliwa — helu. Wydawałoby się, że powinna teraz nastąpić reakcja:

$$^{4}_{2}\mathrm{He} + ^{4}_{2}\mathrm{He} \rightarrow ^{8}_{4}\mathrm{Be}. \qquad (8)$$

Możliwość przebiegu tej reakcji można by zakwestionować w związku z tym, że jest to reakcja endoenergetyczna, o energii progowej rzędu 96 keV. Nie to jest jednak zasadniczą przeszkodą, gdyż w temperaturze 10^8 K spora część jąder helu ^{4}He ma energię wyższą od tej wartości progowej. Natomiast z danych doświadczalnych wiadomo, że jądro ^{8}Be z okresem połowicznego zaniku rzędu $2{,}6 \cdot 10^{-16}$ s rozpada się z powrotem na dwa jądra ^{4}He. Czy wobec tego reakcja (8) może odgrywać jakąkolwiek rolę?

Okazuje się, że tak, bo jedynie w warunkach ziemskich jądro ^{8}Be ma tylko jedną możliwość zaniku: rozpad na dwie cząstki α. W warunkach zaś bardzo gęstej materii wnętrza gwiazdy oraz w wysokiej temperaturze zachodzi tyle zderzeń między jądrami helu, że się w końcu ustala pewna równowaga między procesami powstawania i rozpadu berylu, a więc i równowaga między stężeniami helu i berylu. Równowaga ta zależy od gęstości i temperatury. Przy gęstości 10^5 g/cm^3 i w temperaturze 10^8 K przypada jedno jądro ^{8}Be na miliard jąder ^{4}He, co umożliwia wychwyt helu przez beryl (zanim ten ostatni zdąży się rozpaść):

$$^{8}\mathrm{Be} + ^{4}\mathrm{He} \rightarrow ^{12}\mathrm{C} + \gamma. \qquad (9)$$

Innymi słowy — w warunkach tak ogromnego stężenia cząstek α, jakie istnieje wewnątrz gwiazdy, zanik nietrwałego jądra ^{8}Be może następować również przez wychwyt cząstki α, co prowadzi do powstania trwałego jądra węgla ^{12}C. Nic więc dziwnego, że węgiel jest czwartym pod względem rozpowszechnienia pierwiastkiem (po wodorze helu i tlenie).

Dwuetapowy proces złożony z reakcji (8) i (9) zapisać można krótko:

$$3\alpha \rightarrow ^{12}\mathrm{C} + \gamma, \qquad (10)$$

proces 3α

co uzasadnia jego nazwę: proces 3α. Należy podkreślić, że reakcja (9) jest reakcją rezonansową w tym przedziale wartości energii, jakimi dysponują jądra ^{4}He wewnątrz gwiazd spalających hel. Przekrój czynny na tę reakcję jest więc tak duży, że jądra węgla powstają z dużą wydajnością — mimo niewielkiego stężenia jąder berylu. Jednocześnie możliwe jest przyłączanie dalszych cząstek α do wytworzonych już jąder węgla; powstaje przy tym tlen:

$$^{12}_{6}\mathrm{C} + ^{4}_{2}\mathrm{He} \rightarrow ^{16}_{8}\mathrm{O} + \gamma. \qquad (11)$$

Gdy w centrum gwiazdy hel jest już całkowicie zużyty, sytuacja wygląda podobnie jak po wypaleniu wodoru. Raz jeszcze nastąpić musi kontrakcja grawitacyjna i zapalenie się kolejnego paliwa, będącego jednocześnie „popiołem" jądrowym z poprzedniej fazy spalania. Przykłady reakcji jądrowych spalania węgla i tlenu zawiera tabela 2.

Wszystkie reakcje jądrowe wytwarzające energię w gwieździe (z wyjątkiem lekko endoenergetycznej reakcji 8) są to dość podobne do siebie reakcje egzoenergetyczne między cząstkami naładowanymi. W każdej z nich z jąder słabiej związanych tworzą się jądra o większej energii wiązania, a jednocześnie wydziela się energia. Produkcji energii w gwieździe towarzyszą zmiany składu chemicznego jej wnętrza, popioły poprzedniej fazy spalania stają się paliwem w następnej. Prowadzi to do jąder pierwiastków coraz cięższych, aż do jąder żelazowców, które mają

Tabela 2. Niektóre reakcje jądrowe zachodzące po wypaleniu helu

I. Spalanie węgla:

$$^{12}_{6}\mathrm{C} + ^{12}_{6}\mathrm{C} \rightarrow \begin{cases} ^{20}_{10}\mathrm{Ne} + ^{4}_{2}\mathrm{He} \\ ^{23}_{11}\mathrm{Na} + ^{1}_{1}\mathrm{H} \\ ^{24}_{12}\mathrm{Mg} + \gamma \\ ^{23}_{12}\mathrm{Mg} + n \\ ^{16}_{8}\mathrm{O} + 2\,^{4}_{2}\mathrm{He} \end{cases} \qquad (23)$$

II. Spalanie tlenu:

$$^{16}_{8}\mathrm{O} + ^{16}_{8}\mathrm{O} \rightarrow \begin{cases} ^{28}_{14}\mathrm{Si} + ^{4}_{2}\mathrm{He} \\ ^{31}_{15}\mathrm{P} + ^{1}_{1}\mathrm{H} \\ ^{31}_{16}\mathrm{S} + n \\ ^{32}_{16}\mathrm{S} + \gamma \\ ^{24}_{12}\mathrm{Mg} + 2\,^{4}_{2}\mathrm{He} \end{cases} \qquad (24)$$

III. Spalanie krzemu:

$$^{28}_{14}\mathrm{Si} + ^{4}_{2}\mathrm{He} \rightarrow ^{32}_{16}\mathrm{S} + \gamma \qquad (25)$$

$$^{28}_{14}\mathrm{Si} + ^{28}_{14}\mathrm{Si} \rightarrow ^{56}_{28}\mathrm{Ni} + \gamma \qquad (26)$$

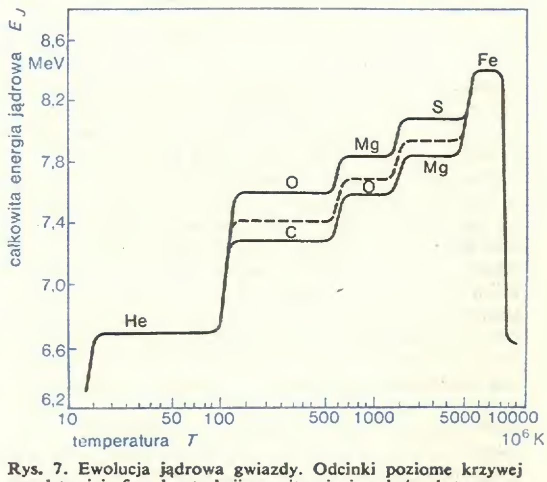
(praktycznie nie zachodzi, ze względu na wysoką barierę kulombowską)

$$^{28}_{14}\mathrm{Si} + \gamma \rightarrow \begin{cases} ^{24}_{12}\mathrm{Mg} + ^{4}_{2}\mathrm{He} \\ ^{27}_{13}\mathrm{Al} + ^{1}_{1}\mathrm{H} \end{cases} \qquad (27)$$

największą energię wiązania na jeden nukleon.

powstawanie jąder żelazowców

Dalsze reakcje nukleosyntezy między jądrami żelazowców (i ewentualnie pierwiastków cięższych) musiałyby być reakcjami endoenergetycznymi i nie mogłyby stanowić źródła energii gwiazdy (rys. 7). Ponadto z uwagi na coraz wyższą barierę kulombowską między oddziałującymi ze sobą jądrami konieczny byłby coraz większy wzrost energii kinetycznej oddziałujących z sobą jąder, a temu musiałby towarzyszyć prawie nieograniczony wzrost temperatury centralnej. Sprawia to, że powstania jąder atomowych o liczbach masowych $A \geqslant 65$ nie da się powiązać z reakcjami termojądrowymi między cząstkami naładowanymi. Występowanie w przyrodzie takich jąder (pierwiastki od cynku do uranu) wskazuje jednak na to, że procesy ich powstawania odbywały się, a podstawową w nich rolę odgrywały reakcje z udziałem neutronów.



Rys. 7. Ewolucja jądrowa gwiazdy. Odcinki poziome krzywej przedstawiają fazy kontrakcji grawitacyjnej, w których temperatura centralna rośnie, a skład chemiczny gwiazdy nie ulega zmianie; przy każdym z nich podano symbol pierwiastka, którego jądra ulegną „zapłonowi" po zakończeniu fazy kontrakcji. Procesy spalania kolejnych paliw jądrowych przedstawiają odcinki niemal pionowe (spalanie izotermiczne). Krzywe C–O–Mg i O–Mg–S odpowiadają dwom przypadkom skrajnym; gdy jądro gwiazdy po wypaleniu helu składa się wyłącznie z węgla i gdy składa się wyłącznie z tlenu. E_J jest to całkowita energia jądrowa jaka została wydobyta z gwiazdy od chwili jej powstania, przypadająca na jeden jej nukleon. Gdy w jądrze gwiazdy pozostają jądra żelaza, niemożliwe jest dalsze wydobywanie energii jądrowej. Możliwe jest wtedy przejście fazowe Fe → He; energii potrzebnej do tego dostarcza kontrakcja grawitacyjna

Reakcje wytwarzania i wychwytu neutronów

Neutrony mogą się pojawiać jako jeden z produktów końcowych w różnych reakcjach termojądrowych podczas ewolucji gwiazdy. W tabeli 2 podaliśmy przykładowo procesy powstawania neutronów podczas spalania węgla lub tlenu. Przedstawimy jeszcze kilka przykładów.

reakcje wytwarzania neutronów

Kiedy we wnętrzu masywnej gwiazdy, w której przebiegał cykl CNO, wypalił się wodór, wtedy obok jąder ^{4}He pozostaje pewna domieszka jąder ^{13}C — pozostałości katalizatora z tego cyklu. W fazie kontrakcji grawitacyjnej wnętrza gwiazdy jądra helu wstępują w reakcję

$$^{13}_{6}C+^{4}_{2}He \rightarrow ^{16}_{8}O+n, \qquad (12)$$

która jest znacznie bardziej prawdopodobna od dwuetapowego procesu 3α.

Niezależnie od tego, jaka była proporcja wzajemna różnych odmian węgla i azotu przed rozpoczęciem spalania wodoru w cyklu CNO, w czasie jego trwania ustala się następujący stosunek ilościowy jąder-katalizatorów:

$$(^{14}_{7}N):(^{12}_{6}C):(^{13}_{6}C) = 95:4:1. \qquad (13)$$

Widać stąd, że najobfitszą pozostałością ciężką po zakończeniu cyklu CNO jest azot. Jądra jego wychwytują znaczną część neutronów powstających w reakcji (12), co obniża wydajność innych reakcji z neutronami. Ale już po zapaleniu się helu, w wyższej temperaturze, możliwa jest reakcja wychwytu cząstki α przez jądro ^{14}N, w której się tworzy jądro ^{18}O, zdolne do kolejnego wychwytu cząstki α. Oto ciąg dopuszczalnych procesów:

$$^{14}_{7}N+^{4}_{2}He \rightarrow ^{18}_{9}F+\gamma, \qquad (14)$$

$$^{18}F \rightarrow ^{18}O+e^{+}+\nu_e, \qquad (15)$$

$$^{18}O+^{4}He \rightarrow ^{22}Ne+\gamma, \qquad (16)$$

w których się tworzą dwa jądra (^{18}O i ^{22}Ne), niezbędne do dalszego wytwarzania neutronów. Obie reakcje produkcji neutronów:

$$^{18}_{8}O+^{4}_{2}He \rightarrow ^{21}_{10}Ne+n, \qquad (17)$$

$$^{22}_{10}Ne+^{4}_{2}He \rightarrow ^{25}_{12}Mg+n, \qquad (18)$$

są endoenergetyczne. Progi ich są jednak tak niskie (0,705 MeV reakcji 17 i 0,48 MeV reakcji 18), że mogą one przebiegać wydajnie w temperaturze, jaka panuje w jądrze gwiazdy pod koniec fazy spalania helu.

Powyższe procesy wytwarzania neutronów związane są z fazą spalania helu. Można by jeszcze wymienić inne procesy produkcji neutronów, np. spalanie krzemu, siarki, ale ograniczymy się do stwierdzenia, że neutrony pojawiają się wewnątrz gwiazdy w wyniku kilku stosunkowo rzadkich reakcji. Stanowi to konsekwencję albo niewielkiego rozpowszechnienia w materii gwiazdowej tych jąder-substratów, które są źródłem neutronów (czyli jąder ^{13}C, ^{18}O i ^{22}Ne), albo też niewielkiej stosunkowo częstości występowania kanałów neutronowych wśród kanałów wyjściowych zachodzących reakcji. Neutrony pojawiają się w niewielkich stosunkowo ilościach, ale w ciągu dość długiego przedziału czasu. Można w zasadzie stwierdzić, że począwszy od fazy spalania helu istnieją w gwieździe pewne stacjonarne źródła neutronów. Wytwarzanie neutronów odbywa się w skali czasu rzędu milionów do setek milionów lat (w zależności od masy gwiazdy, a więc i od tempa jej ewolucji).

pochłanianie neutronów

„Losy" powstałych neutronów zależą od obecności w materii gwiezdnej jąder, które mogą je pochłaniać. Tak więc wspomniane już kilkakrotnie jądro ^{14}N w reakcji

$$^{14}_{7}N+n \rightarrow ^{14}_{6}C+p \qquad (19)$$

pochłania neutrony powstające w procesach (α, n), np. w reakcji (12) i (17). Tworzące się w reakcji (19) nietrwałe jądro izotopu węgla ^{14}C ulega przemianie β^- (z okresem połowicznego zaniku $t_{1/2}$ ok. 5,5 tys. lat) i odtwarza „truciznę neutronową" ^{14}N.

powstawanie pierwiastków ciężkich

Do znacznie ciekawszych konsekwencji prowadzić może obecność jąder mających duże przekroje czynne na reakcje (n, γ). Jądrami takimi są jądra żelazowców i pierwiastków cięższych. Na przykład jądro najbardziej rozpowszechnionego izotopu żelaza, $^{56}_{26}Fe$, po wychwycie pierwszego neutronu przekształca się w trwałe jądro $^{57}_{26}Fe$, po wychwycie drugiego — również w trwałe jądro $^{58}_{26}Fe$, dopiero następny wychwyt neutronu tworzy nietrwałe jądro $^{59}_{26}Fe$, rozpadające się przez przemianę β^- ($t_{1/2} = 45$ dni):

$$^{59}_{26}Fe \rightarrow ^{59}_{27}Co+e^{-}+\bar{\nu}_e. \qquad (20)$$

Istotne jest to, że trzy kolejne wychwyty neutronu wraz z następującą po nich przemianą β doprowadzają do przemiany jądra żelaza w jądro następnego po nim w układzie okresowym pierwiastka. Wychwyt kolejnego neutronu przez jądro $^{59}_{27}Co$ tworzy nietrwałe jądro $^{60}_{27}Co$, z którego w wyniku przemiany β^- powstaje trwałe jądro dalszego pierwiastka — niklu $^{60}_{28}Ni$. Tak oto poprzez kolejne wychwyty neutronów tworzyć się mogą pierwiastki o coraz większych liczbach atomowych (→ Powstawanie pierwiastków chemicznych).

szybki wychwyt neutronów

Zwróćmy jeszcze uwagę na to, że poza mechanizmem powolnego wychwytu neutronów, który się wiąże ze stacjonarnymi źródłami neutronów w gwieździe, istnieje jeszcze inny sposób powstawania pierwiastków ciężkich: szybki wychwyt neutronów występujący przy pojawieniu się dużego strumienia neutronów w ciągu krótkiego czasu. Właśnie taki strumień pojawia się w końcowej fazie ewolucji gwiazd o dużej masie, gdy eksplodują jako supernowe. W centrum takiej gwiazdy ciąg kolejnych reakcji egzoenergetycznych doprowadza prawdopodobnie do powstania jąder najsilniej związanych — jąder żelazowców. Dalsza produkcja energii w reakcjach termonuklearnych nie jest już możliwa, następuje kontrakcja grawitacyjna jądra gwiazdy, podwyższenie gęstości i temperatury. W efekcie możliwa się staje endoenergetyczna przemiana jąder żelazowców w hel i neutrony:

$$^{56}_{26}Fe \rightarrow 13^{4}_{2}He+4n, \qquad (21)$$

a wraz z dalszym zagęszczaniem się jądra gwiazdy możliwy jest i rozpad helu. W rezultacie powstają ogromne ilości neutronów, z których przynajmniej część opuszcza centralny obszar gwiazdy. Neutrony te napromieniają materię wyrzuconą z obszarów bardziej zewnętrznych, o niższej gęstości i temperaturze, zawierającą jądra żelazowców (i jądra cięższe), zdolne do wychwytu neutronów pochodzących z centrum gwiazdy. W tych warunkach w czasie rzędu minuty przypada od kilkudziesięciu do kilkuset neutronów na jedno jądro, które je może wychwycić.

Astrofizyka neutrinowa i śledzenie reakcji jądrowych w gwiazdach

Hipoteza jądrowych źródeł energii gwiazd wydaje się obecnie najbardziej przekonująca. Przemawia za nią zgodność wyprowadzonych na jej podstawie wniosków z obserwacją. Reakcje jądrowe przebiegają jednak w niewielkiej, całkowicie ukrytej przed obserwatorem, centralnej części gwiazdy i nikt ich bezpośrednio nie obserwował.

Weźmy np. pod uwagę najbliższą gwiazdę — Słońce. Obserwowane promieniowanie dostarcza bezpośredniej informacji tylko o atmosferze Słońca mającej temperaturę ok. 6 tys. stopni, podczas gdy

promieniowanie Słońca

temperatura w jego jądrze wynosi kilkanaście mln stopni. Wiemy poza tym, że Słońce wypromieniowuje ok. $3{,}86 \cdot 10^{26}$ J/s, ale obserwacje promieniowania elektromagnetycznego, wydobywającego się z jego powierzchni, nie mówią nic o reakcjach jądrowych w centrum Słońca, umożliwiających wypromieniowywanie takiej energii. O tych reakcjach wnosić możemy jedynie pośrednio, na podstawie teoretycznego modelu wnętrza Słońca. Poprawny model musi uwzględniać to, że wkład każdej reakcji termojądrowej do łącznej produkcji energii zależy od temperatury, gęstości w centrum, składu chemicznego itp.

Udział poszczególnych gałęzi cykli *p-p* i CNO we wnętrzu Słońca można dopasować do jego jasności powierzchniowej, typu widmowego i innych znanych własności, pozostanie jednak pewna wątpliwość, nad którą przyrodnikowi nigdy nie wolno przejść do porządku dziennego: A może u podstaw teorii ewolucji gwiazd tkwi jakaś fałszywa przesłanka dotycząca procesów zachodzących w ich wnętrzu? Dylemat ten można by rozstrzygnąć, gdyby się udało uzyskać bezpośrednią informację o procesach wewnątrz gwiazd. Źródłem takiej informacji mogą być neutrina, które powstają jako „produkt uboczny" w różnych reakcjach spalania wodoru (tabela 1). Bez przeszkód opuszczają Słońce i poruszając się z prędkością światła, docierają do Ziemi w ciągu mniej więcej 8 minut od chwili, gdy w jądrze słonecznym zostały wytworzone. Mają przy tym niemal tę samą energię i kierunek lotu co w chwili powstania. Wynika to z niezwykle słabego oddziaływania neutrin z materią (→ Oddziaływania słabe).

średnia droga swobodna neutrin

Średnia droga swobodna neutrin o energiach do kilku MeV w materii o normalnej gęstości (rzędu 1 g/cm³) sięga milionów miliardów kilometrów. Natomiast średnia droga swobodna kwantu γ pochodzącego z wnętrza Słońca jest miliardy razy mniejsza od rozmiarów Słońca. Zatem obserwowane widmo promieniowania elektromagnetycznego Słońca nie odzwierciedla widma kwantów γ powstających w reakcjach jądrowych w obszarach centralnych. Oddziaływania elektromagnetyczne fotonów, przedzierających się z wnętrza Słońca ku jego powierzchni, opóźniają o tysiące lat ostateczne wypromieniowanie przez Słońce energii wytworzonej w jego wnętrzu. Ponadto ewentualne oscylacje w wytwarzaniu energii w centrum Słońca ulegają znacznemu rozmyciu w czasie i mogą się w ogóle nie odbijać na jasności fotonowej Słońca. Natomiast obserwacje neutrin umożliwiłyby w zasadzie wyznaczenie udziału różnych procesów jądrowych w produkcji energii wewnątrz Słońca oraz śledzenie zmian tego udziału z upływem czasu.

Neutrina z różnych reakcji w Słońcu charakteryzują się różnym widmem energetycznym. Na przykład reakcja (1) z tabeli 1 jest wywoływana oddziaływaniem słabym, zbliżona jest więc swym charakterem do przemiany β. Widmo neutrin z tej reakcji jest ciągłe, podobnie jak widmo neutrin z przemiany β^+. Sięga ono od granicy dolnej, równej zeru, do energii maksymalnej, równej 0,42 MeV. Reakcja (1) występuje we wszystkich gałęziach cyklu *p-p*, pochodzą z niej neutrina o stosunkowo małej energii. Z reakcją (1) konkuruje proces (1a), analogiczny do procesu wychwytu elektronu. Prawdopodobieństwo procesu (1a) jest mniej więcej 400 razy mniejsze od prawdopodobieństwa reakcji (1). Proces (1a) dostarcza neutrin monoenergetycznych, o energii 1,44 MeV. W porównaniu z liczbą neutrin otrzymywanych z reakcji (1) jest ich jednak bardzo mało.

widmo energetyczne neutrin

W tabeli 3 zestawione są dane o procesach powstawania neutrin wewnątrz Słońca. Łatwo dostrzec, że gdyby głównym źródłem energii była II gałąź boczna cyklu *p-p*, to neutrina ze Słońca miałyby najwyższą możliwą wartość energii średniej — połowę neutrin stanowiłyby bowiem neutrina borowe, pochodzące z procesu (8), o energii maksymalnej 14,06 MeV. Neutrina takie najłatwiej byłoby wykryć, gdyż przekrój czynny na oddziaływanie (słabe!) neutrin z materią szybko rośnie ze wzrostem energii neutrin. Co do neutrin pochodzących z I gałęzi bocznej cyklu *p-p*, to energia jednego z neutrin (z reakcji 1) nie przekracza 0,42 MeV; drugie neutrino, pochodzące z wychwytu elektronu (5), może mieć energię 0,861 MeV albo 0,383 MeV — zależnie od tego, czy jądro ^{7}Li tworzy się w stanie podstawowym czy też wzbudzonym. Oddziaływanie neutrin litowych z materią słabsze jest średnio od oddziaływania neutrin borowych, silniejsze natomiast od neutrin *pp* lub *pep*.

Bruno Pontecorvo, uczeń E. Fermiego i jeden z najwybitniejszych znawców problematyki neutrinowej, zaproponował ponad 30 lat temu wykorzystanie reakcji odwrotnej przemiany β do radiochemicznej detekcji neutrin docierających ze Słońca do Ziemi. W reakcjach tych wykrywane byłyby neutrina wywołujące przemianę neutronu w proton w jakimś trwałym jądrze atomowym. Zestawienie proponowanych reakcji zawiera tabela 4. Wszystkie reakcje odwrotne przemiany β są reakcjami progowymi, tj. wymagają pewnej minimalnej energii neutrina. Podaliśmy je w kolejności odpowiadającej zwiększaniu progu

detekcja neutrin

Tabela 3. Energia unoszona przez neutrina pochodzące z reakcji termojądrowych

Cykl	Reakcje (z tab. 1), w których powstają neutrina	Nazwa grupy neutrin i rodzaj widma energetycznego	Energia neutrin MeV	Energia (MeV) unoszona średnio w 1 cyklu przez		Ułamek energii całkowitej unoszony przez neutrina
				promieniowanie elektromagnetyczne	neutrina	
p-p, gałąź główna	(1) (1a) — nieznaczny udział w produkcji energii i neutrin	neutrina *pp* (widmo ciągłe) neutrina *pep* (widmo liniowe)	0 do 0,42 1,44	26,22	0,51	2%
p-p, gałąź boczna I	(1) i (1a) — nieznaczny udział (jw.) (5)	neutrina *pp* i *pep* neutrina berylowe (widmo liniowe)	0–0,42; 1,44 {0,861 (90%) i 0,383 (10%)	25,67	1,06	4%
p-p, gałąź boczna II	(1) i (1a) — nieznaczny udział (jw.) (8) (8a) — nieznaczny udział (jw.)	neutrina *pp* i *pep* neutrina borowe (widmo ciągłe) (widmo liniowe)	0–0,42; 1,44 0 do 14,06 15,08	19,1	7,63	28%
CNO, gałąź główna	(11) (14) (11a) i (14a) — nieznaczny udział w produkcji energii	neutrina azotowe (widmo ciągłe) neutrina tlenowe (widmo ciągłe) (widmo liniowe)	0 do 1,20 0 do 1,74 2,22 i 2,76	24,97	1,76	6%

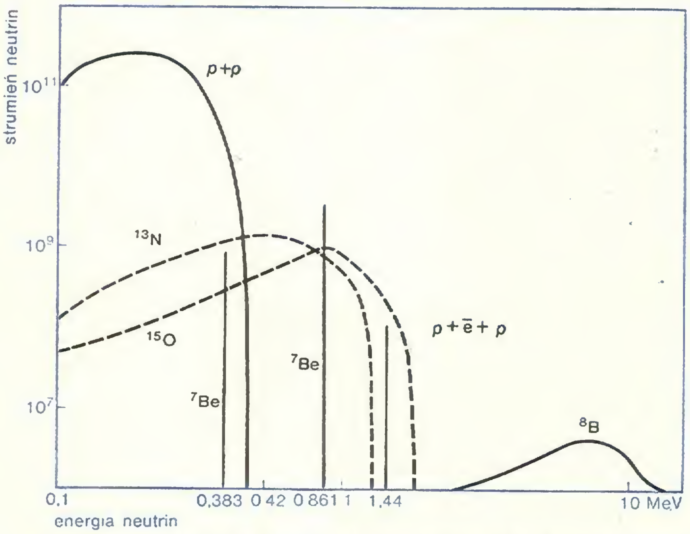
Tabela 4. Reakcje odwrotnej przemiany β zaproponowane do detekcji neutrin słonecznych

Reakcje		Rozpowszechnianie wyjściowego nuklidu w % (w stosunku do danego pierwiastka)	Energia progowa reakcji MeV	Rodzaj rozpadu i okres połowicznego zaniku powstającego w danej reakcji nuklidu promieniotwórczego	Rząd wielkości masy pierwiastka (nie nuklidu), w której zachodzi 1 wychwyt neutrina ze Słońca na dobę
${}^{87}_{37}Rb+\nu_e \to {}^{87m}_{38}Sr+e^-$	(28)	27,85	0,115	wychwyt elektronu, γ, 2,8 h	10 t
${}^{55}_{25}Mn+\nu_e \to {}^{55}_{26}Fe+e^-$	(29)	100	0,231	wychwyt elektronu, 2,6 lat	300 t
${}^{71}_{31}Ga+\nu_e \to {}^{71}_{32}Ge+e^-$	(30)	39,6	0,233	wychwyt elektronu, 11 d	10 t
${}^{51}_{23}V+\nu_e \to {}^{51}_{24}Cr+e^-$	(31)	99,76	0,752	wychwyt elektronu γ, 27,8 d	300 t
${}^{37}_{17}Cl+\nu_e \to {}^{37}_{18}Ar+e^-$	(32)	24,47	0,816	wychwyt elektronu, 35,1 d	450 t
${}^{7}_{3}Li+\nu_e \to {}^{7}_{4}Be+e^-$	(33)	92,58	0,862	wychwyt elektronu, γ, 53,4 d	4 t
${}^{9}_{4}Be+\nu_e \to {}^{9}_{5}B+e^-$	(34)	100	0,884	trwałe produkty końcowe: $p+2\alpha$	?

energetycznego. Tylko pierwsze trzy reakcje z tabeli 4 pozwalają na wykrycie neutrin *pp*; reakcja (34) i dalsze (których nie podajemy) mają zbyt wysoką energię progową, by się nadawały do wykrycia nawet neutrin berylowych. Jeśli więc energia wytwarzana jest niemal tylko w cyklu *p-p*, to reakcja (34) i inne odbywać się mogą jedynie pod działaniem neutrin borowych (i niewielkiej liczby neutrin *pep*).

strumień neutrin

Znając wartość energii wypromieniowywanej przez Słońce, możemy oszacować produkcję neutrin: powstaje ich ok. 10^{38} w ciągu sekundy. Ich widmo energetyczne po dotarciu do powierzchni Ziemi przedstawia rysunek 8. Całkowity strumień neutrin padających na powierzchnię Ziemi wynosi ok. $7 \cdot 10^{10}$ na 1 cm² w ciągu sekundy. Wobec niezwykłej słabości oddziaływania neutrin z materią trzeba by użyć ogromnych ilości substancji do detekcji neutrin przez odwrotną przemianę β.



Rys. 8. Przewidywane (na podstawie standardowych modeli Słońca) widmo neutrin słonecznych po dotarciu ich do powierzchni Ziemi. Linie ciągłe — neutrina z cyklu *p-p*, linie przerywane — niewielka domieszka neutrin z cyklu CNO, odgrywającego podrzędną rolę w Słońcu. Strumienie neutrin podane są w 1/(cm²·s·MeV) dla widma ciągłego i w 1/(cm²·s) dla widma liniowego (neutrina monoenergetyczne z procesów (1a) i (5) z tabeli 3)

doświadczenie Davisa

Aby zobrazować, jak niezmiernie trudne jest doświadczalne wykrycie neutrin słonecznych, przedstawimy w zarysie doświadczenie Davisa, trwające już od kilkunastu lat. Raymond Davis, wybitny radiochemik z Brookhaven, wykorzystuje do detekcji neutrin reakcję (32). Detektor Davisa to ogromny zbiornik zawierający ok. 610 t (380 tys. l) czterochloroetylenu C_2Cl_4. Aby wyeliminować w maksymalnym stopniu wpływ promieniowania kosmicznego, Davis umieścił ten zbiornik na głębokości ponad 1,5 km w szybie kopalni złota w Południowej Dakocie. Czterochloroetylen został przedtem oczyszczony z ewentualnych, najmniejszych nawet domieszek argonu i innych zanieczyszczeń, zbadano również eksperymentalnie możliwości wydzielania się argonu ze ścianek zbiornika. Montaż zbiornika (z części transportowanych do szybu dźwigiem kopalnianym), sprawdzanie jego szczelności i inne prace przygotowawcze trwały ponad rok (rys. 9 oraz il. 220, tabl. 59).

przemiany służące detekcji neutrin

Idea doświadczenia jest prosta. Neutrina słoneczne przenikają bez trudu warstwę ziemi nad detektorem (a w nocy — warstwę pod nim, bo dochodzą wtedy z drugiej strony — od dołu) i wywołują przemianę chloru ^{37}Cl w argon promieniotwórczy ^{37}Ar ($t_{1/2}$ = = 35,1 dni). Rozpad argonu to reakcja (32) biegnąca od strony prawej ku lewej. Po wychwycie elektronu przez jądro ^{37}Ar następuje emisja elektronu przez atom (zjawisko Augera); takie właśnie elektrony augerowskie służą w warunkach laboratoryjnych do rejestracji rozpadów argonu.

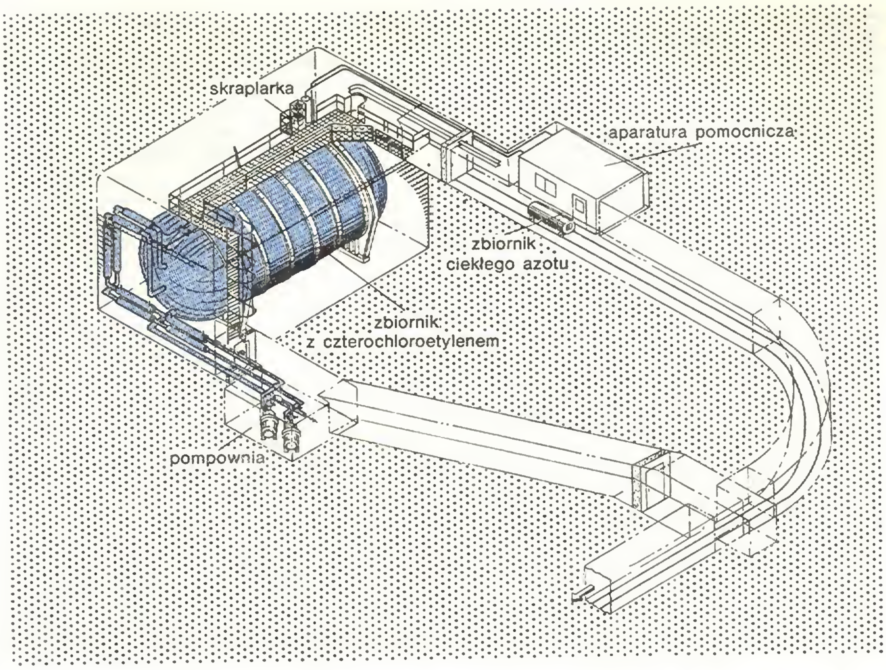
Co 100–140 dni odbywa się przedmuchiwanie zbiornika helem, który unosi ze sobą argon (oba pierwiastki są gazami szlachetnymi o podobnych właściwościach chemicznych). Do helu dodany jest trwały izotop argonu ^{36}Ar, co ułatwia ocenę stopnia usunięcia argonu ze zbiornika (a przy tym — ze względu na niewielką ilość powstającego i nietrwałego argonu ^{37}Ar — izotop trwały stanowi dla niego nośnik). Następnie argon gazowy oddziela się od helu kondensującego się na węglu drzewnym chłodzonym ciekłym azotem. Hel przepuszcza się przez zbiornik wielokrotnie, aż się odzyska co najmniej 90% wprowadzonego niepromieniotwórczego argonu ^{36}Ar. Argon mierzy się licznikiem wewnętrznego napełnienia, połączonym z analizatorem energii impulsów (w celu eliminowania impulsów o pochodzeniu innym niż od elektronów Augera).

Liczba atomów argonu wytworzonych w zbiorniku przez neutrina ze Słońca w ciągu 100–140 dni ekspozycji jest od lat liczbą najwyżej dwucyfrową (!). Podkreślamy, że wartości otrzymywane przez Davisa okazują się wciąż niższe i to aż kilkakrotnie, od ocen teoretycznych, wynikających ze standardowych modeli wnętrza Słońca. A przecież wartości doświadczalne mogłyby być większe od teoretycznych uzyskanych przy założeniu, że argon ^{37}Ar tworzy się wyłącznie w oddziaływaniach neutrin słonecznych z chlorem ^{37}Cl. Łatwiej byłoby przystać na istnienie dodatkowych mechanizmów wytwarzania jąder ^{37}Ar w zbiorniku Davisa niż na nieskuteczność normalnych sposobów jego powstawania, związanych z wielokrotnie już sprawdzonymi (w innych sytuacjach) mechanizmami oddziaływań jąder atomowych i cząstek elementarnych.

Wielokrotnie sprawdzano różne etapy doświadczenia, m.in. badano wydajność metody wydobywania

Rys. 9 Schemat obserwatorium neutrinowego Davisa

detektor Davisa



argonu ze zbiornika oraz badano niezależnie reakcje jądrowe związane choćby pośrednio z procesami produkcji neutrin w Słońcu albo z procesem (32). Nic jednak nie wskazuje na to, aby w tym tkwiła przyczyna niezgodności między przewidywaniami teoretycznymi a wynikami doświadczenia Davisa. Raczej może za mało jeszcze wiadomo o wnętrzu Słońca i może w standardowych modelach Słońca przyjmuje się zbyt wysoką temperaturę centralną. Znaczyłoby to, że wnętrze Słońca jest zbyt chłodne, by protony mogły choć w niewielkim stopniu przechodzić przez wysoką barierę potencjału jądra berylu ^{7}Be w reakcji (7, tabela 1). W takim razie po prostu nie istniałyby wysokoenergetyczne neutrina borowe z procesu (8, tabela 1), które stanowią wprawdzie nieznaczną część strumienia neutrin słonecznych, ale które się powinny najbardziej przyczyniać do produkcji argonu (ok. 90%) w reakcji (32). Wszystkie jednak proponowane dotychczas modele Słońca z niższą temperaturą w centrum wydają się sprzeczne z nagromadzoną dotychczas informacją o Słońcu i z wyobrażeniami o ewolucji gwiazd.

Najważniejsze wnioski z doświadczenia Davisa są następujące: 1) Cykl CNO nie może być podstawowym cyklem energetycznym Słońca. Gdyby bowiem tak było, to szybkość powstawania argonu w zbiorniku Davisa musiałaby o rząd wielkości przewyższać szybkość obserwowaną. 2) Temperatura centralna nie może przekraczać ok. 15 mln stopni. Gdyby była wyższa, wtedy znacznie większą rolę odgrywałaby II gałąź boczna cyklu *p-p* i znacznie więcej powstawałoby wysokoenergetycznych neutrin borowych, które w zbiorniku Davisa produkowałyby dodatkowe jądra argonu ^{37}Ar.

Obecnie rozważa się i próbuje konstruować detektory neutrin słonecznych wykorzystujące inne reakcje z tabeli 4.

inne detektory neutrin

Detektor galowy np. miałby pewną przewagę nad detektorem chlorowym, gdyż reakcja (30) ma stosunkowo niski próg energetyczny, tak że neutrina *pp*, tj. z reakcji (1, tabela 1), które nie wytwarzają argonu w detektorze Davisa, produkowałyby w detektorze galowym prawie $^{2}/_{3}$ łącznej ilości tworzącego się germanu. Neutrina *pp* muszą powstawać w Słońcu, niezależnie od temperatury jego wnętrza, jeśli tylko hipoteza jądrowych źródeł energii jest słuszna. Tak więc detektor galowy (albo rubidowy czy manganowy) o wiele bardziej niż obecnie stosowany chlorowy nadawałby się do wykazania, że reakcje jądrowe rzeczywiście muszą przebiegać wewnątrz Słońca. Natomiast znacznie trudniejszy niż w detektorze Davisa byłby problem wydzielenia niewielkiej spodziewanej liczby atomów germanu ^{71}Ge z kilku czy kilkunastu ton preparatu galowego (wydzielenie argonu — gazu szlachetnego nie tworzącego związków — jest nieporównanie łatwiejsze).

Astrofizyka neutrinowa Słońca stawia dopiero pierwsze kroki. W sferze projektów jest astrofizyka neutrinowa innych gwiazd. Mówi się np. o możliwości wykrywania olbrzymich strumieni neutrin tworzących się w końcowych fazach ewolucji gwiazd, tuż przed ich eksplozją jako supernowych. Przewiduje się, że detektory, czy może raczej absorbenty owych neutrin będą miały rozmiary o rzędy wielkości przewyższające niemały przecież zbiornik Davisa.

Jak można jeszcze inaczej przekonać się bezpośrednio lub prawie bezpośrednio, czy we wnętrzach gwiazd przebiegają reakcje jądrowe? Obserwując nietrwałe produkty tych reakcji trafiające na powierzchnię gwiazdy. Przykładem jest odkrycie w widmach gwiazd typu *S* linii widmowych nietrwałego pierwiastka technetu wśród pasm ZrO i linii wielu pierwiastków ciężkich. Gwiazdy typu *S* zalicza się do czerwonych olbrzymów, w których wnętrzach przypuszczalnie zachodzi proces powolnego wychwytu neutronów, prowadzący do stopniowego powstawania atomów pierwiastków coraz cięższych. W ten sposób tworzyć się może znaczna liczba trwałych jąder molibdenu $^{98}_{42}$Mo, które po wychwycie neutronu przekształcają się w cięższy izotop molibdenu, $^{99}_{42}$Mo ($t_{1/2}$ ok. 67 h). W przemianie β^{-} tworzy się z niego technet ^{99}Tc:

technet w gwiazdach typu *S*

$$^{99}_{42}\text{Mo} \rightarrow {}^{99}_{43}\text{Tc} + e^{-} + \nu_e, \qquad (22)$$

którego $t_{1/2}$ wynosi 210 tys. lat. Jest to wartość niewielka w porównaniu ze skalą życia gwiazdy. Przyjęcie hipotezy, że technet powstaje gdzieś we wnętrzu gwiazd-olbrzymów i wyniesiony zostaje na powierzchnię, jest jedynym wyjaśnieniem, skąd się on bierze w atmosferach tych gwiazd. Przypuszczalnie duże rozpowszechnienie różnych pierwiastków ze środka układu okresowego (jak Sr, Y, Zr, Ba, La, Ce, Pr) w atmosferach gwiazd typu *S* jest wynikiem takiego samego procesu.

W 1972 r. stwierdzono występowanie dwóch transuranowców, ameryku i kiuru, w gwieździe osobliwej

ameryk i kiur w gwieździe HD 25354

oznaczonej numerem katalogowym HD 25354. Najdłużej żyjące izotopy obu tych pierwiastków znane są z laboratoriów ziemskich. Oto one:
$^{243}_{95}Am$ — $t_{1/2}$ = 7950 lat, $^{247}_{96}Cm$ — $t_{1/2} \approx$ 16,4 mln lat. Te transuranowce mogły powstać jedynie w procesie szybkiego wychwytu neutronów, który się zazwyczaj wiąże ze stadium supernowej. Wątpliwe, by powstały one w tej gwieździe, w której widmie obserwuje się ich linie. Mogły jednak np. opaść na jej powierzchnię wraz z innymi produktami nukleosyntezy, wyrzuconymi w przestrzeń kosmiczną przez eksplodującą w pobliżu gwiazdę supernową. Obecność stosunkowo krótkożyjącego ameryku wskazuje na to, że proces szybkiego wychwytu neutronów musiał się odbywać niezbyt dawno.

promet w gwieździe HR 463

Zagadkę stanowi pochodzenie bardzo nietrwałego prometu w atmosferze gwiazdy osobliwej HR 465. Najdłużej żyjącymi izotopami prometu są $^{145}_{61}Pm$ ($t_{1/2}$ = 17,7 lat) i $^{146}_{61}Pm$ ($t_{1/2}$ = 5,5 lat). Skąd się bierze promet w takich ilościach, że spośród 153 znanych linii pojedynczo zjonizowanego prometu zaobserwowano w widmie tej gwiazdy aż 110? Albo jądra prometu tworzą się w gwieździe (i to na powierzchni lub niezbyt głęboko pod nią) w rezultacie reakcji jądrowych wywołanych neutronami, np.

$$^{147}_{62}Sm + n \rightarrow {}^{147}_{61}Pm + p, \qquad (23)$$

lub protonami:

$$^{145}_{60}Nd + p \rightarrow {}^{145}_{61}Pm + n, \qquad (24)$$

albo też stanowią one produkt rozszczepienia jąder pierwiastków pozauranowych. Oba wyjaśnienia wywołują dalsze pytania, np. skąd się biorą neutrony, protony o odpowiednio wysokiej energii, by się mogły wbić do jąder pierwiastka o $Z = 60$, oraz pierwiastki pozauranowe. Jeden wniosek jest niewątpliwy: obecność krótkożyjącego prometu w atmosferze gwiazdy można by wytłumaczyć tylko tym, że procesy jądrowe, w których promet powstaje, wciąż jeszcze trwają.

„skamieliny"

Bezpośrednie dowody na to, że różne ważne reakcje jądrowe dziś jeszcze przebiegają w gwiazdach, nie są może kompletne. Istnieją jednak istotne argumenty teoretyczne związane ze strukturą i ewolucją gwiazd. Na ich poparcie można przedstawić dowód rzeczowy: „skamieliny" pozostałe do dziś po owych reakcjach, czyli produkty reakcji jądrowych — różne jądra atomowe. Z dzisiejszego rozpowszechnienia pierwiastków i ich izotopów odczytać można historię materii, a historię tę w znacznym stopniu tworzyły procesy jądrowe przebiegające w gwiazdach.

D.D. Clayton *Principles of Stellar Evolution and Nucleosynthesis*, New York 1968; B. Kuchowicz *Astronomia neutrinowa Słońca*, Post. Astr. 18, 149 (1970) i 263 (1970), oraz 19, 109 (1971); B. Kuchowicz *Neutrinos from the Sun*, Rep. on Prog. in Phys. 39, 291 (1976); B. Kuchowicz *Problemy i osiągnięcia astrofizyki jądrowej. Część II. Nukleosynteza pierwiastków chemicznych jako rezultat wytwarzania energii w gwiazdach*, Post. Fiz. 22, 622 (1971); *Fizika kosmosa*, red. S. B. Pikelner), Moskwa 1976; H. Reeves *Stellar Evolution and Nucleosynthesis*, New York–London–Paris 1968.

Powstawanie pierwiastków chemicznych

Bronisław Kuchowicz

Do połowy lat czterdziestych sądzono, że pierwiastki chemiczne powstały w pewnej określonej chwili przed utworzeniem się gwiazd. Usiłowano więc wyjaśnić ich powstanie za pomocą jednego, uniwersalnego mechanizmu. Większość owych teorii uniwersalnej, kosmicznej syntezy pierwiastków (czy raczej nukleosyntezy — w odróżnieniu od syntezy chemicznej) może już dziś zainteresować jedynie historyków nauki. Przypuszczenia związane z różnymi hipotetycznymi mechanizmami powstania pierwiastków nie zgadzały się z wynikami obserwacji. Próbę czasu przetrwała tylko teoria alfa-beta-gamma, wysunięta przez R. A. Alphera, H. Bethego i G. A. Gamowa w 1948 r. Twórcy jej próbowali wyjaśnić przebieg uniwersalnej krzywej rozpowszechnienia (rys. 2 i 3 z artykułu → Rozpowszechnienie pierwiastków chemicznch i molekuł we Wszechświecie) zachodzeniem reakcji jądrowych we wczesnych fazach ewolucji Wszechświata. Obecnie, teoria ta pozwala zrozumieć, na gruncie kosmologii oraz fizyki jądra atomowego i cząstek elementarnych, jak utworzyły się we Wszechświecie pierwsze jądra atomowe. Były to jądra najlżejsze — tylko wodór i hel, z których, w kolejnych pokoleniach gwiazd, powstały i do dziś powstają dalsze pierwiastki. Procesy nukleosyntezy wciąż jeszcze trwają w przyrodzie, i wiążą się ściśle z reakcjami jądrowymi przebiegającymi w gwiazdach (→ Reakcje jądrowe w gwiazdach); (powołując się na wzory lub rysunki z tego artykułu podawać będziemy ich kolejny numer wraz z literą R).

teoria alfa–beta–gamma

Kosmiczna synteza pierwiastków

Dwa podstawowe fakty obserwacyjne (rozszerzanie się Wszechświata i wypełnienie przestrzeni kosmicznej promieniowaniem tła), w powiązaniu z ekstrapolacją znanych praw fizycznych na minione epoki, wskazują na ogromne zagęszczenie materii w pierwszych fazach ewolucji Wszechświata. Wówczas nie mogły istnieć ani atomy, ani oddzielne jądra atomowe. Nie było oczywiście i tych struktur astronomicznych (gwiazd, galaktyk), które są tak istotnymi składnikami obecnego Wszechświata. Aktualny stan fizyki cząstek elementarnych, teorii grawitacji, kosmologii i termodynamiki pozwala nam na cofnięcie w czasie do chwili, gdy gęstość materii sięgała niewyobrażalnej wartości rzędu 10^{26} g/cm^3. (Nie znamy praw opisujących materię wcześniejszą, o jeszcze większej gęstości). Była to gęstość o dziesięć rzędów wielkości przekraczająca gęstość materii jądrowej w centralnych obszarach pulsarów i w jądrach atomowych. Skład owej pierwotnej materii nadgęstej był znacznie bogatszy niż skład materii w jądrach pulsarów; występowały w niej różne nietrwałe cząstki elementarne i w równowadze z nimi cząstki o zerowej masie spoczynkowej: fotony i neutrina czyli, jak będziemy mówić, promieniowanie wysokoenergetyczne. W miarę jak Wszechświat się rozszerzał, obniżała się gęstość materii i temperatura, zmieniał się też skład materii nadgęstej. Rozpadały się krótkożyjące hiperony i mezony, zwiększał się udział nukleonów. W temperaturze ok. 10^{11} K (i przy gęstości ok. 10^{10} g/cm^3) dominującym składnikiem ylemu (jak ową materię nadgęstą nazwał Gamow) były neutrony. gdy zaś temperatura opadła do 10^9 K, protonów było już prawie siedmiokrotnie więcej niż neutronów. Nic w tym dziwnego, wraz z rozszerzaniem się Wszechświata trwał bowiem nieustanny rozpad neutronów (o których wiadomo, że w warunkach laboratoryjnych mają okres połowicznego zaniku ok. 10,8 min) na protony, elektrony i antyneutrina elektronowe:

nadgęsta pramateria

$$n \rightarrow p + e^- + \bar{\nu}_e. \qquad (1)$$

Jednocześnie protony wychwytywały neutrony, tworząc pierwsze jądra złożone — deuterony:

$$n+p \rightarrow {}^{2}_{1}D+\gamma. \quad (2)$$

Dopóki średnia energia fotonów znajdujących się w równowadze z innymi składnikami materii była większa od energii wiązania deuteronu (2,2 MeV), oddziaływanie z fotonami niszczyło powstające deuterony. Dopiero w temperaturze 10^9 K i niższej reakcja (2) prowadzi do szybkiego utworzenia deuteronów — kosztem szybko wyczerpujących się neutronów. Powstanie deuteronów umożliwia dalsze reakcje tzw. syntezy pierwotnej, odbywające się w ciągu pierwszej godziny ekspansji kosmicznej:

synteza pierwotna

$$ {}^{2}_{1}D+p \rightarrow {}^{3}He+\gamma, \quad {}^{2}_{1}D+n \rightarrow {}^{3}H+\gamma,$$

$$ {}^{2}_{1}D+{}^{3}_{1}D \rightarrow {}^{4}_{2}He+\gamma, \quad {}^{2}_{1}D+{}^{3}_{2}He \rightarrow {}^{4}_{2}He+p, \quad (3)$$

$$ {}^{2}_{1}D+{}^{3}_{1}H \rightarrow {}^{4}_{2}H+n, \quad {}^{3}_{2}He+{}^{4}_{2}He \rightarrow {}^{7}_{4}Be+\gamma.$$

W miarę ekspansji Wszechświata rozpadają się neutrony; już po upływie godziny zostaje zaledwie $^1/_{64}$ część ich początkowej ilości (faktycznie jest ich jeszcze mniej, bo ulegają wychwytom w reakcjach 2 i 3). Przeszkodę w powstawaniu kolejnych jąder przez wychwyt neutronów i protonów stanowi nieistnienie jakiegokolwiek, choćby krótkożyjącego, jądra o liczbie masowej $A = 5$; następną przeszkodą jest niezwykła nietrwałość jąder o liczbie masowej $A = 8$. Gdyby kolejne pierwiastki miały się tworzyć w reakcjach między jądrami złożonymi, w rodzaju czterech ostatnich reakcji z grupy (3), wtedy w kolejnych reakcjach musiałaby być pokonywana coraz to wyższa bariera kulombowska. Tymczasem z upływem czasu Wszechświat stygnie, energia kinetyczna cząstek stale się obniża, tak że ciągu reakcji (3) nie da się już kontynuować.

Choć przy obliczaniu wydajności syntezy pierwotnej w rozszerzającym się Wszechświecie wzięto pod uwagę wszystkie możliwe reakcje jądrowe między kilkudziesięcioma lekkimi jądrami, okazało się, że w syntezie tej nie powstają właściwie jądra o liczbie masowej większej niż 7. Oto typowe wyniki pierwotnej syntezy kosmicznej podane w procentach wagowych składu otrzymanej materii: ^{3}He — ok. 0,00004%, ^{4}He — ok. 30%, ^{7}Li — ok. 0,00004%, jądra o liczbie masowej $A \geqslant 12$ — ok. 0,000006%, reszta — wodór ^{1}H. Wyniki liczbowe różnią się nieco, zależnie od szczegółów rozważanej wersji modelu rozszerzającego się Wszechświata (np. ułamek wagowy wytworzonego helu ^{4}He może się wahać od 0,25 do 0,40), ale zasadnicze znaczenie ma to, że w syntezie pierwotnej mogły powstać jedynie wodór i hel oraz lit (ten ostatni głównie z przemiany β^- wytworzonego berylu ^{7}Be). Większa część obserwowanego dziś w gwiazdach helu — to nie hel powstały w ich wnętrzach, lecz produkt procesów jeszcze dawniejszych, gdy gwiazdy nie istniały. Trzy najlżejsze pierwiastki chemiczne — to „skamieliny" pochodzące z ery radiacyjnej ewolucji Wszechświata (→ Kosmologia).

powstanie jąder najlżejszych

Powstawanie pierwiastków chemicznych jako produkt uboczny wytwarzania energii w gwiazdach

Gwiazda czerpie energię z egzoenergetycznych reakcji termojądrowych, z wyjątkiem krótkotrwałych faz kontrakcji grawitacyjnej (następujących po wypaleniu określonego rodzaju paliwa jądrowego w centrum).

Gwiazdy pierwszego pokolenia powstały z materii zawierającej produkty kosmicznej syntezy pierwotnej. Najdłużej trwający etap ich ewolucji — to spalanie wodoru w cyklu *pp* (materia nie zawierała w swoim składzie jąder-katalizatorów cyklu CNO). Zależnie od masy gwiazdy (→ Ewolucja gwiazd), po wypaleniu wodoru, mógł przebiegać dłuższy lub krótszy ciąg procesów termojądrowych, w których z jąder lżejszych tworzyły się jądra coraz to cięższe. Ale nawet w gwiazdach o największej masie nie mogły powstać tym sposobem jądra atomowe o liczbach masowych większych niż 60. W egzoenergetycznych reakcjach termonuklearnych jądra słabiej związane nie mogą się tworzyć z jąder silniej związanych.

spalanie wodoru

Spalanie wodoru w gwieździe zwiększa zawartość helu. Następny pod względem ważności proces termonuklearny to spalanie helu, podczas którego gwiazda opuszcza ciąg główny i zatacza charakterystyczne pętle na wykresie Hertzsprunga-Russella w obszarze olbrzymów (→ Ewolucja gwiazd, rys. 2), przechodząc przez fazy niestabilności pulsacyjnej. Jest wtedy nawet możliwa powolna utrata materii przez gwiazdę.

spalanie helu

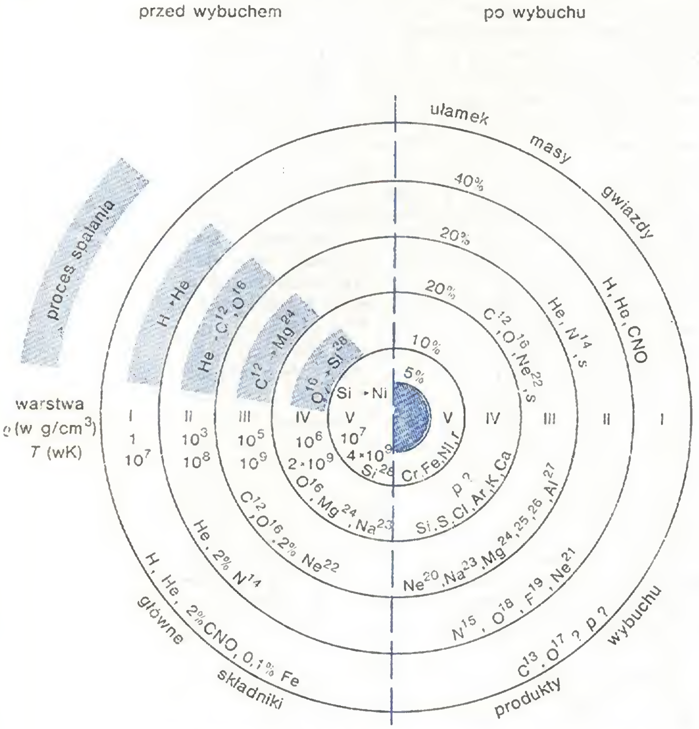
W procesie spalania helu tworzy się nie tylko najpospolitszy izotop węgla ^{12}C w reakcjach (R-8) i (R-9), ale i najpospolitsze izotopy tlenu i neonu. Jądra ^{16}O powstają przez przyłączenie dalszej cząstki α w reakcji (R-11) do już wytworzonego jądra węgla ^{12}C; analogicznie, z tlenu ^{16}O tworzy się neon ^{20}Ne. Powyższe procesy nukleosyntezy odbywają się w stadium olbrzymów, przez które przechodzi znaczna część gwiazd. Fakt ten tłumaczy większe rozpowszechnienie jąder „alfowych" ^{12}C, ^{16}O i ^{20}Ne niż jąder z nimi sąsiadujących. Jest też zrozumiałe, dlaczego tlen, węgiel i neon zajmują trzecie, czwarte i szóste miejsce w kolejności rozpowszechnienia (str. 982 rys. 2). Poza tym widać, dlaczego na uniwersalnej krzywej rozpowszechnienia występują charakterystyczne minima lokalne dla trzech lekkich pierwiastków: litu, berylu i boru. Nie mogą one powstać z helu, który spala się od razu na węgiel i dalsze pierwiastki (parzyste!). Nie mogą też powstać w większych ilościach podczas spalania wodoru, mimo że powstają np. w gałęziach bocznych cyklu *pp*, mają jednak duże przekroje czynne dla różnych procesów z tego cyklu i powstają jedynie przejściowo podczas przemiany wodoru w hel.

Produkty spalania helu w centrum gwiazdy mogą same „wstąpić" w reakcje jądrowe postaci (R-23 — tabela 2) i (R-24 — tabela 2), jeśli tylko masa gwiazdy jest wystarczająco duża, by pod wpływem sił grawitacyjnych mogło nastąpić kolejne zagęszczenie i podgrzanie jądra gwiazdy. Analogiczna sytuacja może się powtarzać jeszcze kilkakrotnie, po zużyciu kolejnego paliwa w jądrze gwiazdy. Do zapłonu następnego paliwa dochodzi tym łatwiej, im większa jest masa gwiazdy, przy czym stale rośnie temperatura i gęstość w centrum. W reakcjach między dwoma jądrami mogą powstawać coraz to bardziej różnorodne produkty końcowe. W reakcjach spalania węgla (R-23 — tabela 2) mieliśmy aż pięć końcowych możliwości; zwróćmy uwagę na pojawienie się swobodnych neutronów i fotonów. W dalszych fazach spalania termonuklearnego powstaje i oddziałuje z sobą coraz to większa liczba różnych rodzajów jąder, a także swobodne neutrony, protony i kwanty γ, wreszcie zbliżamy się do równowagi między nimi, i w składzie mieszaniny jąder zaczynają przeważać te, które są najtrwalsze. Takimi najtrwalszymi jądrami są jądra żelazowców. Zespół procesów, przebiegających w warunkach równowagi, nazwano procesem *e* (z ang. *equilibrium* — równowaga). Wydaje się, że proces ten jest końcową fazą aktywnego okresu ewolucji gwiazdy. Faza ta przebiega bardzo szybko i kończy się wybuchem rozrzucającym w przestrzeń dużą część produktów tego procesu, a także procesów poprzednich, których produkty zawarte są w bardziej zewnętrznych warstwach gwiazdy (rys. 1). Dziś nazwa procesu *e* obejmuje nie jakiś hipotetyczny, odrębny od innych procesów jądrowych w gwieździe mechanizm powstawania jąder żelazowców, lecz jest raczej skrótem, pod który podstawiamy wiele już poznanych i zlokalizowanych w różnych obiektach procesów spalania pierwiastków o liczbach atomowych $Z \geqslant 14$ (a więc krzemu, siarki, wapnia itd.), w których powstają jądra coraz to cięższe aż do żelazowców.

proces *e*

Maksimum rozpowszechnienia żelazowców (str. 983, rys. 3) jest lokalne, a nie bezwzględne. Stąd wynika, że tylko niewielka część materii we Wszech-

świecie przeszła przez proces *e*. Mimo że skład chemiczny materii zmienia się w określonym kierunku, oraz że stale wzrasta udział pierwiastków cięższych od wodoru i helu, to niepodona sobie wyobrazić, by nawet za sto miliardów lat rozpowszechnienie żelazowców było największe ze wszystkich pierwiastków.



Rys. 1. Struktura i skład chemiczny masywnej gwiazdy ($10\ M_{\odot} \leqslant M \leqslant 60\ M_{\odot}$) tuż przed wybuchem (po lewej) oraz skład chemiczny materii po wybuchu (po prawej). Dla stanu przed wybuchem podano (po lewej stronie): u góry — reakcje stacjonarnego spalania przebiegające w kolejnych warstwach; pośrodku — gęstość średnią i temperaturę każdej z tych warstw; u dołu — skład chemiczny warstwy. Dla stanu po wybuchu podano (z prawej strony): u góry — procent masy gwiazdy, zawarty w danej warstwie (wyrzucany w przestrzeń), a także ważniejsze nuklidy, które przeżyły wybuch; u dołu — nuklidy wytworzone w danej warstwie podczas wybuchu oraz odbywające się podczas niego procesy.
Obszar ciemnoniebieski w środku oznacza ewentualną pozostałość po gwieździe (pulsar, czarna dziura). Litery *p* i *r* wskazują na odpowiednie procesy nukleosyntezy, litera *s* na obecność produktów procesu *s*

ewolucja gwiazd II pokolenia

Naszkicowany schemat powstawania pierwiastków w trakcie ewolucji gwiazd I pokolenia można przenieść — z pewną modyfikacją — na dalsze pokolenia gwiazd, uwzględniając przede wszystkim zmianę składu chemicznego materii międzygwiazdowej, o której przyjmuje się, że powstają z niej kolejne pokolenia gwiazd. Różnego typu niestabilności w późnych stadiach ewolucji gwiazd prowadzą do powolnej lub gwałtownej utraty masy (np. wybuch gwiazdy supernowej), zatem ośrodek międzygwiazdowy wzbogaca się stopniowo w pierwiastki coraz cięższe. Gwiazdy drugiego pokolenia mogą już zawierać pierwiastki chemiczne aż do żelazowców włącznie. Jeśli gwiazda drugiego pokolenia ma dość dużą masę, to w jej wnętrzu zawierającym jądra węgla, wodór będzie się spalał w cyklu CNO. Stosunek ilościowy jąder-katalizatorów ustala się wtedy zgodnie z wyrażeniem (R-13), czyli 95% początkowej liczby jąder węgla ^{12}C przechodzi w jądra azotu ^{14}N, 1% tej liczby — w jądra rzadszego izotopu węgla ^{13}C. Duże rozpowszechnienie azotu we Wszechświecie (piąte miejsce) jest wynikiem procesów spalania wodoru w gwiazdach II i następnych pokoleń.

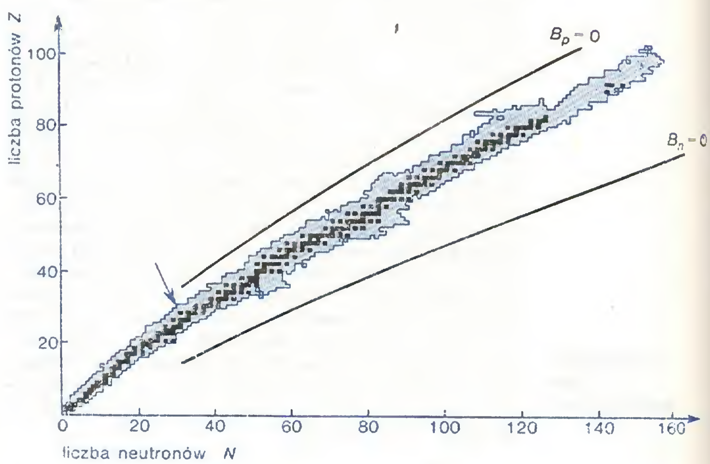
Istotną cechą jąder żelazowców są duże przekroje czynne na reakcje typu (n, γ). Zatem w gwieździe II pokolenia neutrony powstające w reakcjach (R-23–24 — tabela 2) i innych mogą wytwarzać z jąder żelazowców jądra jeszcze cięższe.

Powstawanie pierwiastków ciężkich w reakcjach wychwytu neutronów

tablica nuklidów

Rozważanie procesów wychwytu neutronów ułatwia tablica nuklidów — graficzne zestawienie wszystkich znanych trwałych i nietrwałych odmian jąder atomowych. Tablicę taką w zmniejszeniu przedstawia rys. 2. Czarne kwadraciki oznaczają jądra trwałe i długożyjące, linia łamana otacza obszar, w którym znajdują się znane jądra promieniotwórcze. Dwie linie grube odpowiadają jądrom, w których energia wiązania ostatniego neutronu bądź ostatniego protonu jest równa zeru ($B_n = 0$ lub $B_p = 0$); obejmują one obszar, wewnątrz którego mogą istnieć jądra atomowe. Czarne kwadraciki łączą się z ścieżką trwałości (→ Jądra atomowe i ich wzbudzenia), która dzieli obszar istnienia jąder atomowych na dwie części. W części górnej (między linią $B_p = 0$ a ścieżką trwałości) jądra promieniotwórcze mają na ogół za dużo protonów. Przemiana β^+ albo wychwyt elektronu sprowadzają jądra takie bliżej ścieżki. W części dolnej (między linią $B_n = 0$ a ścieżką trwałości) jądra promieniotwórcze zawierają za dużo neutronów. Przemiana β^-, w której wyniku jeden z neutronów w jądrze zamienia się w proton, przesuwa takie jądro skośnie w lewo ku górze, zbliżając je ku ścieżce trwałości od dołu.

Począwszy od liczby masowej $A = 50$ trwałe jądra atomowe mają zazwyczaj duże przekroje czynne (rzędu barna i większe) na wychwyt neutronu (n, γ). Wyjątkowo małe przekroje charakteryzują jedynie niektóre jądra magiczne. Strzałka na rys. 2 wskazuje na



Rys. 2. Tablica nuklidów w schematycznym ujęciu. Z liczba protonów, N liczba neutronów

trwałe jądra żelazowców — końcowe produkty łańcucha reakcji termojądrowych, stanowiące zarazem punkt wyjściowy dla dalszej nukleosyntezy. Ponieważ procesy wychwytu neutronów zaczynają się od jąder — zarodzi z grupy żelazowców, wystarczy w dalszych rozważaniach ograniczyć się do tej części tablicy nuklidów, która zawiera jądra o liczbie protonów $Z \geqslant 25$ i liczbie neutronów $N \geqslant 30$ (rys. 3). Wychwyt neutronów, zaczynający się od trwałych jąder-zarodzi, prowadzi zawsze na prawo od ścieżki trwałości, tj. w obszar jąder promieniotwórczych mających za dużo neutronów. Nawet jeśli po kilku pierwszych wychwytach neutronów jądro wychwytujące pozostanie trwałe, to i tak, w rezultacie któregoś z kolei wychwytu, powstanie jądro wykazujące promieniotwórczość β

powolny i szybki wychwyt neutronów

W poprzednim artykule rozważaliśmy sytuację, w której z trwałego jądra-zarodzi $^{56}_{26}Fe$ dopiero po wychwycie trzeciego neutronu utworzyło się nietrwałe jądro $^{59}_{26}Fe$. W przemianie β^- (R-20) powstało z niego trwałe jądro kobaltu $^{59}_{27}Co$, zdolne do wychwytu czwartego neutronu. Zwróćmy uwagę na wartość liczbową okresu połowicznego zaniku nietrwałego jądra $^{59}_{26}Fe$: 45 dni. Fakt, że jądro to zdążyło się rozpaść, i że dopiero trwały produkt jego przemiany β wychwycił czwarty neutron, dowodzi powolności procesu wychwytu. Inaczej mówiąc, średni odstęp czasu t_n między kolejnymi wychwytami neutronu przez to samo jądro jest dosyć duży. Oczywiście odstęp ten musi być większy niż okres połowicznego zaniku jądra $^{59}_{26}Fe$ ze względu na przemianę β^-. Ale w innych przypadkach czas t_n może wcale nie być większy od t_β — okresu połowicznego zaniku ze względu na rozpad β^- powstających w tych procesach jąder. Ze względu na relacje między t_n i t_β przebiegający zespół reakcji i rozpadów stanowi łącznie albo proces *s* (z ang. *slow* — powolny), albo proces *r* (z ang. *rapid* — szybki):

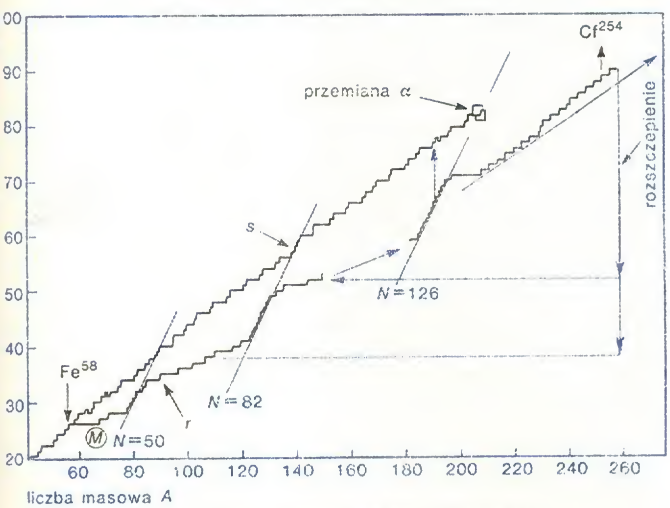
1) Gdy $t_n \gg t_\beta$, to wychwyt kolejnych neutronów odbywa się stosunkowo powoli. Neutrony są wychwytywane przez jądro trwałe, gdyż jeśli powstanie jądro nietrwałe, to i tak zdąży się rozpaść przed wychwytem następnego neutronu. Jest to proces *s*. Jego ilustracją jest omówiony wyżej przykład.

2) Gdy $t_n \ll t_\beta$, wtedy wychwyt odbywa się szybko. Odstęp czasu między kolejnymi aktami wychwytu neutronów przez to samo jądro jest tak krótki, że jądro nie zdąży się rozpaść, dopóki trwa wychwyt. Sytuacja ta odpowiada procesowi *r*.

Każdy z tych procesów przebiega w charakterystycznych dla siebie warunkach w odpowiedniej fazie ewolucji gwiazdy.

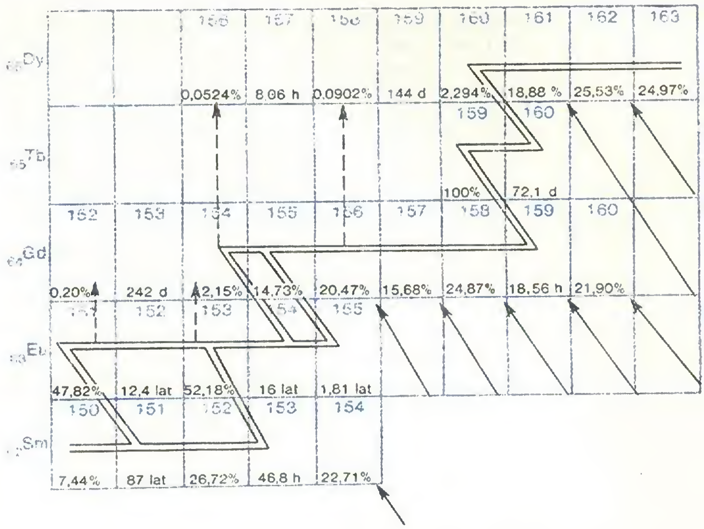
proces *s*

Nukleosynteza pierwiastków ciężkich w procesie *s* — to jak gdyby powolne wdrapywanie się jąder-zarodzi z lewego dolnego rogu na rys. 3 wzdłuż ścieżki trwałości w górę na prawo, ku coraz wyższym wartościom liczby masowej *A* i liczby porządkowej *Z*. Podczas trwania tego procesu jądra atomowe nie oddalają się od ścieżki trwałości o więcej niż jedną lub



Rys. 3. Schemat ogólny przebiegu procesów *s* i *r*

dwie wartości liczby neutronów *N*. Gdy tylko się od niej oddalą, przemiana β^- sprowadza je na nią z powrotem. Proces ten, schematycznie przedstawiony linią łamaną „*s*" na rys. 3, ilustrujemy powiększonym fragmentem tablicy nuklidów, zawierającym odcinek ścieżki trwałości między liczbami masowymi 151 i 161. Najlżejszym nuklidem na rysunku jest samar $^{150}_{62}Sm$. W rezultacie wychwytu neutronu tworzy się z niego nietrwały $^{151}_{62}Sm$, który może przed swym rozpadem wychwycić następny neutron, przechodząc



Rys. 4. Fragment tablicy nuklidów ze ścieżką procesu *s*(=)) oraz zaznaczeniem procesów *r* (strzałki ukośne) i *p* (strzałki pionowe i poziome -→ i ⇝). Liczba neutronów wzrasta od strony lewej ku prawej, liczba protonów — ku górze. Dla nuklidów trwałych podano rozpowszechnienie (zawartość procentowa poszczególnych izotopów danego pierwiastka), dla nietrwałych — okresy połowicznego zaniku w stanie podstawowym

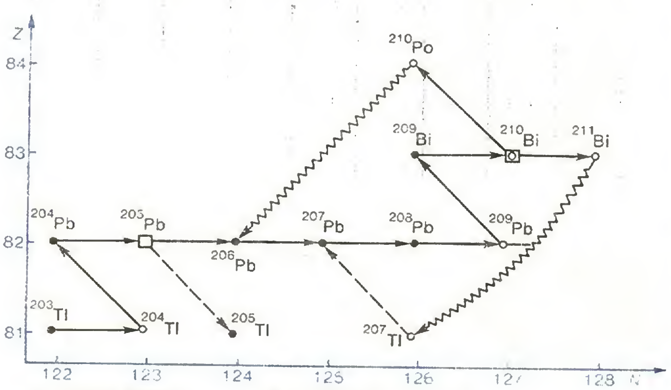
w trwały $^{152}_{62}Sm$. Jeśli natomiast jądro $^{151}_{62}Sm$ zdąży się rozpaść przed wychwytem neutronu, to dalszy neutron zostanie wychwycony już przez trwały europ $^{151}_{63}Eu$. Dla jądra $^{151}_{62}Sm$ należy uwzględnić dwa możliwe procesy — przemianę β^- i reakcję (n, γ) — gdyż dla niego nie jest spełniony warunek $t_n \gg t_\beta$ (natomiast oba charakterystyczne czasy są wielkościami niemal tego samego rzędu); w miejscu tym szlak procesu *s* rozgałęzia się. Na rys. 4 widać jeszcze jedno rozgałęzienie procesu *s*.

Rozgałęzienia nie zmieniają w istotny sposób przebiegu procesu *s*. Jego cechą charakterystyczną jest pojawianie się przez dłuższy czas (setki tysięcy a nawet miliony lat) niewielkiej liczby neutronów, tak że średnio na jedno jądro-zarodź przypada wychwyt neutronu raz na kilka miesięcy albo kilka do kilkudziesięciu czy nawet kilkuset lat. Proces *s* nie wytwarza jednak wszystkich izotopów pierwiastków ciężkich. Trwałe izotopy pierwiastków parzystych, które znajdują się najdalej na prawo na rys. 4, nie mogą w nim powstać. Tak więc, np. najcięższy izotop gadolinu, $^{160}_{64}Gd$, nie może się wytworzyć z lżejszego izotopu gadolinu przez wychwyt neutronu, gdyż $^{159}_{64}Gd$ jest jądrem o niezwykle krótkim okresie połowicznego zaniku i po powstaniu w procesie *s* przechodzi przez przemianę β^- w trwały izotop terbu ($^{159}_{65}Tb$). Z terbu $^{159}_{65}Tb$ po wychwycie neutronu i przemianie β^- tworzy się dysproz $^{160}_{66}Dy$, nie może natomiast powstać gadolin $^{160}_{64}Gd$. Szlak procesu *s* nie przebiega także przez najlżejsze izotopy pierwiastków parzystych, w rodzaju gadolinu $^{152}_{64}Gd$ czy $^{156}_{66}Dy$ z rys. 4.

Jak wynika z tego rysunku w procesie *s* nie mogą powstać ani najcięższe, ani też najlżejsze izotopy niektórych pierwiastków parzystych. Również nie mogą powstać nietrwałe pierwiastki ciężkie, tworzą się jedynie pierwiastki trwałe (tj. takie, które mają przynajmniej jeden trwały izotop). Ostatnim pierwiastkiem trwałym w układzie okresowym jest bizmut, który ma tylko jeden trwały izotop $^{209}_{83}Bi$. Obszar niezwykle krótkożyjących pierwiastków między bizmutem a torem nie pozwala na przedłużenie szlaku procesu *s* do nietrwałych wprawdzie, ale długożyjących izotopów uranu i toru, tworzących tak zwaną pierwszą wyspę trwałości (oczywiście względnej), leżącą na przedłużeniu ścieżki trwałości. Przyczyną tego jest przemiana α — dominujący sposób rozpadu jąder nietrwałych cięższych od bizmutu. W wyniku tej przemiany

pierwsza wyspa trwałości

następuje „zawracanie" powstających jąder ciężkich w obszar jąder lżejszych, wytworzenie zamkniętej pętli, kończącej proces *s*. Spójrzmy na rys. 5. Jeśli trwałe jądro $^{209}_{83}$Bi wychwyci neutron, wtedy powstaje $^{210}_{83}$Bi w jednym z dwóch możliwych stanów izomerycznych. Ze stanu o okresie połowicznego zaniku ok. 5 dni jądro to przechodzi przez przemianę β^- w znany emiter cząstek α, polon $^{210}_{84}$Po ($t_{1/2} \approx$ 138 dni). W przemianie α tworzy się z niego ołów $^{206}_{82}$Pb. Jeśli $^{210}_{83}$Bi powstanie w stanie długożyjącym ($t_{1/2}$ = 2,6 mln lat), wtedy jeszcze przed rozpadem zdąży wychwycić następny neutron. W ten sposób powstałe jądro $^{211}_{83}$Bi



• – nuklidy trwałe □ – nuklidy długo żyjące
o – nuklidy krótko żyjące, ⟶ (poziomo) reakcja (n, γ)
⟵ (skośnie w lewo) przemiana β^-
⟵ – – (skośnie w prawo) wychwyt elektronu
⟵∿∿ (skośnie w dół) rozpad α

Rys. 5. Przebieg końcowej części procesu *s*

($t_{1/2}$ = 2,15 min przez przemianę α przechodzi w tal $^{207}_{81}$Tl, to zaś jądro przez przemianę β^- przechodzi w ołów $^{207}_{82}$Pb. Przyłączanie neutronów przez bizmut tworzy aż dwie pętle obiegu zamkniętego pod koniec szlaku procesu *s*. Choćby proces ten trwał miliony lat, w rezultacie powolnego dołączania neutronów do jąder o liczbie masowej $A \leqslant 210$ można z jąder tych otrzymać na koniec co najwyżej jądra ołowiu i bizmutu, nigdy jednak nie uda się „przepompować" ich do obszaru pozabizmutowego.

proces r

„Przepompowanie" takie jest natomiast możliwe w procesie *r*. W rezultacie neutronizacji zagęszczającego się wnętrza gwiazdy przez reakcje typu (R-21) powstają ogromne ilości swobodnych neutronów, stężenie ich sięga wartości 10^{25} neutronów w 1 cm³ i większych. Część tych neutronów, uchodząca na zewnątrz, napromieniowuje jądra-zarodzie z wyrzucanych w przestrzeń warstw zewnętrznych. Na jedno takie jądro może nawet przypadać kilkaset neutronów. Jeśli znów wystartujemy z lewego dolnego narożnika rys. 3, gdzie znajdują się trwałe jądra żelazowców, to w ciągu pierwszych paru sekund jądro-zarodź wychwyci kolejno tyle neutronów, że przesunie się aż pod granicę obszaru, w którym mogą jeszcze istnieć jądra, a więc energia wiązania ostatniego neutronu w tym jądrze będzie wielkością rzędu 1–2 MeV. Położenie jądra odpowiadać będzie miejscu tuż przy krzywej $B_n = 0$ z rys. 2. Dopóki jądro się tak bardzo nie przesunęło w prawo, wciąż spełniony był warunek $t_n \ll t_\beta$. Jednak w owym skrajnym położeniu (które odpowiada miejscu oznaczonemu literą *M* na rys. 3) okres połowicznego zaniku przemiany β^- dla owego jądra staje się wielkością porównywalną z odstępem czasu między kolejnymi wychwytami neutronu, a obie te wartości nie przekraczają ułamków sekundy. Do jądra w tym miejscu nie da się już dołączyć dalszego neutronu, zanim nie nastąpi przemiana β^-. Produkt tej przemiany znów będzie mógł przyłączyć jeden lub parę neutronów w czasie rzędu ułamków sekundy, zanim stanie się tak krótkożyjący, że zdąży doznać następnej przemiany β^- przed wychwytem kolejnego neutronu. Sytuacja przypomina w rezultacie przebieg procesu *s* z pewną istotną różnicą. Szlak procesu *r* oddala się najpierw od ścieżki trwałości i przybliża do hipotetycznej linii $B_n = 0$, by biec następnie wzdłuż tej ostatniej, mniej więcej równolegle do szlaku procesu *s*. Na uwagę zasługuje stromy kształt przebiegu szlaku procesu *r* przy liczbach magicznych neutronów N = 50, 82 i 126. Wiąże się to z trudnością dołączania w tych miejscach kolejnego neutronu do zamkniętej powłoki.

Podczas gdy szlak procesu *s* biegnie na przemian przez jądra trwałe i β^--promieniotwórcze, szlak procesu *r* prowadzi przez jądra wyłącznie nietrwałe, o znikomych okresach połowicznego zaniku. Gdy proces *r* się skończy (a trwa on zapewne nie więcej niż kilka minut poczas eksplozji gwiazdy), wszystkie jego produkty są nietrwałe i dopiero po pewnej liczbie kolejnych przemian β^- mogą się stać jądrami trwałymi. Ową stabilizację produktów procesu *r* zaznaczyliśmy na rysunku 3 linią przerywaną dla jednego jądra; nie sposób zrobić to dla wszystkich produktów procesu *r*, gdyż wtedy rysunku stałby się bardzo nieczytelny.

Jeśli tylko strumień neutronów w procesie *r* był dostatecznie duży, wtedy mogły powstać jądra o dużych wartościach liczby masowej, być może nawet $A \approx 300$. Z badań laboratoryjnych nad pierwiastkami transuranowymi wiemy, że im cięższe jest jądro, tym łatwiej ulega samorzutnemu rozszczepieniu, jak i rozszczepieniu pod działaniem neutronów. Oba te zjawiska nie pozwalają na przedłużanie szlaku procesu *r* ku dowolnie dużym wartościom liczb masowych. Ekstrapolacja danych laboratoryjnych pozwala przypuszczać, że w wyniku rozszczepienia (samorzutnego i wywołanego przez neutrony) proces *r* kończy się gdzieś w pobliżu liczby masowej $A \approx 300$. Na skutek rozszczepienia tworzą się jądra lżejsze — fragmenty rozszczepienia; to zawracanie fragmentów rozszczepienia w środkowe rejony tablicy nuklidów (zaznaczone strzałkami na rys. 3) stanowi odpowiednik pętli kończącej proces *s*.

W wyniku procesu *r* i następujących po nim przemianach β^- można dotrzeć w pobliże pierwszej wyspy trwałości. Tak też mogły powstać pierwiastki transuranowe, znacznie krócej żyjące niż uran i tor. Nie dotrwały one na Ziemi do dziś, mogą jednak występować w promieniowaniu kosmicznym (powstającym w eksplozjach gwiazd supernowych, podczas których tworzą się produkty procesu *r*), a może i gdzie indziej w kosmosie, gdzie proces *r* nie tak dawno się odbył.

druga wyspa trwałości

Proces *r* może doprowadzić do powstania pewnej grupy jąder z tzw. drugiej wyspy trwałości, w pobliżu hipotetycznego jądra podwójnie magicznego o Z = = 114 i N = 184. Istnienie takiej wyspy przewidują od lat teoretycy (→ Jądra w stanie ekstremalnym). Hipotetyczne pierwiastki superciężkie z tej wyspy byłyby oczywiście nietrwałe, mogłyby jednak być na tyle długożyjące, że warto ich poszukiwać, jeśli nie w materii ziemskiej, to wśród produktów reakcji jądrowych w laboratorium, także w pierwotnym promieniowaniu kosmicznym pochodzącym z eksplozji odległych gwiazd.

Produkty procesu *r* po jego zakończeniu zbliżają się do ścieżki trwałości na rys. 3 skośnie w górę na lewo. Kierunki kilku końcowych przemian β są także zaznaczone na rys. 4. Łatwo zauważyć, że w procesie *r* tworzą się najcięższe izotopy pierwiastków parzystych, które jak $^{154}_{62}$Sm, i $^{160}_{64}$Gd z rys. 4 nie mogą powstać w procesie *s*. Produktami procesu *r* jest też znaczna część tych jąder trwałych, przez które przebiega szlak procesu *s*. Ale niektóre jądra powstające w procesie *s*, jak np. $^{154}_{62}$Gd czy $^{160}_{66}$Dy z rys. 4, nie mogą być produktami procesu *r*. Są one przed procesem *r* „osłaniane" przez swoje izobary o niższym ładunku, a więc odpowiednio przez $^{154}_{62}$Sm i $^{160}_{64}$Gd. Wreszcie najlżejsze izotopy pierwiastków parzystych, które nie mogą powstać w procesie *s*, tym bardziej nie będą tworzyć się w procesie *r*.

Procesy nukleosyntezy trzeciego rzędu

Pozostały jeszcze dwie nieliczne grupy jąder o niewielkim stosunkowo rozpowszechnieniu, których powstanie nie wiąże się ani z procesami pierwszego rzędu — egzoenergetycznymi reakcjami termonuklearnymi, ani też z procesami drugiego rzędu — wychwytem neutronów. Pierwszą grupę tworzą tzw. jądra pominięte — najlżejsze izotopy niektórych parzystych pierwiastków ciężkich. Rozpowszechnienie każdego z nich jest małe (patrz np. wartości liczbowe dla $^{152}_{64}Gd$, $^{156}_{66}Dy$ i $^{158}_{66}Dy$ na rys. 4), co wskazuje na powstanie ich w jakimś procesie o niezbyt wysokim prawdopodobieństwie. Proces ten nazwano kiedyś procesem *p*, przypuszczano bowiem, iż w wyniku jednej lub dwóch reakcji przyłączenia protonu, tj. reakcji (p, γ), jądra te mogą powstać np. z produktów procesu *s* (niekiedy nawet i z produktów procesu *r*). Proces *p* byłby więc procesem trzeciego rzędu. Zasadniczą trudnością w analizie tego procesu jest wyjaśnienie, skąd biorą się w gwieździe protony o tak wielkiej energii, że mogą przejść przez wysoką barierę kulombowską wokół jąder ciężkich.

proces p

Powstanie jąder pominiętych znacznie łatwiej wyjaśnić, odwołując się do reakcji fotojądrowych, które mogłyby przebiegać w końcowych fazach ewolucji niektórych gwiazd po osiągnięciu wysokiej temperatury rzędu kilku milionów stopni. W warunkach tych pojawiłaby się ogromna ilość kwantów γ, zdolnych do wywołania całego ciągu reakcji (γ, n), (γ, p) i (γ, α) w obecnych w materii gwiazdnej produktach wychwytu neutronów. Na rys. 4 podano jako przykład, w jaki sposób w wyniku dwóch kolejnych reakcji (γ, n) jądro $^{160}_{66}Dy$ przechodzi w jądro pominięte $^{158}_{66}Dy$. Chociaż obecnie przeważa pogląd, że jądra pominięte powstają raczej w reakcjach fotojądrowych niż przez wychwyt protonów, to nadal utrzymuje się nazwa procesu *p*; odzwierciedla ona to, że proces *p* prowadzi do wytworzenia jąder ciężkich, o możliwie największym udziale protonów.

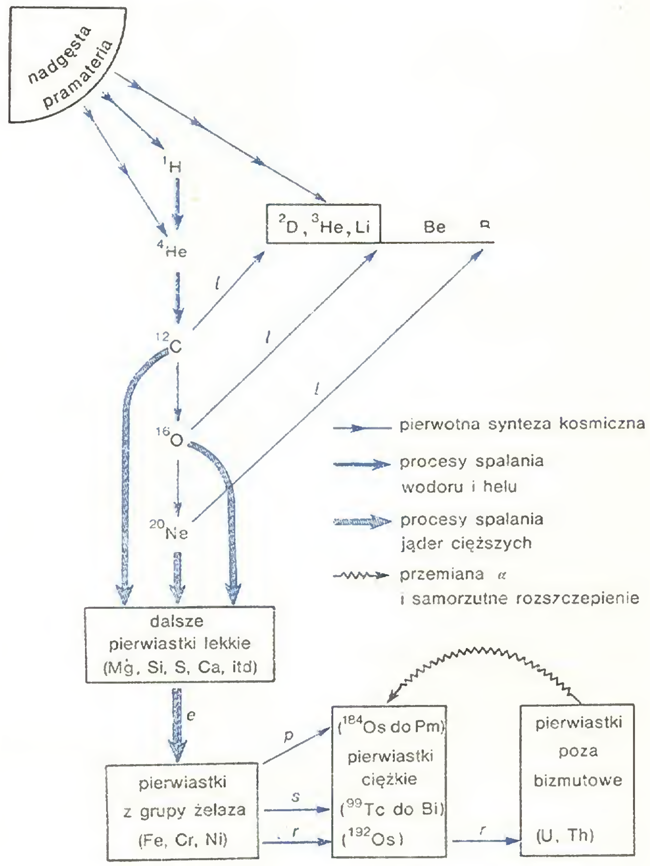
Następna grupa jąder o niewielkim rozpowszechnieniu, która nie tworzy się ani w reakcjach termojądrowych, ani poprzez wychwyt neutronów — to jądra trzech pierwiastków lekkich: litu, berylu i boru. Do grupy tej zalicza się niekiedy dwa rzadkie izotopy dwu najlżejszych pierwiastków: $^{2}_{1}D$ i $^{3}_{2}He$. Wszystkie jądra z tej grupy są bardzo nietrwałe ze względu na reakcje termonuklearne. Choć np. deuter, lekki izotop helu ^{3}He i lit mogły już powstać w niewielkich ilościach podczas kosmicznej syntezy pierwotnej, w zasadzie ulegają one całkowitemu wypaleniu we wnętrzu gwiazdy jeszcze przed rozpoczęciem spalania wodoru. Podczas spalania wodoru jądra z tej grupy odgrywają tylko rolę produktów pośrednich, stąd na krzywej rozpowszechnienia pierwiastków mamy wyraźne minimum lokalne dla litu, berylu i boru, a na krzywej rozpowszechnienia izobarów dołączają się minima dla $A = 2$ i $A = 3$. Występuje natomiast charakterystyczne podwyższenie zawartości owych pierwiastków w pierwotnym promieniowaniu kosmicznym w takim stopniu, że wspomniane minimum w promieniowaniu tym praktycznie nie występuje.

Tworzenie się jąder wspomnianej grupy nazwano krótko procesem *l* (z ang. *light* — lekki). Jest to jedyny proces nukleosyntezy, którego miejscem jest przede wszystkim przestrzeń kosmiczna. Produkty nukleosyntezy wyrzucane podczas eksplozji gwiazd w przestrzeń ulegają zderzeniom z jądrami najbardziej rozpowszechnionego pierwiastka, wodoru. Są to zderzenia przy niezwykle wielkich energiach, jeśli więc już nastąpią, jądra cięższe ulegną pokruszeniu na fragmenty lżejsze. Gdy symulowano takie zderzenia w laboratoriach, stosując wysokoenergetyczne protony z akceleratorów do rozbijania jąder tlenu, węgla, neonu i azotu (pierwiastków od trzeciego do szóstego w kolejności rozpowszechnienia), obfitymi produktami kruszenia okazały się jądra ^{2}D, ^{3}He oraz izotopy litu, berylu i boru. Wszystkim tym jądrom nie „grozi" w przestrzeni kosmicznej wypalenie w procesach termojądrowych, stąd też stężenie ich w materii międzygwiazdowej może być większe niż w gwiazdach. Dodajmy, że analogiczne procesy kruszenia przebiegać mogą w warstwach powierzchniowych niektórych gwiazd, gdzie kruszenie mogą powodować np. protony rozpędzone w zmiennym polu magnetycznym albo fala uderzeniowa.

proces l

Za ostatni proces nukleosyntezy trzeciego rzędu, modyfikujący skład chemiczny materii otrzymanej z innych procesów, można uważać naturalne przemiany promieniotwórcze. W ich wyniku w skorupie ziemskiej wciąż powstają krótkożyjące pierwiastki pozabizmutowe (np. astat lub frans czy bardziej znany rad) z długożyjącego uranu i toru. Może dzięki samorzutnemu rozszczepieniu transuranowców na powierzchni gwiazdy osobliwej HR 465 wciąż na nowo tworzą się nader krótkożyjące jądra prometu?

naturalne przemiany promieniotwórcze



Rys. 6. Schemat nukleosyntezy pierwiastków chemicznych we wczesnej fazie ewolucji rozszerzającego się Wszechświata oraz w gwiazdach i przestrzeni kosmicznej. W kilku miejscach umieszczono typowe produkty zaznaczonych na rysunku procesów, spotykane w przyrodzie. Trzy typy procesów w gwiazdach zaznaczone są odmiennymi typami linii: a) procesy pierwszego rzędu — reakcje egzoenergetyczne prowadzące do wytwarzania energii (strzałki pionowe), b) procesy wychwytu neutronów (strzałki poziome), c) procesy trzeciego rzędu (strzałki biegnące ukośnie)

Rysunek 6 stanowi zaktualizowaną wersję schematu nukleosyntezy, wysuniętego w 1957 r. przez twórców współczesnej astrofizyki jądrowej: małżonków E. M. i G. Burbidge'ów, F. Hoyle'a i W. A. Fowlera.

R. J. TAYLER *The Origin of the Chemical Elements*, London 1972 (ros. Moskwa 1975); B. KUCHOWICZ *Problemy i osiągnięcia astrofizyki jądrowej. Część III. Powolny wychwyt neutronów w gwiazdach*, oraz *Część IV. Szybki wychwyt neutronów i procesy trzeciorzędowe nukleosyntezy*, Post. Fiz. 24, 669 (1973) i 25, 475 (1974); B. KUCHOWICZ, *Kosmochemia*, Warszawa 1980.

Skorowidz

N

O

P

R

S

Ś

Ź

Ż



1. Kryształy granatów otrzymane metodą Czochralskiego: b), d) i f) prawidłowe kształty kryształów, a) i c) widoczny wpływ zmiany temperatury tygla w czasie i asymetrii jej rozkładu

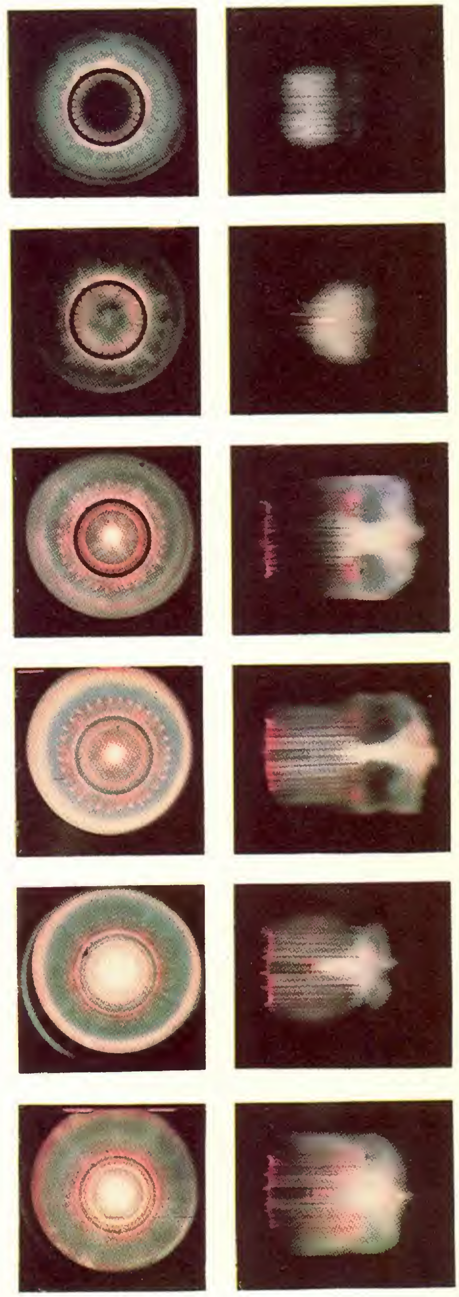
2. Pierścienie barwne obserwowane w wyniku wymuszonego rozpraszania Ramana w benzenie (otrzymane przez R. W. Terhune'a)

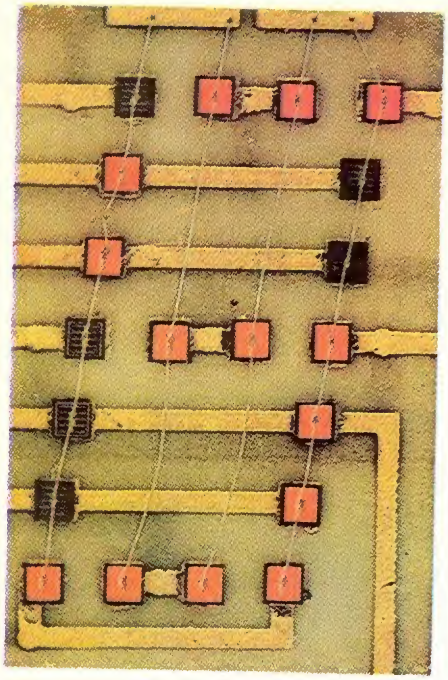




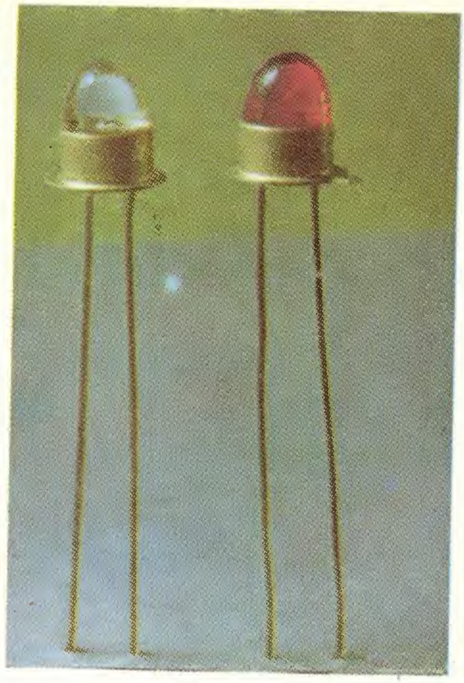
3. Kryształ kwarcu otrzymany metodą hydrotermiczną (oświetlenie światłem niebieskim)

4. Kolejne fazy powstawania plazmy w iniektorze prętowym fotografowane ultraszybką kamerą (500 tys./s) wzdłuż osi iniektora i z boku





5. Wnętrze tablicy z DEL do wyświetlania cyfr przez włączenie do obwodu odpowiednich diod. Świecące diody tworzą tu cyfrę 5 (pow. 20 ×)

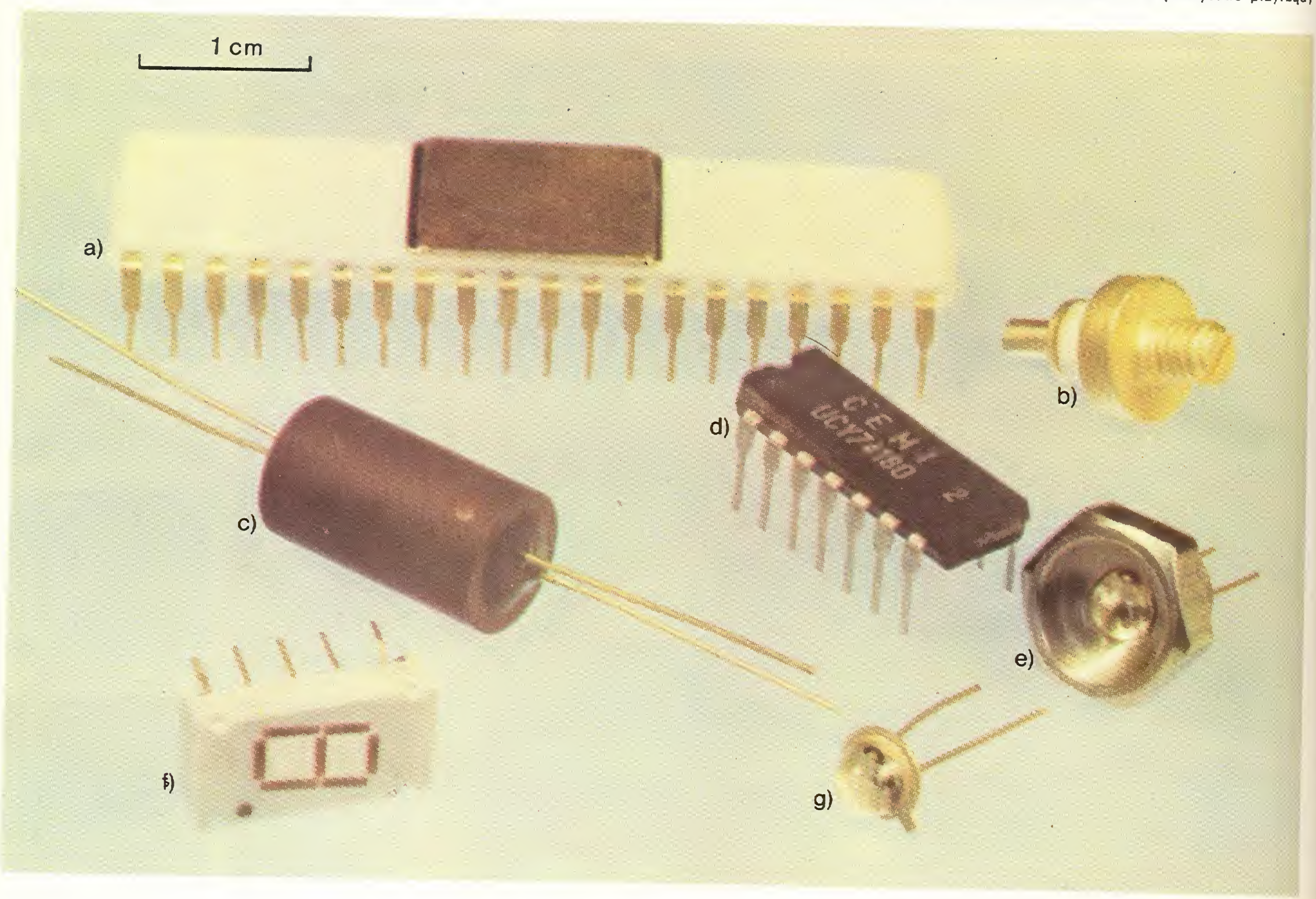


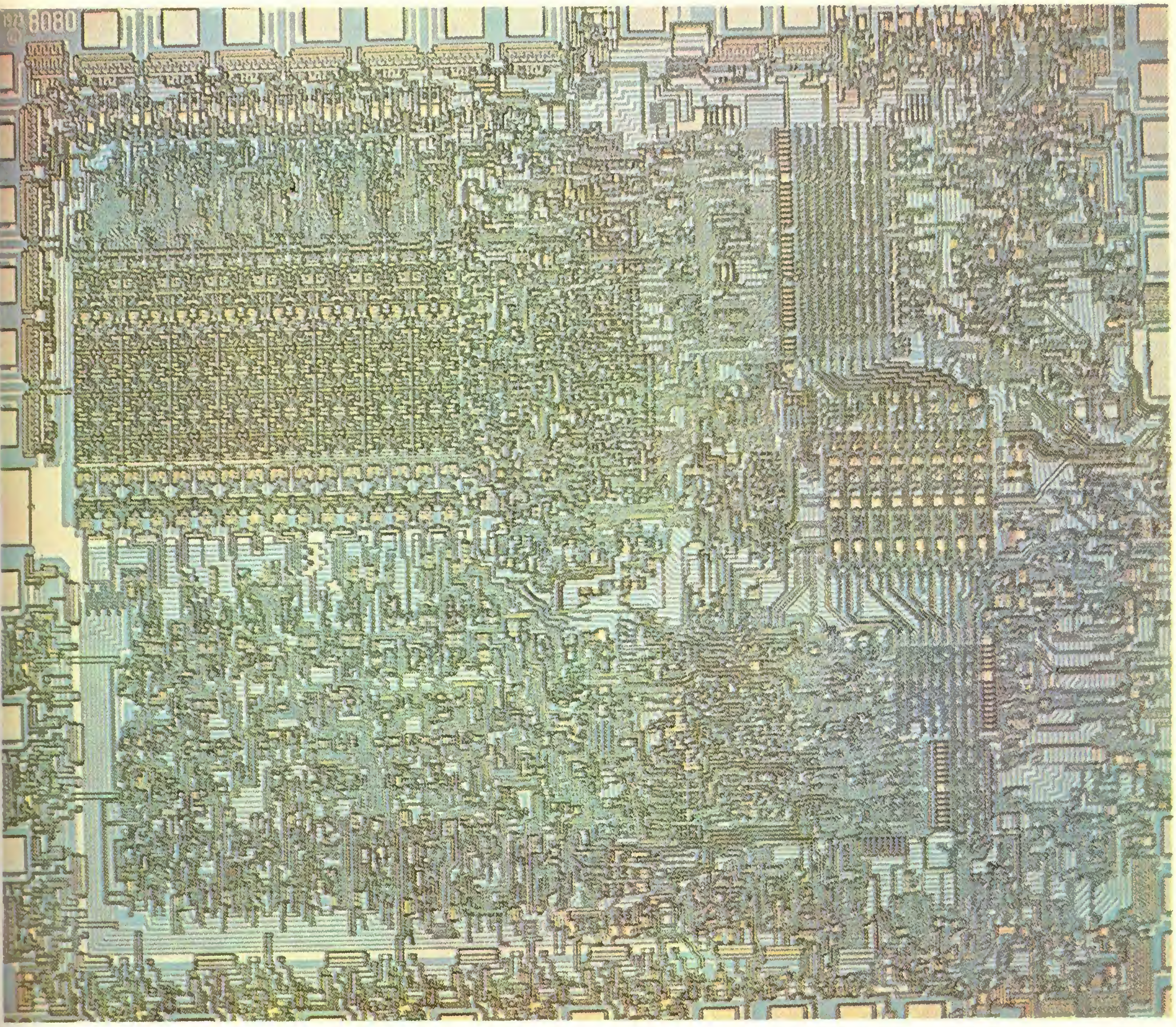
6. Diody elektroluminescencyjne (DEL) CQYP 31 i CQYP 21 (pow. ok. 2 ×; opracowane w ITE)



7. Termistor NTC 230 (pow. ok. 2 ×; opracowany w ITE)

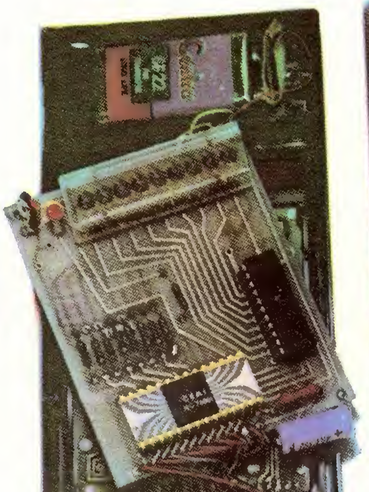
8. Przykłady przyrządów półprzewodnikowych: a) monolityczny układ scalony wielkiej skali integracji, b) dioda mikrofalowa, c) transoptor, d) układ scalony średniej skali integracji, e) dioda elektroluminescencyjna emitująca podczerwień, f) wskaźnik cyfrowy, g) dioda elektroluminescencyjna emitująca światło czerwone (wszystkie przyrządy wykonano w Polsce)

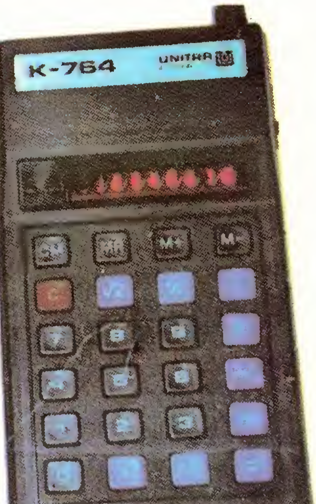




9. Mikrofotografia struktury układu scalonego mikroprocesora 8080 firmy Intel (słowo 8-bitowe, cykl rozkazowy 2 µs, 78 rozkazów). Pow. ok. 40 razy

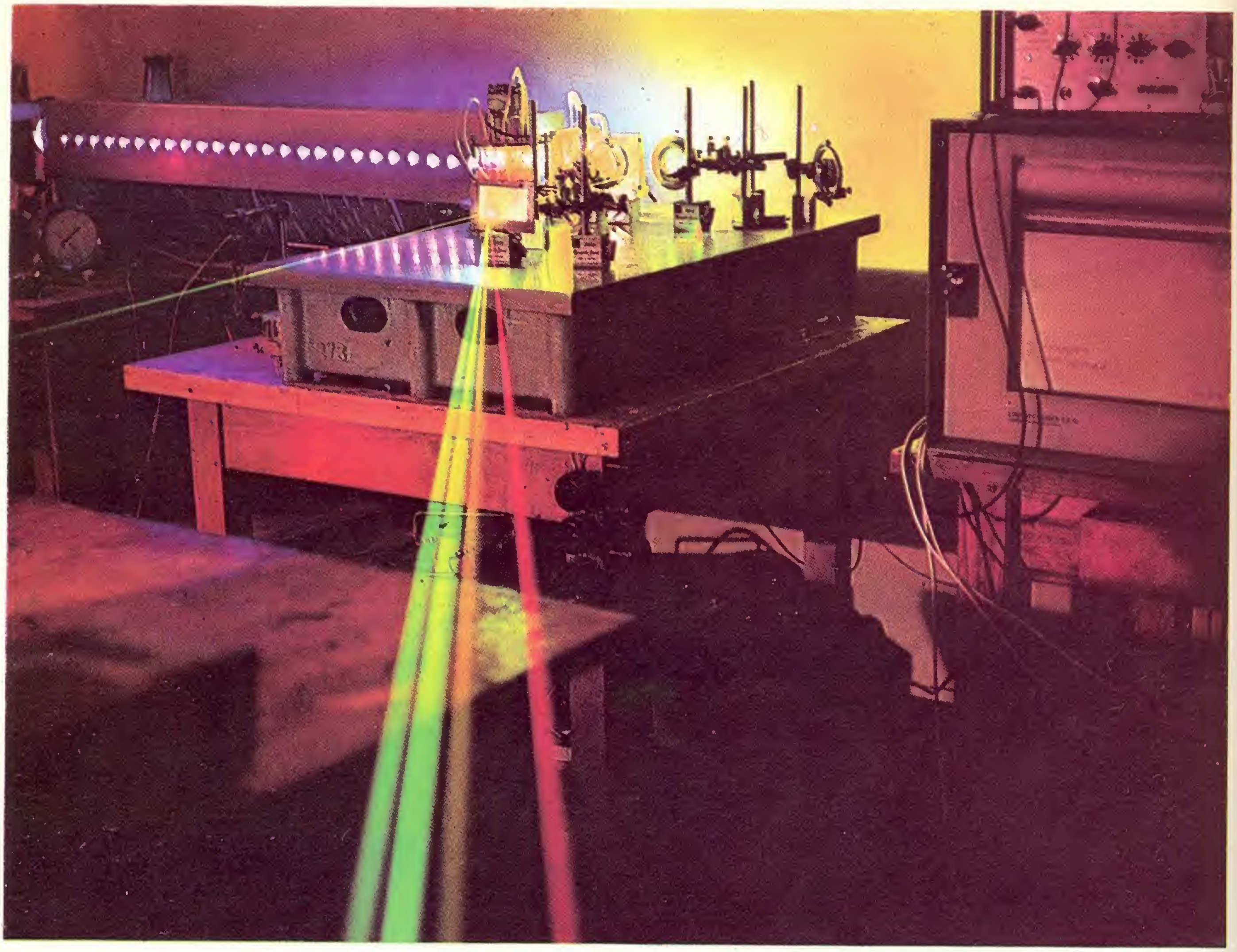




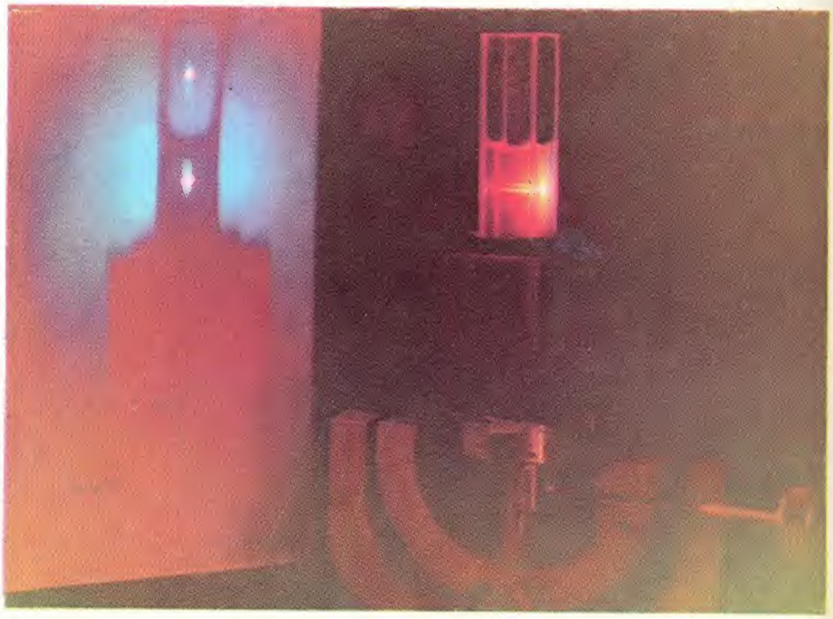


10. Zegarek elektroniczny firmy Unitra zbudowany na układzie scalonym MC X 1201 i układach optoelektronicznych opracowanych w Instytucie Technologii Elektronowej

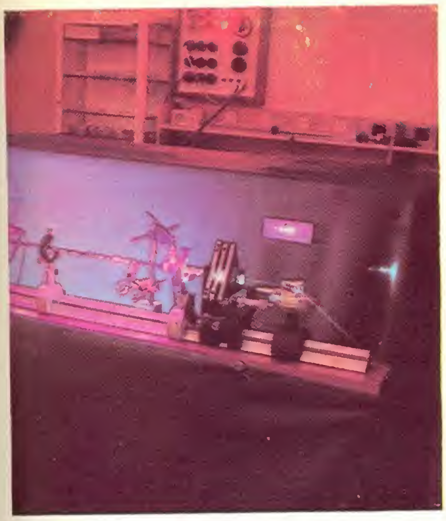
11. Minikalkulator firmy Unitra zbudowany na układzie scalonym MC 74007 i układach optoelektronicznych opracowanych w Instytucie Technologii Elektronowej



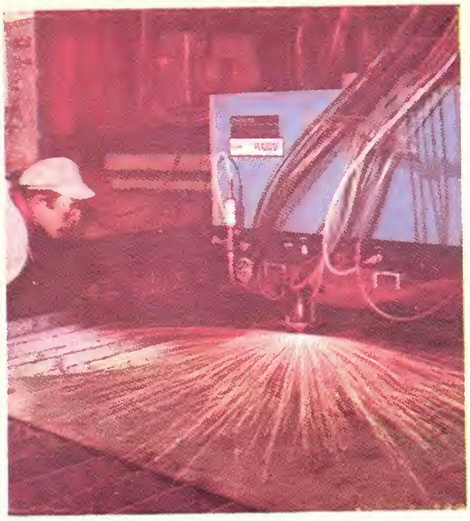
12. Laser barwnikowy zbudowany w Zakładzie Optyki IFD Uniwersytetu Warszawskiego pompowany laserem azotowym. Zdjęcie wykonano techniką wieloekspozycyjną, by pokazać przestrajanie lasera barwnikowego. Laser azotowy: moc 2 MW, czas błysku ok. 12 ns, częstość powtarzania 30/s. Laser barwnikowy: rodamina 6G, moc 500 kW, długość fali od 560 do 640 nm. Zestaw używany jest do badań spektroskopowych



13. Powstanie akcji laserowej w silnie wzmacniającym barwniku organicznym (rodamina B) po wzbudzeniu silnym impulsem (w zakresie nadfioletu) z lasera azotowego. Akcja laserowa rozwija się mimo małej dobroci rezonatora, który stanowią równoległe płytki kwarcowe — ścianki kuwety. (Górna plamka na ekranie powstała przez dodatkowe odbicie)



14. Laser kadmowy w układzie do analizy widmowej



15. Cięcie blachy za pomocą lasera CO_2

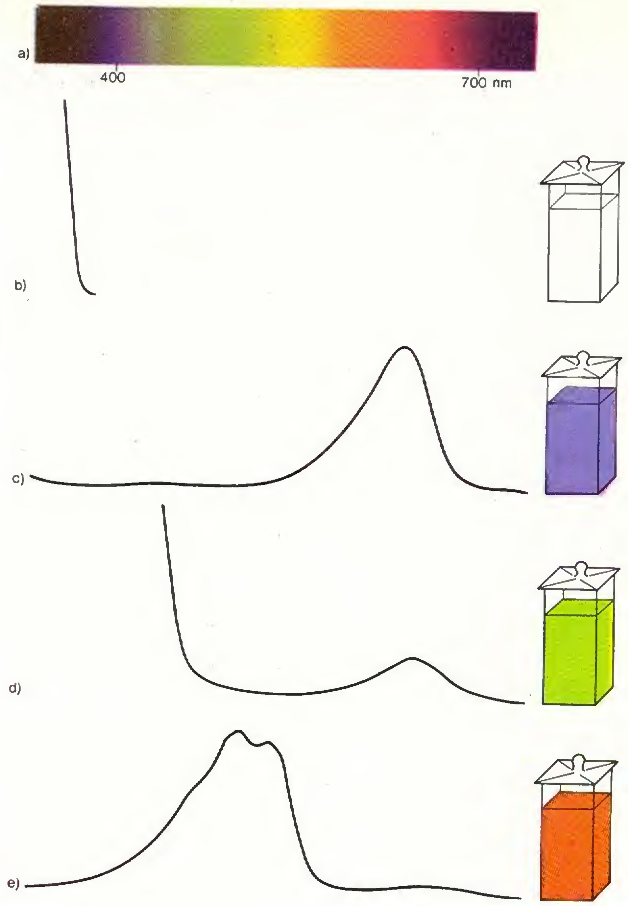


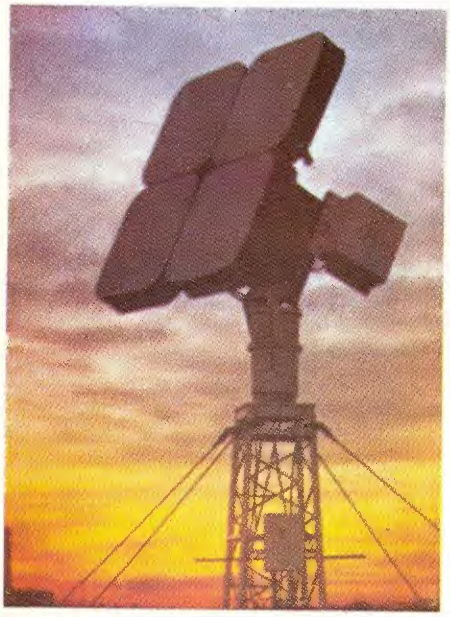
16. Koagulator laserowy



17. Cięcie kwarcu za pomocą lasera CO_2

18. Związek pomiędzy zabarwieniem substancji i jej widmem. Widmo promieniowania elektromagnetycznego z wąskiego obszaru promieniowania widzialnego (światła) ilustruje rys. a. Jeżeli do oka trafia całe promieniowanie z zakresu od ok. 360 do ok. 700 nm, to tę mieszaninę barw oko rejestruje jako barwę białą (światło białe). Widmo absorpcyjne (elektronowe) substancji bezbarwnej, tzn. przepuszczającej całe promieniowanie z zakresu widzialnego, wygląda tak jak na rys. b. Jeżeli próbka absorbuje promieniowanie z pewnej części obszaru widzialnego, to promieniowanie to nie dociera do oka i próbka wykazuje zabarwienie zależne od tego, z jakiego obszaru promieniowanie zostało zaabsorbowane, tak jak to ilustrują rys. c, d i e





19. Radiosonda meteorologiczna

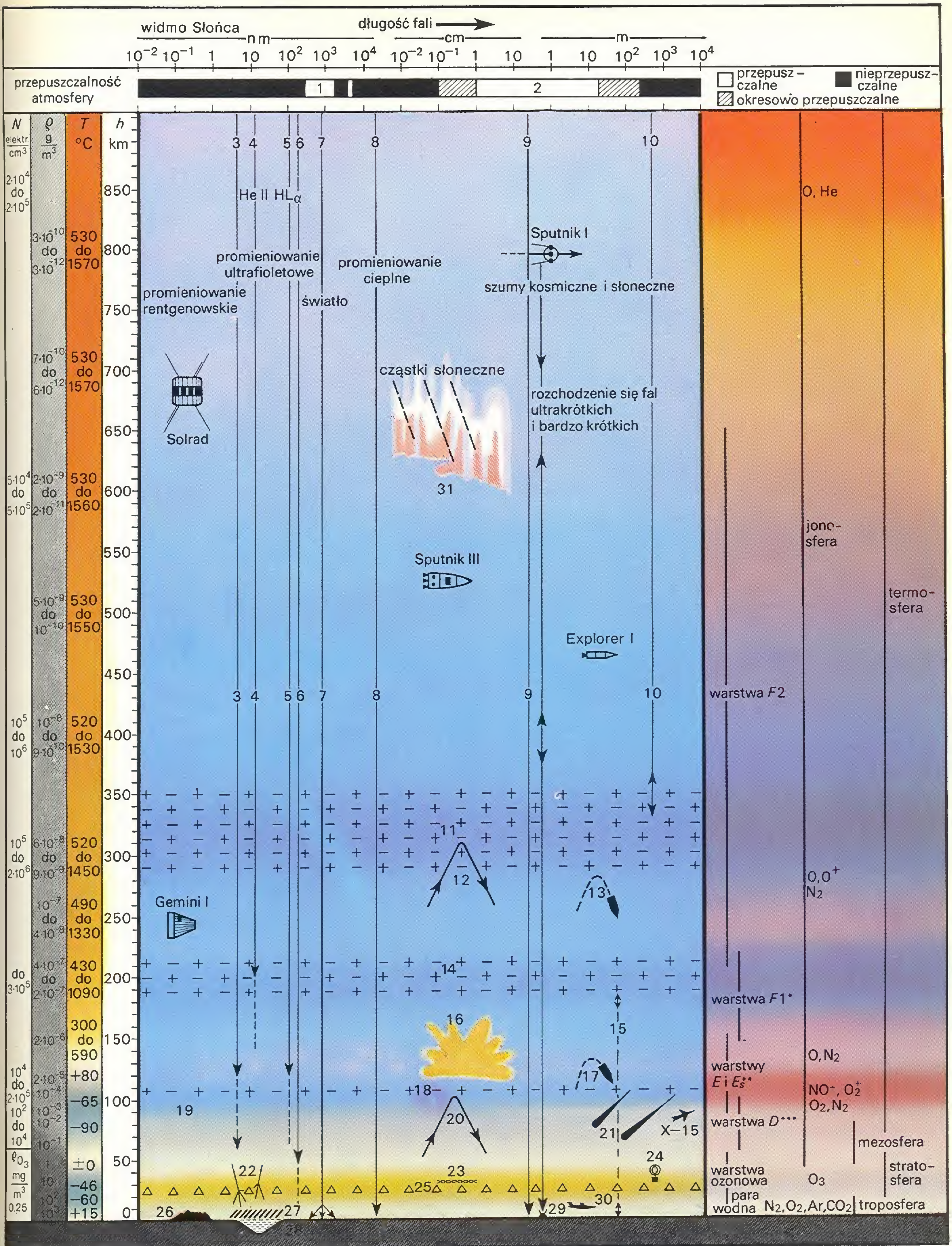


20. Zorza polarna

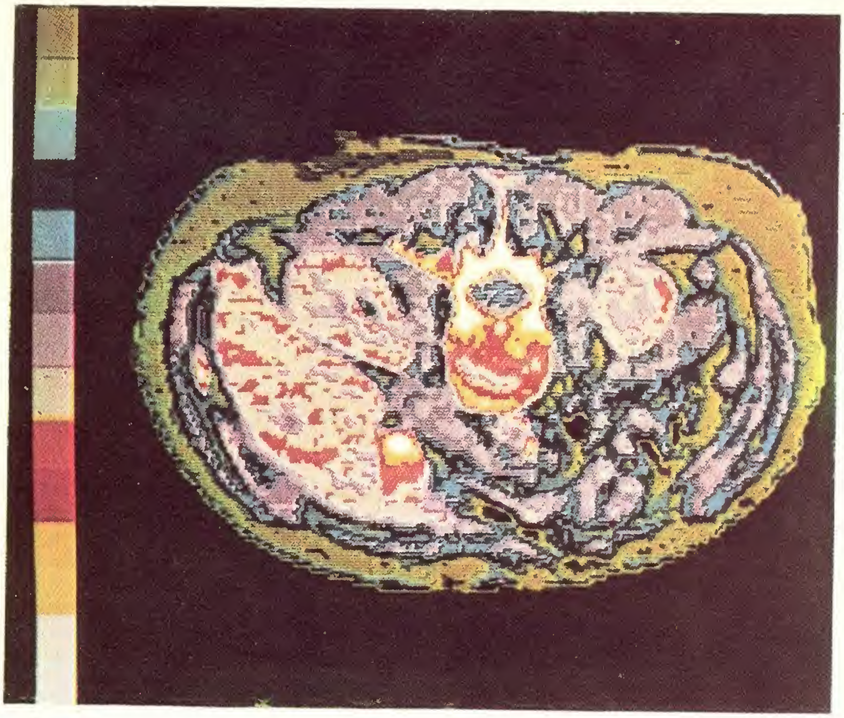
21. Przekrój przez atmosferę Ziemi do wysokości 900 km. Trzy główne obszary atmosfery — troposfera, stratosfera i termosfera — zaznaczone są różnymi odcieniami. Intensywność żółtego koloru proporcjonalna jest do gęstości ozonu, a intensywność koloru niebieskiego do koncentracji elektronów. W prawej kolumnie różnymi kolorami zaznaczone są różne składniki atmosfery. Lewa kolumna: N koncentracja elektronów, ϱ gęstość gazu, ϱ_{O_3} gęstość ozonu, T temperatura, h wysokość. Prawa kolumna: z lewej strony — nazwy różnych warstw jonosferycznych, pośrodku — główne składniki chemiczne, po prawej stronie — nazwy obszarów w atmosferze.

1 okno astronomiczne,
2 okno radioastronomiczne,
3 słoneczne promieniowanie rentgenowskie,
4 linia helu (He II) Słońca (strzałki oznaczają miejsca lub obszary w atmosferze, gdzie promieniowanie o danej długości fali jest pochłaniane),
5 linia wodoru (L_α) Słońca,
6 promieniowanie ultrafioletowe,
7 promieniowanie widzialne (400–750 nm),
8 promieniowanie podczerwone,
9 szumy kosmiczne i słoneczne,
10 szumy kosmiczne i słoneczne odbite od jonosfery,
11 maksimum koncentracji elektronów,
12 odbicie krótkich fal radiowych od jonosfery,
13 wielostopniowe rakiety,
14 warstwa $F1$,
15 sondowanie jonosfery,
16 zorze polarne,
17 jednostopniowe rakiety,
18 warstwy D i E,
19 świecące obłoki,
20 odbicie fal krótkich od dolnej jonosfery (warstwa E tylko w dzień: w nocy od warstwy E_s)
21 meteory,
22 wielkie pęki promieniowania kosmicznego,
23 chmury perłowe,
24 balony stratosferyczne,
25 maksimum gęstości ozonu,
26 Mount Everest,
27 chmury cirrusy,
28 Rów Mariański,
29 antena paraboliczna do łączności z satelitami (tutaj Sputnik I),
30 samoloty,
31 zorze polarne wzbudzone cząstkami słonecznymi,

* tylko w dzień, latem, na średnich szerokościach geograficznych,
** warstwa E tylko w dzień,
*** warstwa D tylko w dzień

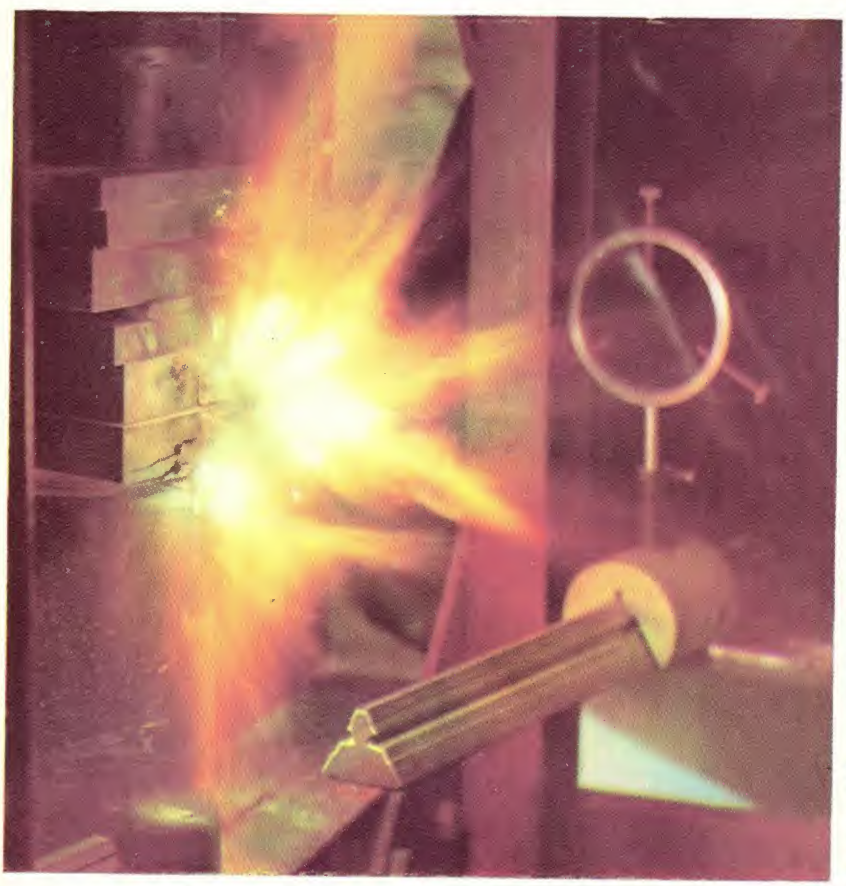


widmo Słońca
długość fali
nm
cm
m
przepuszczalność atmosfery
przepuszczalne
nieprzepuszczalne
okresowo przepuszczalne
N elektr/cm3
ρ g/m3
T °C
h km
He II HLα
promieniowanie ultrafioletowe
promieniowanie cieplne
promieniowanie rentgenowskie
światło
Sputnik I
szumy kosmiczne i słoneczne
cząstki słoneczne
rozchodzenie się fal ultrakrótkich i bardzo krótkich
Solrad
Sputnik III
Explorer I
Gemini I
X-15
O, He
jonosfera
termosfera
warstwa F2
O,O+ N2
warstwa F1*
warstwy E i Es**
O, N2
NO+, O2+ O2, N2
warstwa D***
mezosfera
warstwa ozonowa
O3
stratosfera
para wodna
N2, O2, Ar, CO2
troposfera



22. Obraz przekroju jamy brzusznej człowieka otrzymany za pomocą tomografu komputerowego. Podana skala barw odpowiada kolejnym przedziałom wartości współczynnika pochłaniania promieni rentgenowskich (kolor biały odpowiada wartości największej, czarny — najmniejszej). Biały obszar w środku zdjęcia to kręgosłup, duży jasny obszar po prawej stronie — wątroba, a biała plama na tym obszarze — pęcherzyk żółciowy, dwa okrągłe jaśniejsze miejsca po obu stronach kręgosłupa to nerki

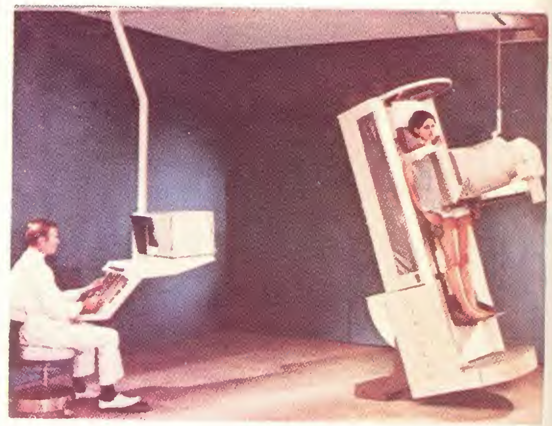
23. Pola magnetyczne o indukcji powyżej 100 T wytwarza się metodami eksplozyjnymi. Zdjęcie przedstawia moment rozerwania cewki (Pracownia Impulsowych Pól Magnetycznych, Instytut Fizyki PAN, Warszawa)





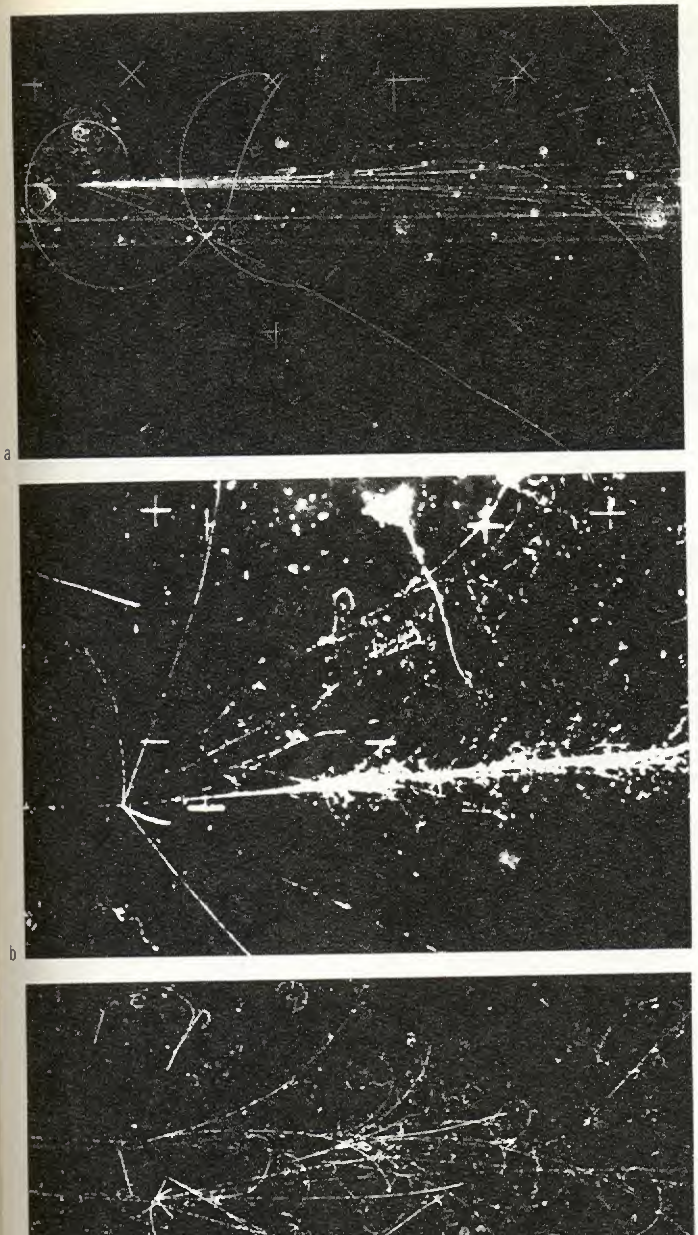
24. Termogram powierzchni dłoni wykonany przy użyciu cholesterolowych ciekłych kryształów

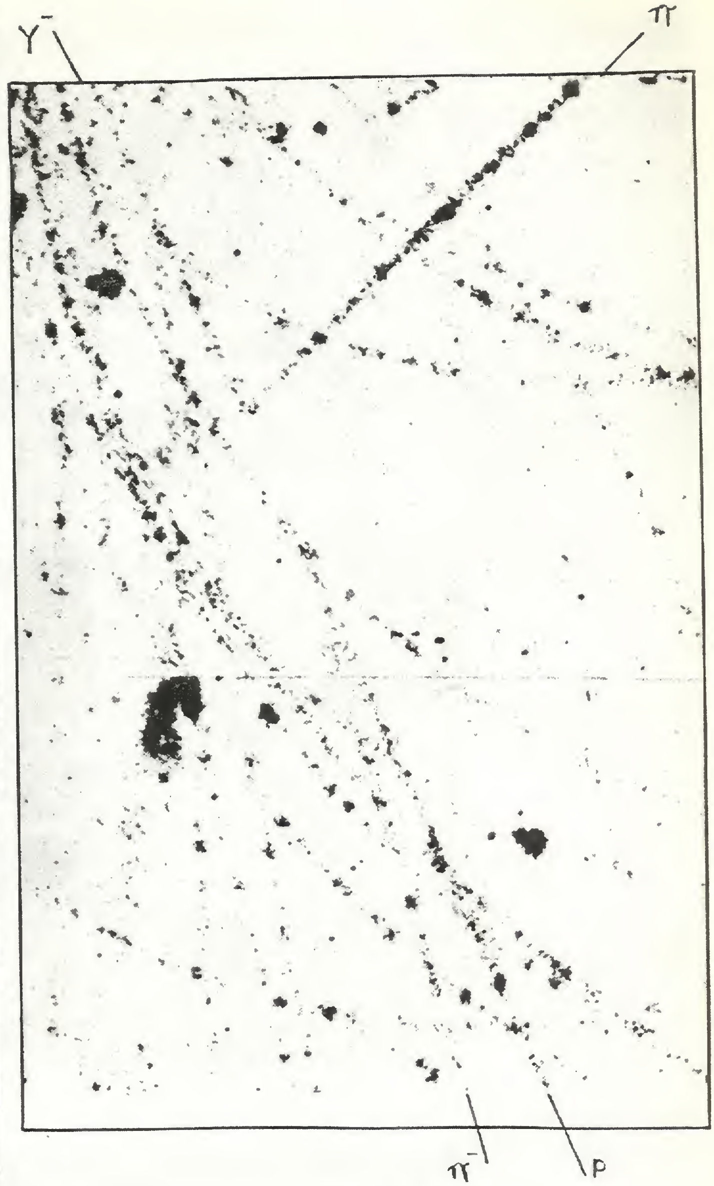
25. Nowoczesny aparat do rentgenodiagnostyki. Zastosowanie wzmacniacza obrazu pozwala na zmniejszenie napromieniowania pacjenta i oddzielenie stanowiska lekarza do pacjenta



26. Hydrofon w toni morskiej (firmy Brüel i Kjaer)





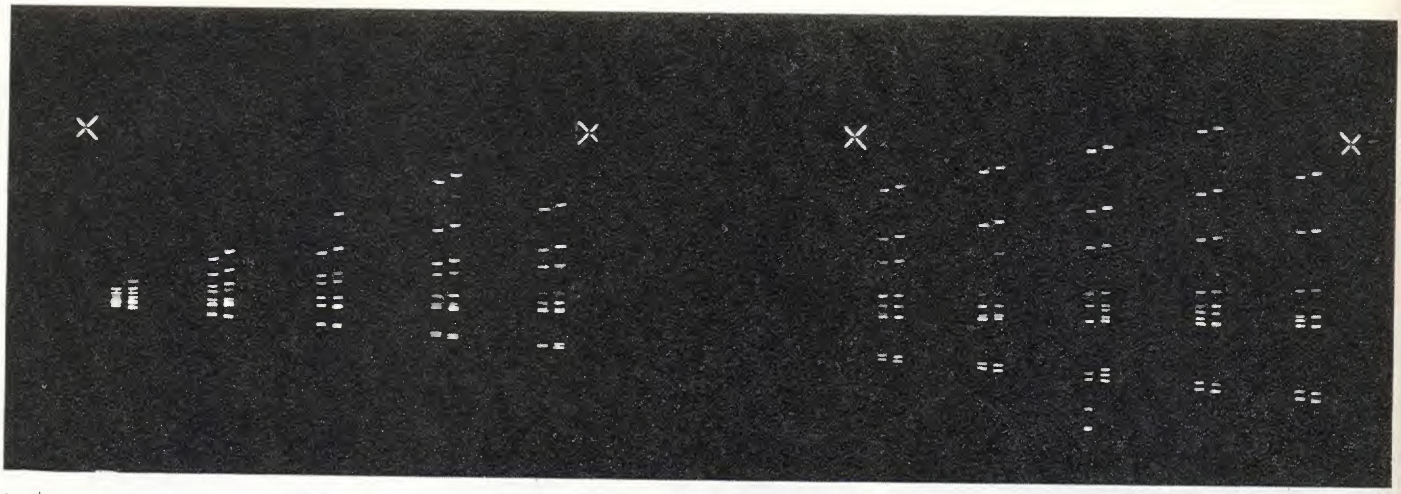


27. Pierwszy zarejestrowany przypadek rozpadu hiperonu Ξ

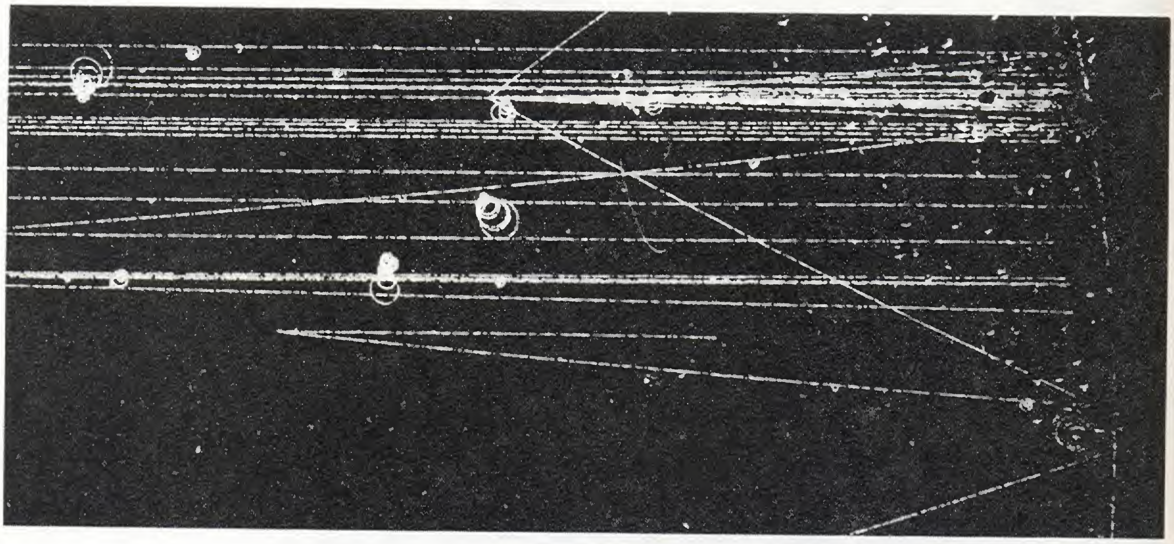
28. Przykłady rejestracji oddziaływań w komorach: a) wodorowej, b) propanowej i c) ksenonowej



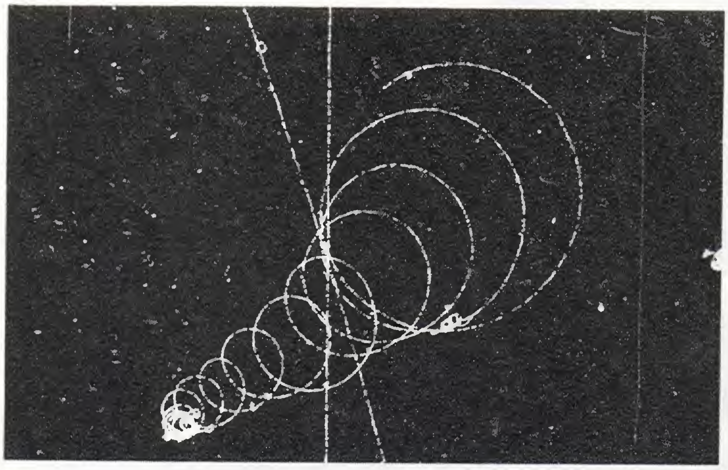
29. Kaskada elektronowo-fotonowa zarejestrowana w komorze ksenonowej



30. Ślady torów zarejestrowane w spektrometrze iskrowym



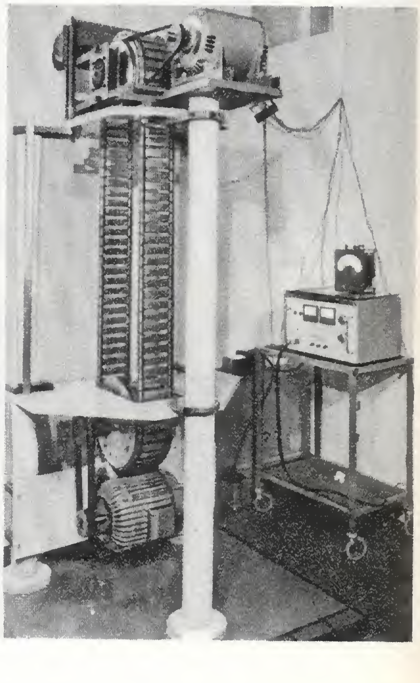
31. Proces powstania i rozpadu hiperonu Λ^0 zarejestrowany w wodorowej komorze pęcherzykowej



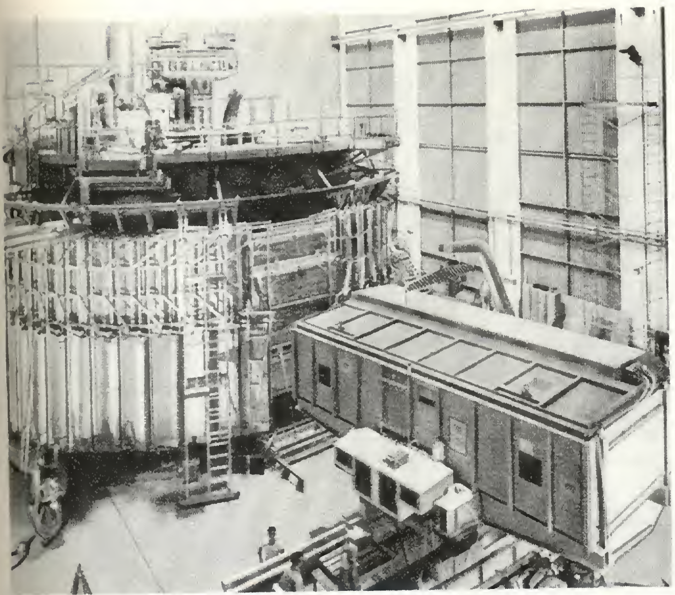
32. Elektron comptonowski zarejestrowany w komorze freonowej



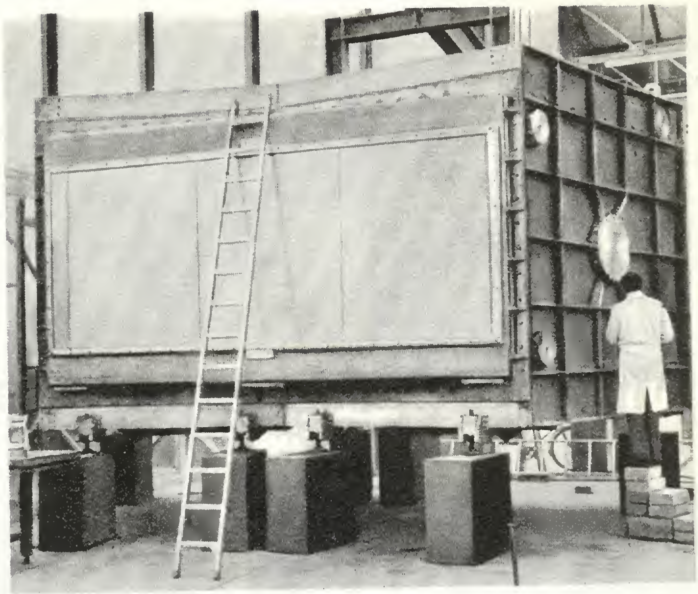
33. Szklana komora pęcherzykowa wypełniona izopentanem o temperaturze 130°C



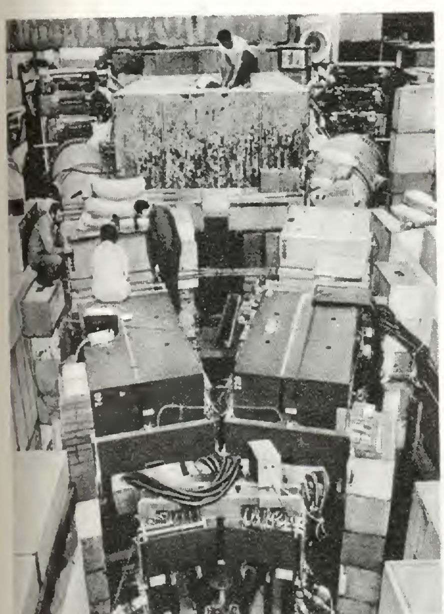
34. Pierwszy akcelerator elektrostatyczny wykorzystujący generator Van de Graaffa. Akcelerator przyspieszał protony i deuterony do energii 0,6 MeV



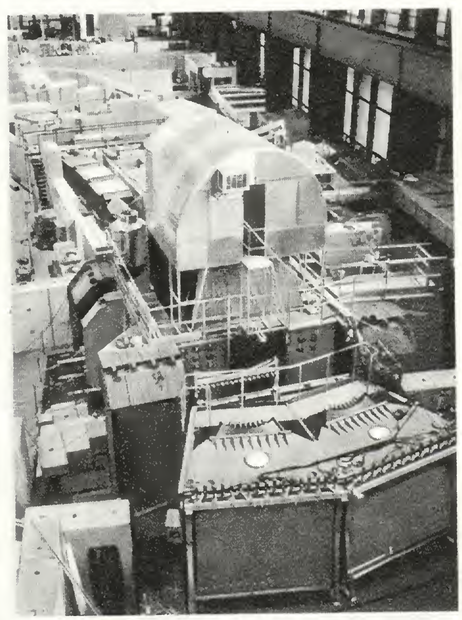
35. Wielka Europejska Komora Pęcherzykowa wraz z zewnętrznym urządzeniem do dentyfikacji cząstek



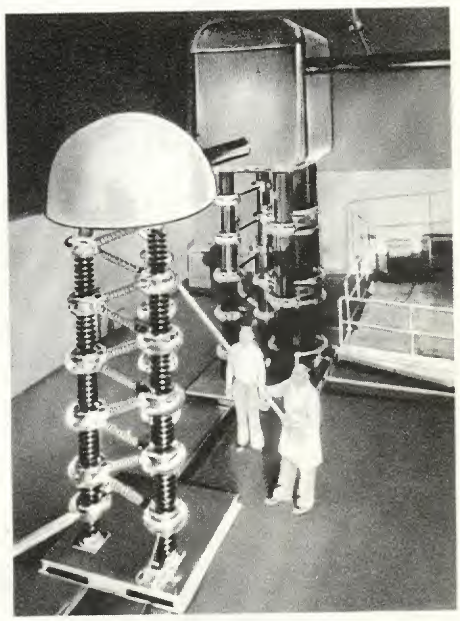
36. Licznik Czerenkowa



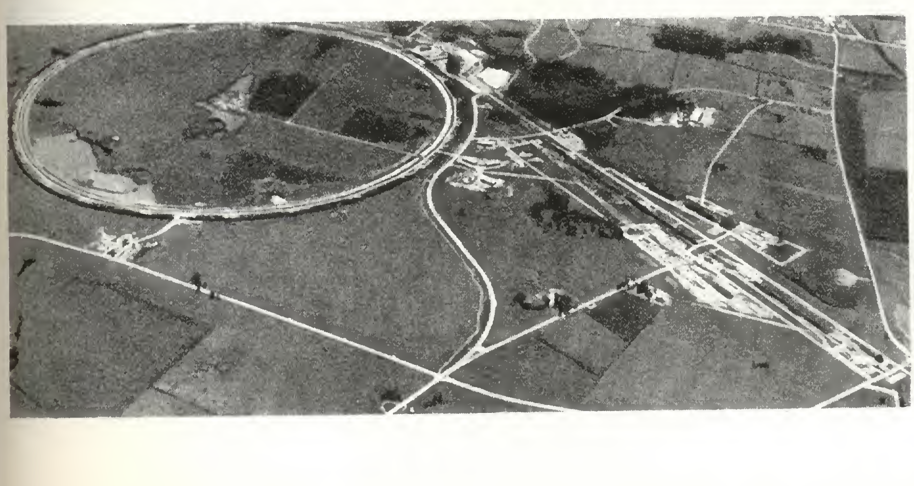
37. Spektrometr dwuramienny



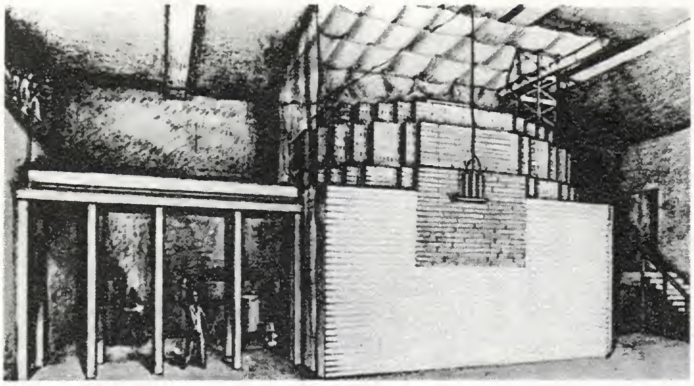
38. Spektrometr Ω



39. Akcelerator elektrostatyczny z generatorem typu Cockrofta i Waltona przyspieszający protony do energii 10 MeV. Uzyskiwana wiązka protonów jest wprowadzana do synchrotronu celem dalszego przyspieszania



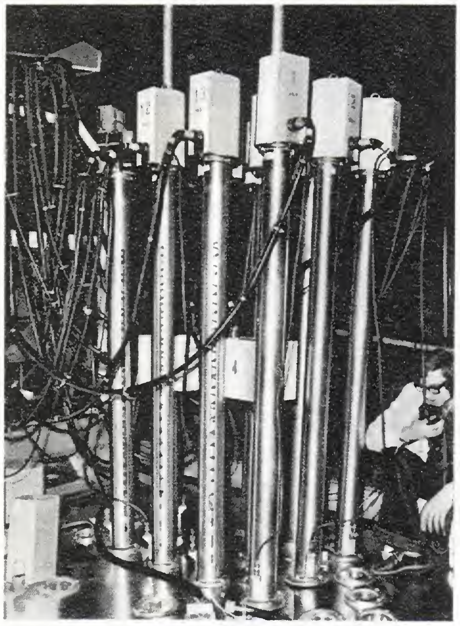
40. Widok ogólny synchrotronu protonów Narodowego Laboratorium Akceleratorowego im. H. Fermiego w Batavii (St. Zjednoczone)



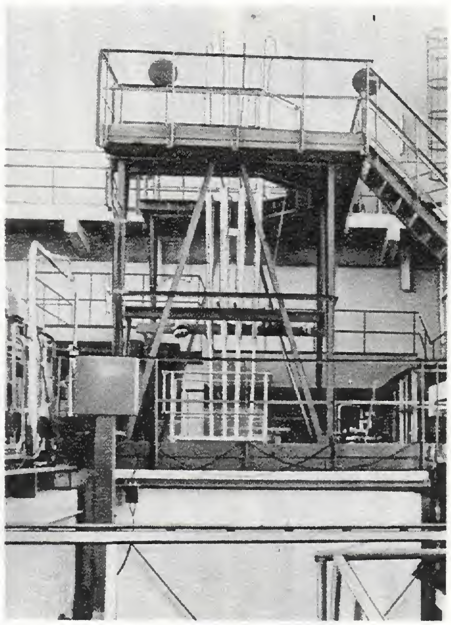
41. Pierwszy reaktor jądrowy uruchomiony pod kierunkiem E. Fermiego w Chicago



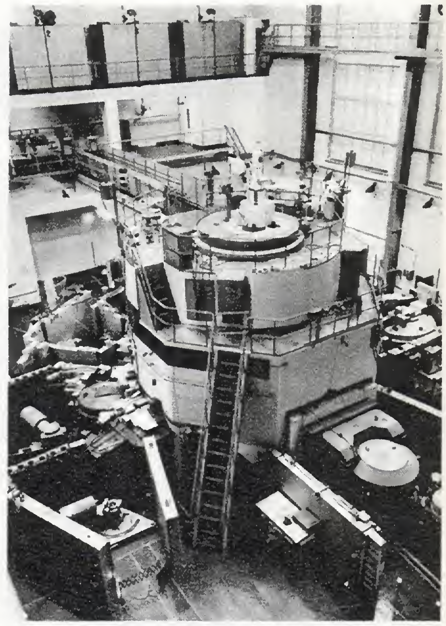
42. Widok ogólny reaktora MARIA w Instytucie Badań Jądrowych w Świerku pod Warszawą



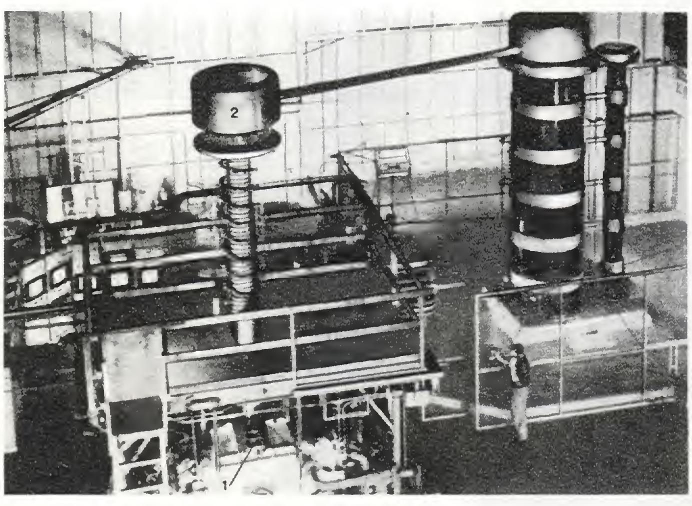
43. Zawieszenie napędów prętów regulacyjnych reaktora MARIA



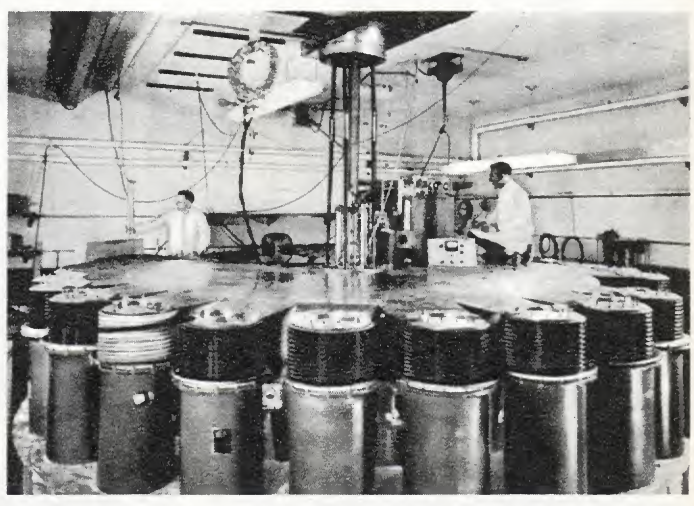
44. Zestaw krytyczny pracujący w Instytucie Badań Jądrowych w Świerku pod Warszawą



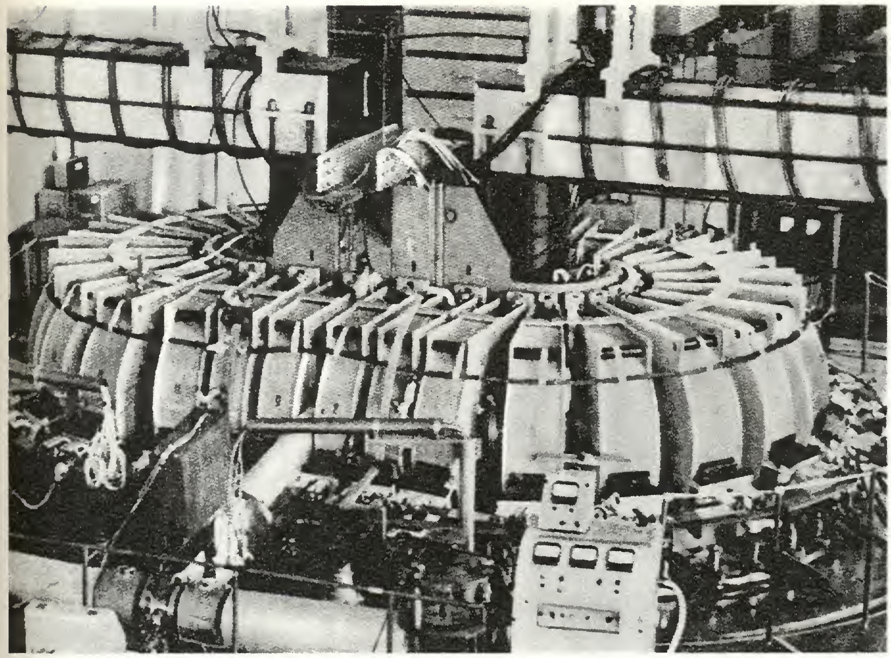
45. Reaktor EWA w Instytucie Badań Jądrowych w Świerku pod Warszawą



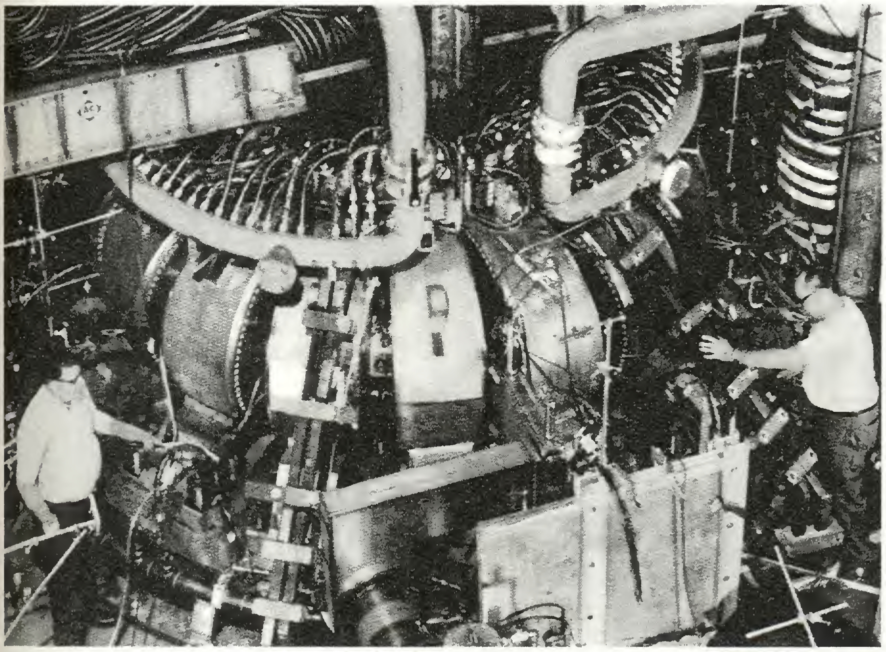
46. Urządzenie DCX (zbudowane w St. Zjednoczonych), w którym osiągnięto rekordowe temperatury rzędu 6 mld K. Obłok gorącej plazmy wytworzony był w pułapce magnetycznej (*1*) przez wiązkę D_2^+ o energii 600 keV, otrzymywaną ze specjalnego akceleratora (*2*)



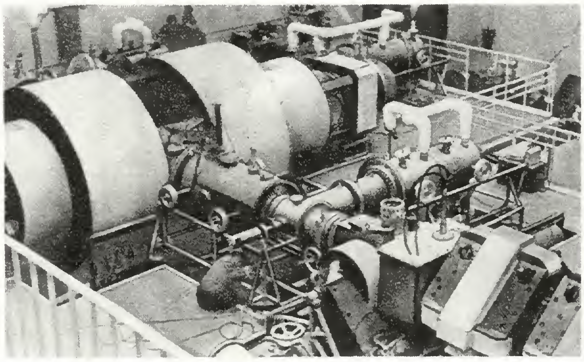
47. Columbus II — jedno z pierwszych urządzeń amerykańskich opartych na zjawisku pinchu liniowego. Na pierwszym planie widać dużą baterię wysokonapięciowych kondensatorów, które służą do zasilania wyładowania. Cylindryczna komora eksperymentalna o osi skierowanej pionowo umieszczona jest w środku układu (widać tylko zakończenia jednej z elektrod)



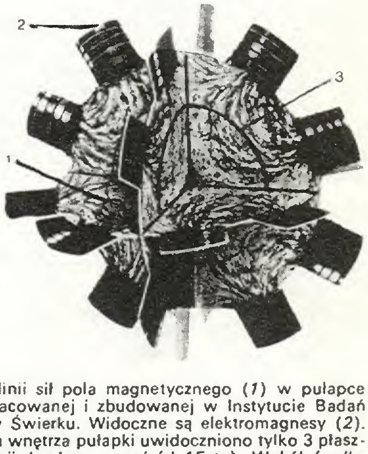
48. Urugan — największy obecnie stellarator (Charków, ZSRR). Na zdjęciu widać komorę w kształcie toru wyścigowego wraz z uzwojeniami służącymi do wytwarzania pola ograniczającego plazmę. U góry widoczna jest część rdzenia potężnego transformatora wykorzystywanego do wytwarzania i grzania plazmy



49. Tokamak-ST. Powstał drogą przebudowy największego amerykańskiego stellaratora (Stellarator-Model C) w Instytucie Plazmy w Princeton. Na zdjęciu widoczny jest fragment toroidalnej komory próżniowej otoczonej uzwojeniami wytwarzającymi stabilizujące pole magnetyczne. W górnej części zdjęcia jest widoczny również rdzeń wielkiego transformatora, który służy do indukowania w tej komorze prądów wyładowania

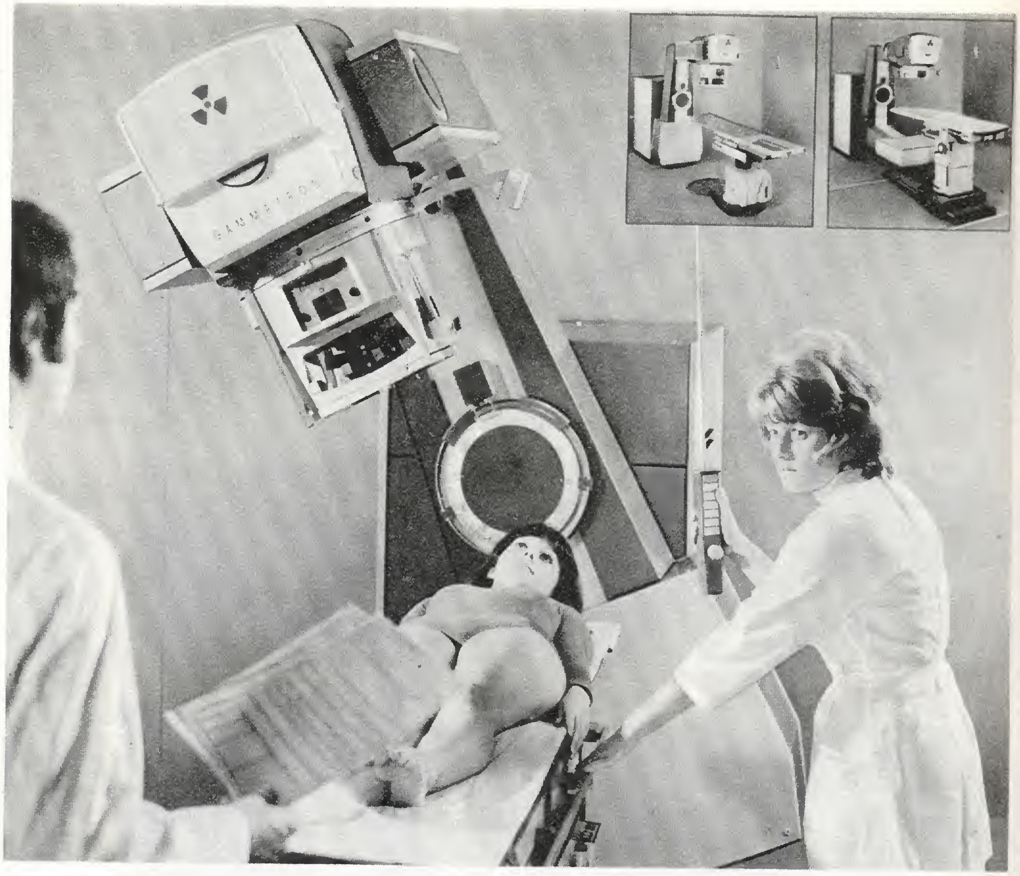


50. Ogra II — pułapka zwierciadlana zbudowana w Instytucie Energii Atomowej w Moskwie. Na zdjęciu widać cylindryczną obudowę komory eksperymentalnej oraz umieszczonych wokół niej prętów stabilizacyjnych. Widoczne są również zgrubienia, w których umieszczono uzwojenia wytwarzające pole zwierciadlane. Na pierwszym planie widać kanał i urządzenia pomocnicze służące do iniekcji wiązek wysokoenergetycznych cząstek neutralnych



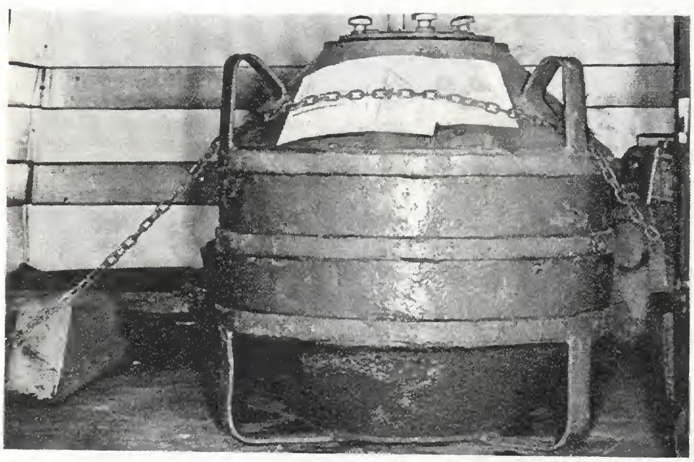
51. Rozkład linii sił pola magnetycznego (*1*) w pułapce typu SM opracowanej i zbudowanej w Instytucie Badań Jądrowych w Świerku. Widoczne są elektromagnesy (*2*). Dla pokazania wnętrza pułapki uwidoczniono tylko 3 płaszczyzny symetrii (wybrane spośród 15-tu). Wokół środka układu zaznaczono granice (*3*) obszaru efektywnego ograniczenia plazmy

52. Gammatron — urządzenie do naświetlania promieniami γ (firmy Siemens)

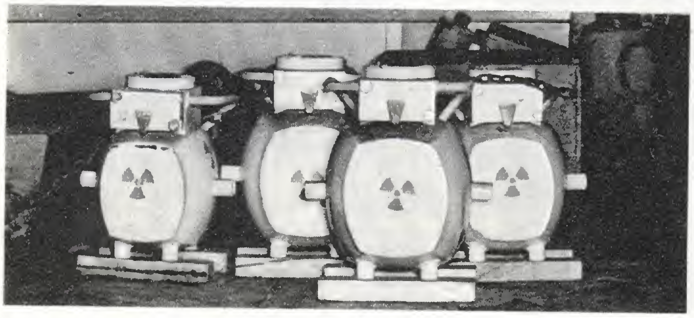


53. Defektoskop izotopowy: a) defektogram zaworu, b) prześwietlanie 6 zaworów na raz

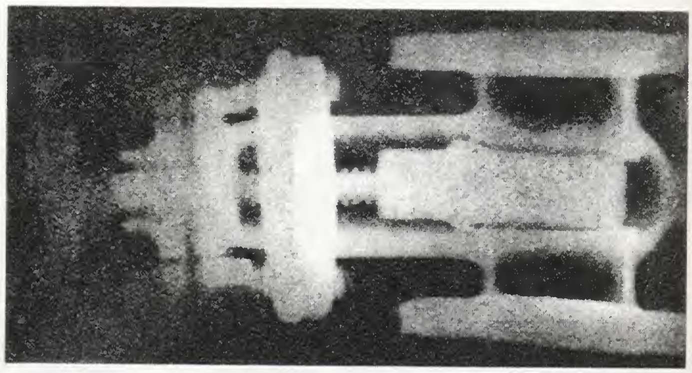
54. Transportowe pojemniki na radioizotopy o aktywnościach większych (a) i niewielkich (b)



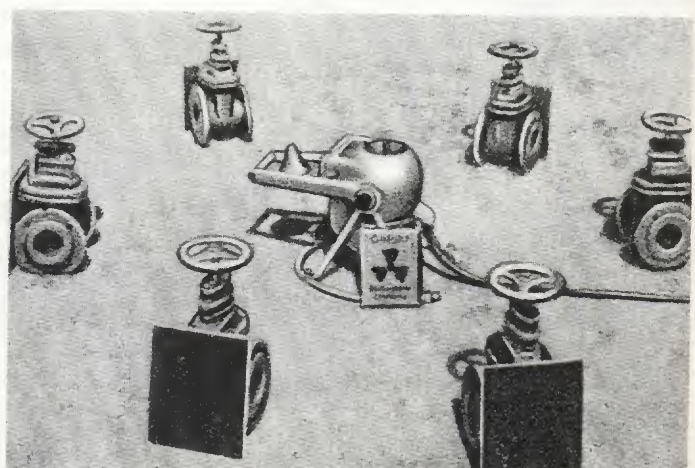
a



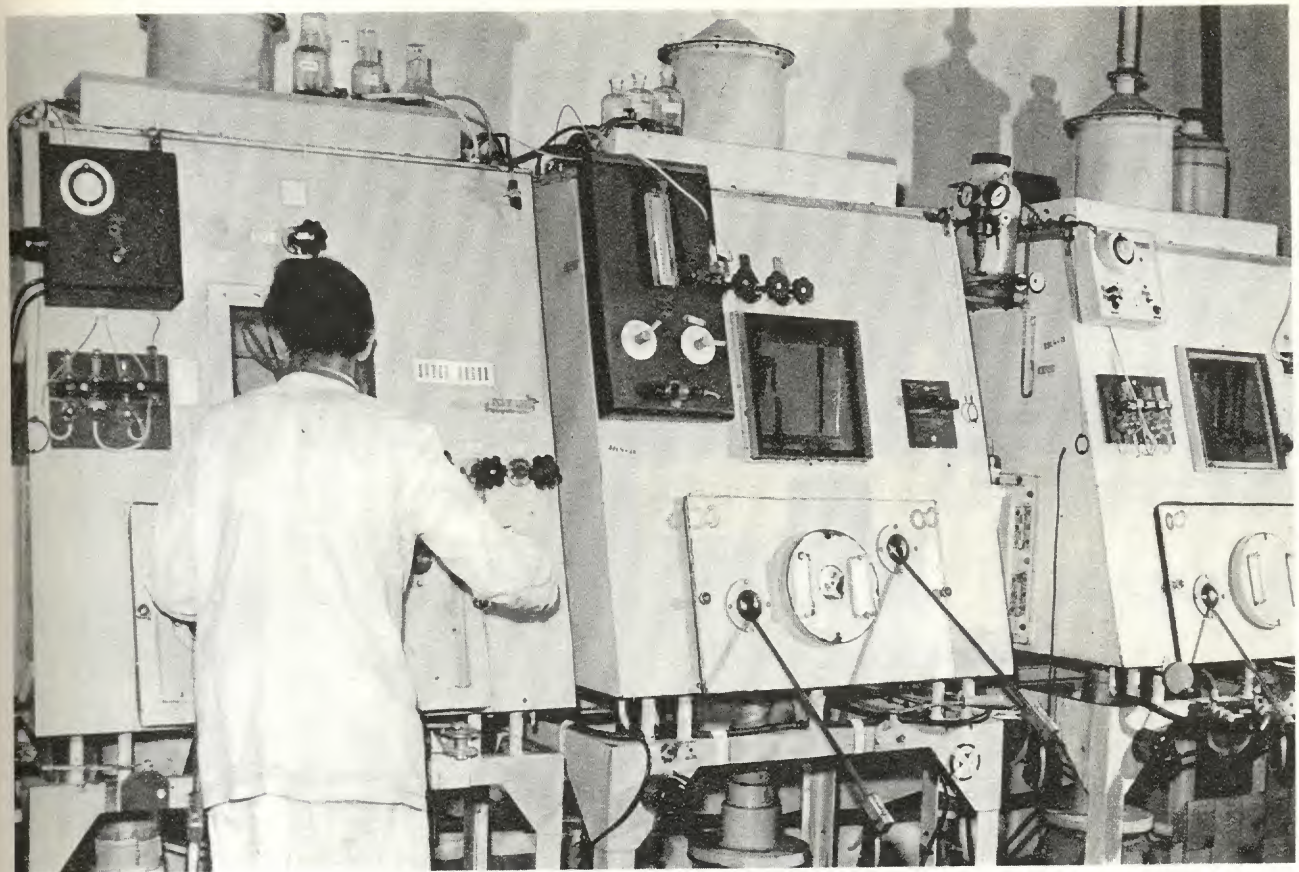
b

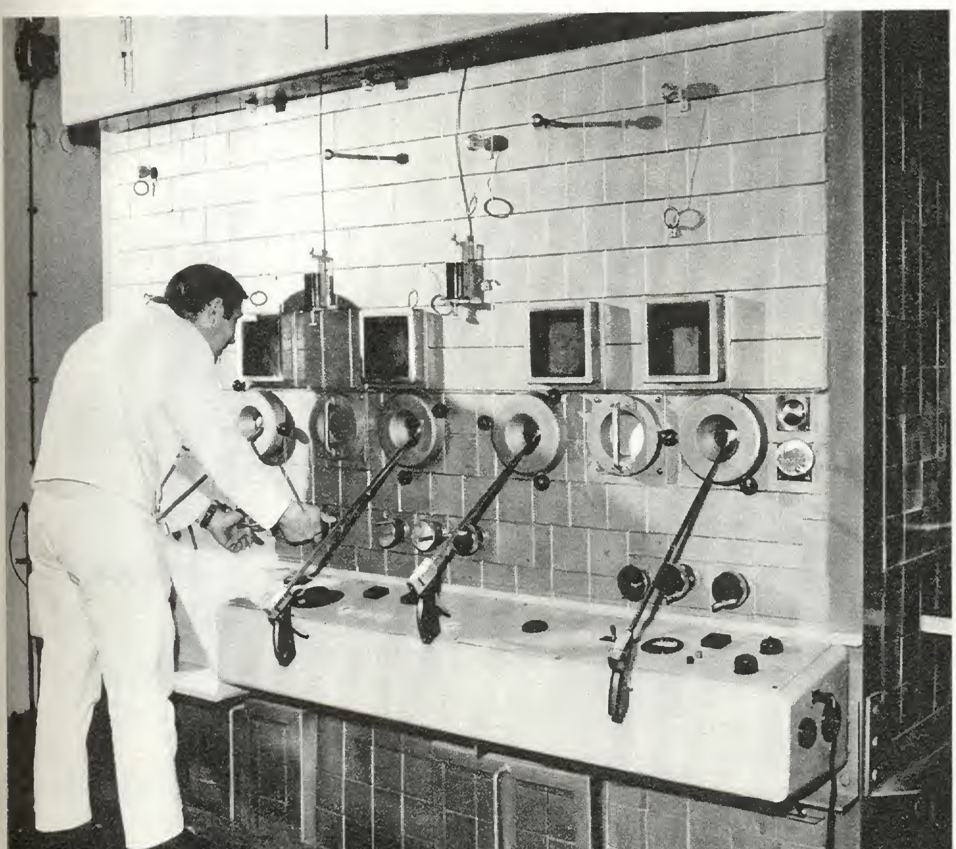


a



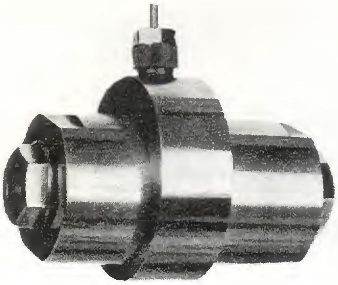
b



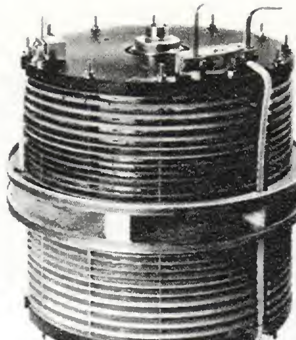


55. Boks do operacji z radioizotopami i związkami znaczonymi o niewielkich aktywnościach

56. Praca przy komorze gorącej (z radioizotopami o większych aktywnościach)

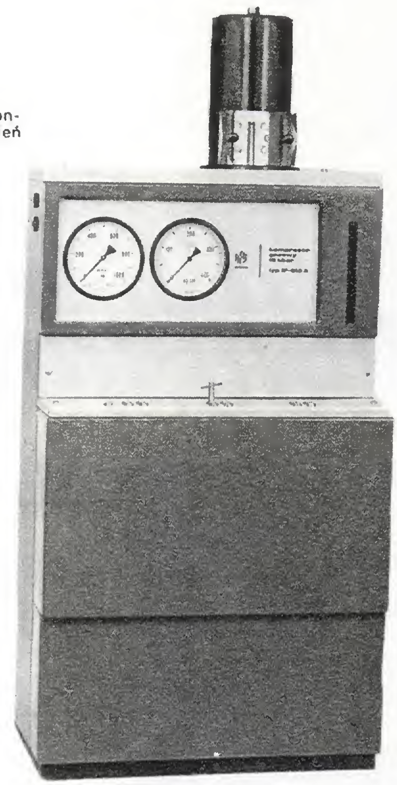


57. Wysokociśnieniowa komora optyczna IF-024 skonstruowana w OBR Wysokich Ciśnień PAN „Unipress" (rzeczywista długość 10 cm)



58. Laboratoryjny elektromagnes nadprzewodnikowy

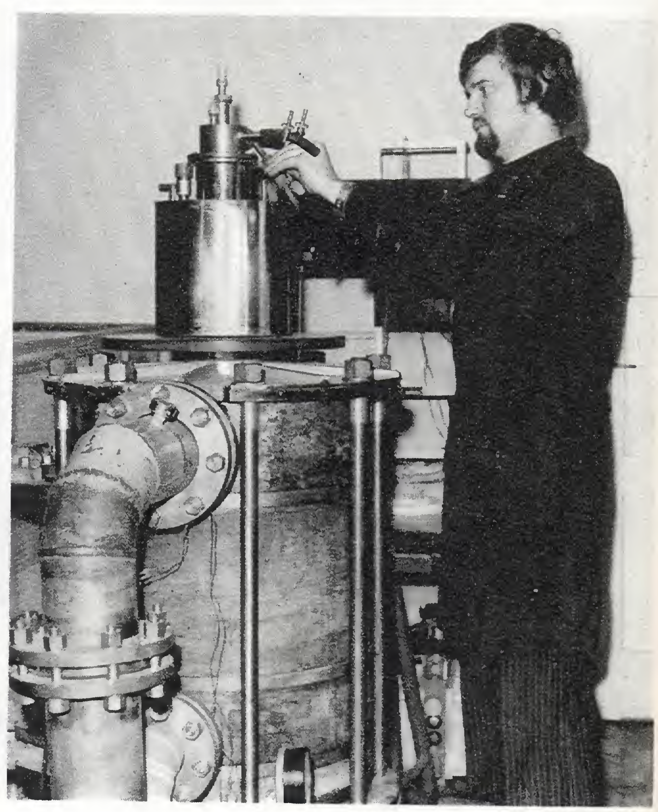
59. Kompresor helowy IF-012A skonstruowany w OBR Wysokich Ciśnień PAN „Unipress"



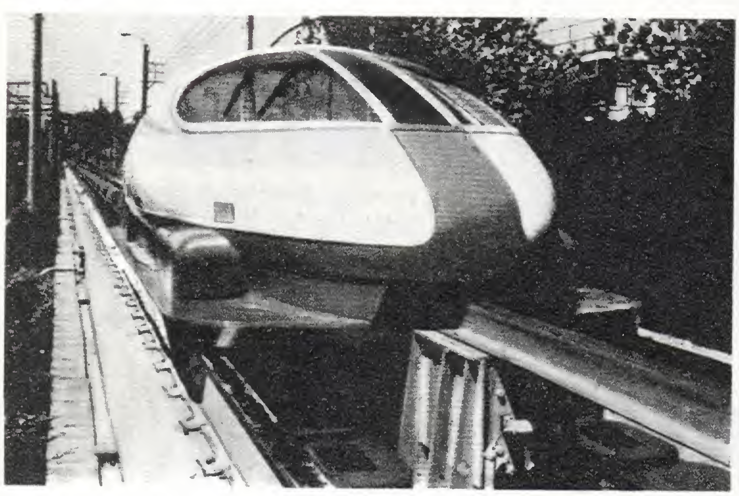


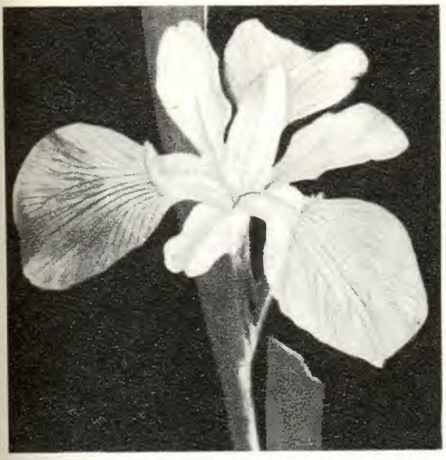
60. Elektromagnes hybrydowy wytwarzający pole magnetyczne 25 T (Instytut Energii Atomowej, Moskwa)

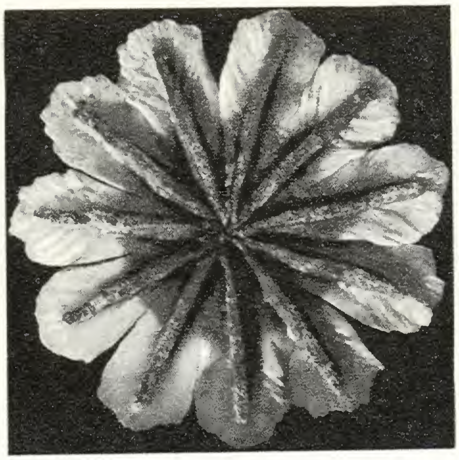
61. Trzycewkowy elektromagnes bezrdzeniowy typu Bittera chłodzony wodą; wytwarza pole magnetyczne 20 T przy mocy zasilania 5,7 MW. (Międzynarodowe Laboratorium Silnych Pól Magnetycznych i Niskich Temperatur, Wrocław). Na elekromagnesie ustawiony jest kriostat pomiarowy



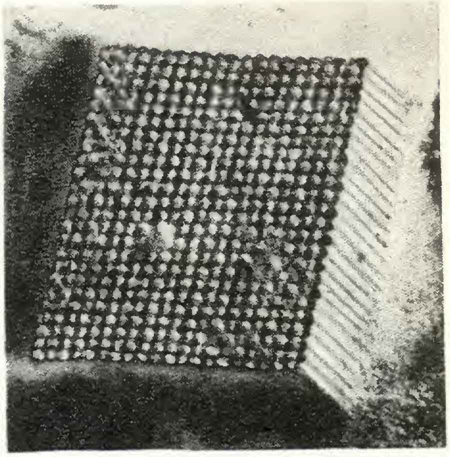
62. Magnetoplan



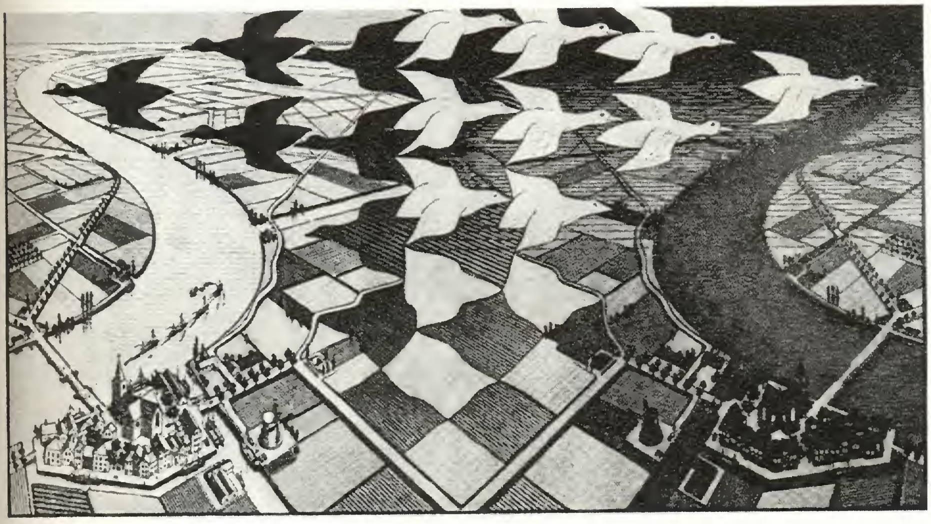




63. Symetria w przyrodzie — irys i widok makówki z góry

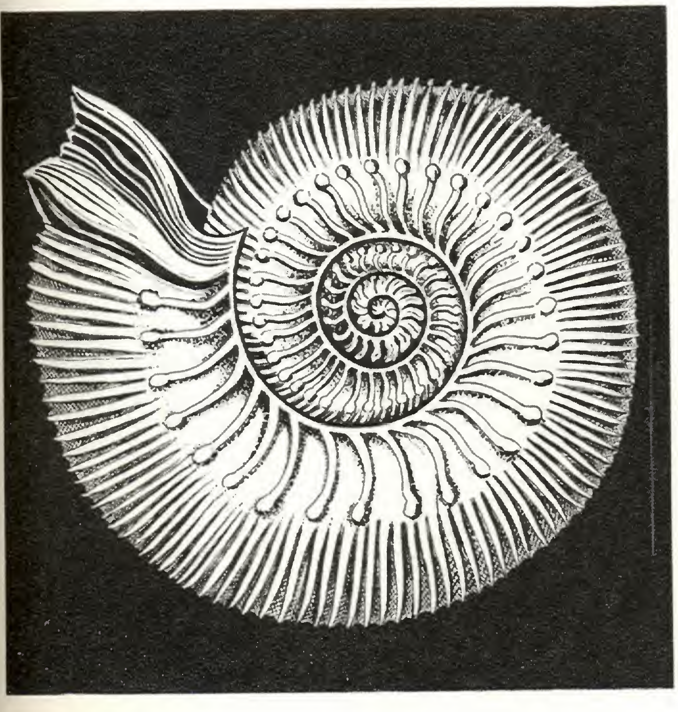


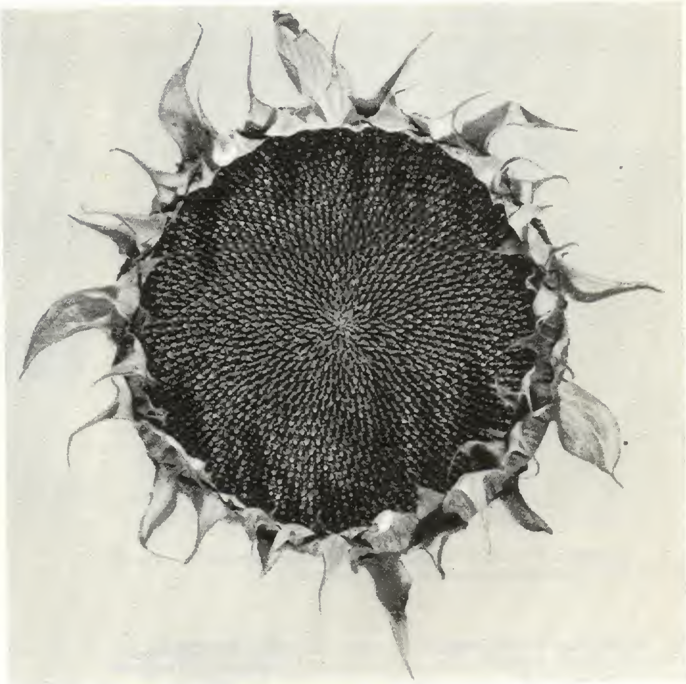
64. Wirusy mozaiki tytoniu ułożone w sieć przestrzenną (obraz otrzymany w mikroskopie elektronowym)

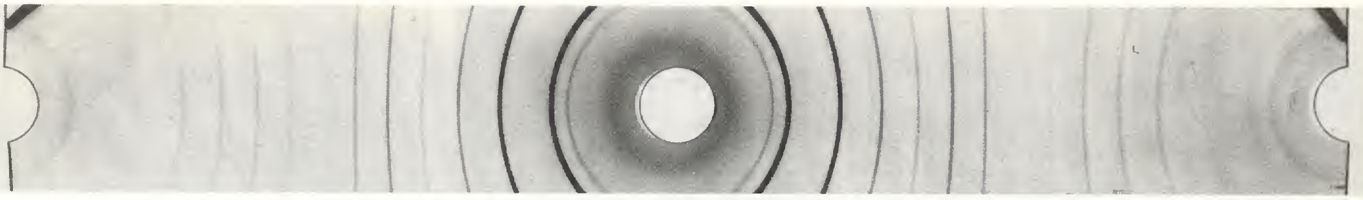


65. Antysymetria w sztuce — M.C. Escher „Dzień i noc"

66. Symetria podobieństwa — muszla amonitu i słonecznik

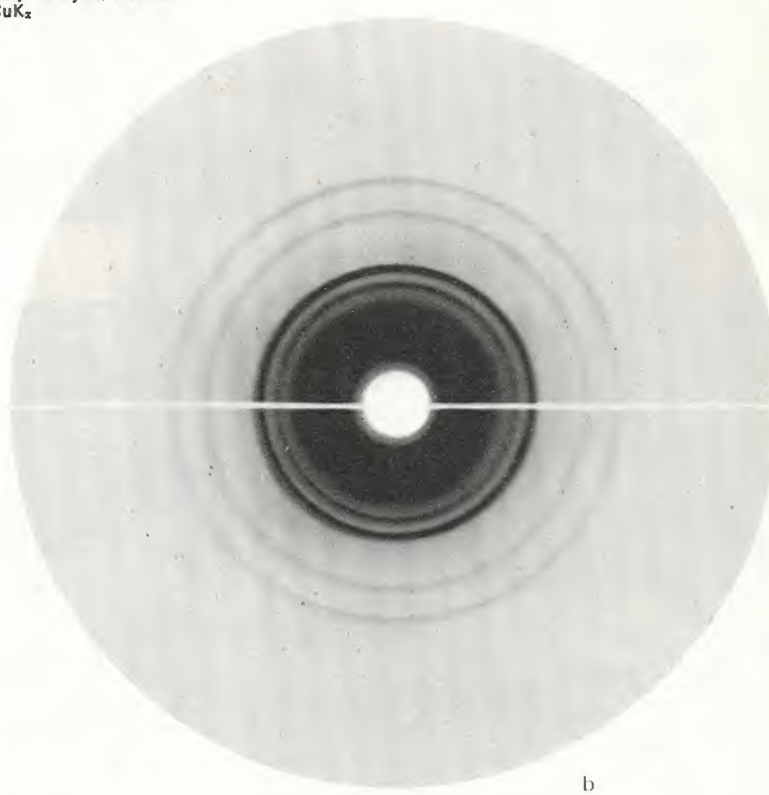




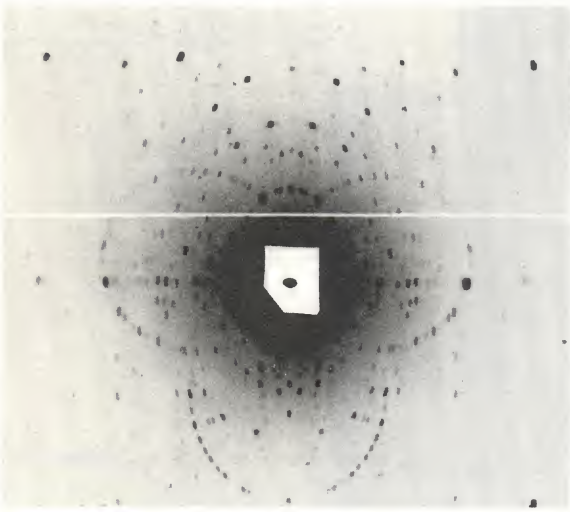


a

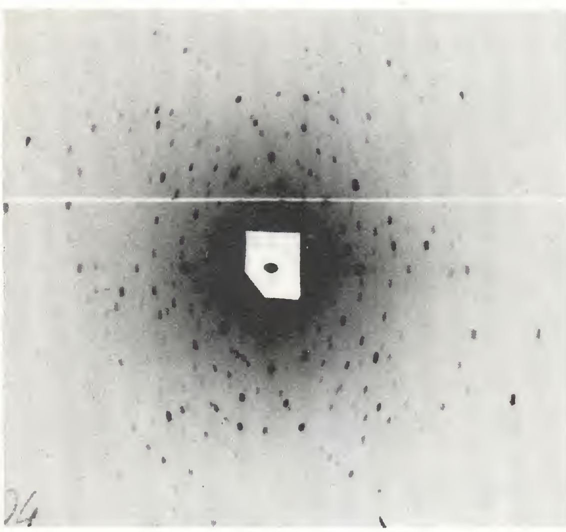
67. Rentgenogram NaCl wykonany metodą Debye'a–Scherrera–Hulla: a) na płaskiej błonie fotograficznej, b) na błonie fotograficznej zwiniętej w walec; promieniowanie CuK_{α}



b

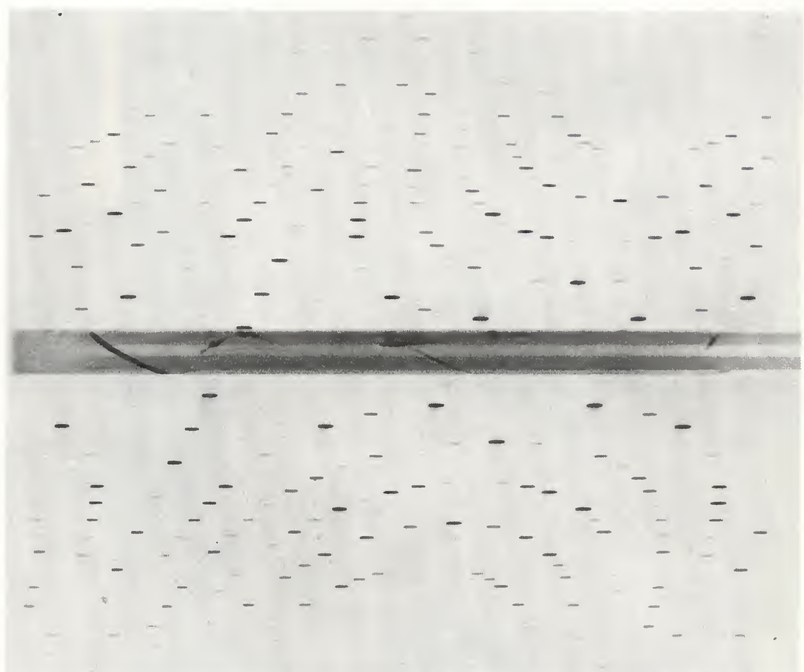


a

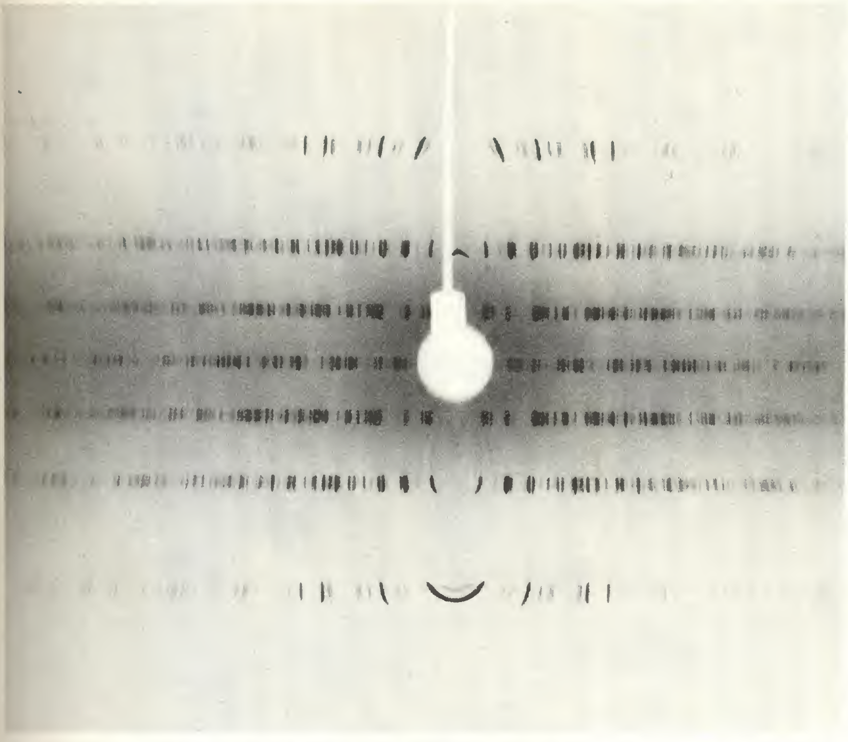


b

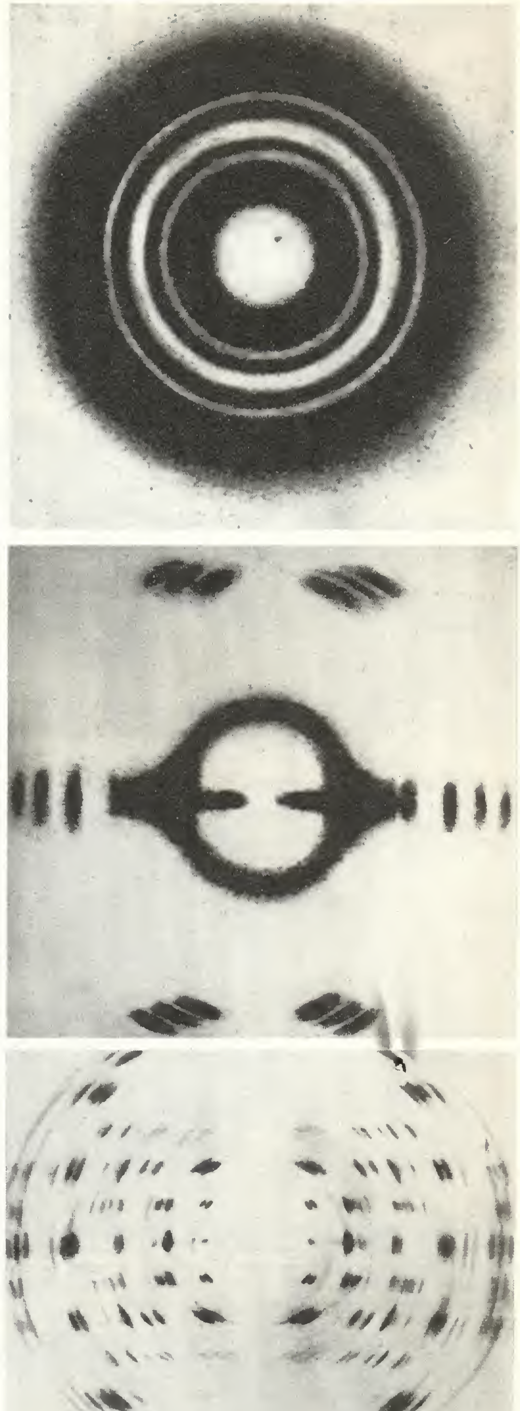
68. Lauegramy: a) widoczna płaszczyzna symetrii (pionowo), symetria *m*; b) widoczna dwukrotna oś symetrii, symetria *2*



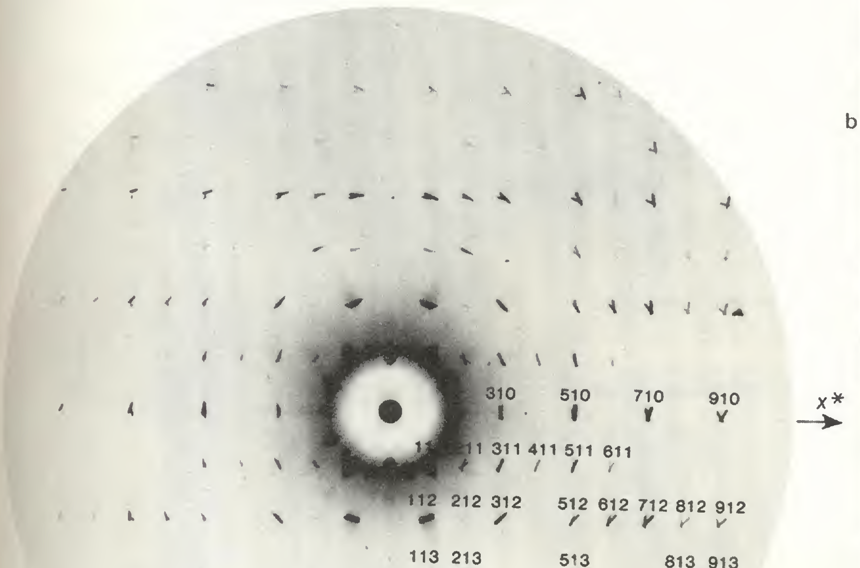
69. Dyfraktogram Weissenberga; kryształ $Zn(NO_3)_2 \cdot 4CO(NH_2)_2 \cdot 2H_2O$ — układ jednoskośny, oś obrotu [001], warstwica zerowa; promieniowanie CuK_{α}



70. Rentgenogram wykonany metodą obracanego kryształu (dyfraktogram warstwicowy); kryształ $ZnB_2 \cdot$ $\cdot 2CO(NH_2)_2$, oś obrotu [100], promieniowanie CuK_α

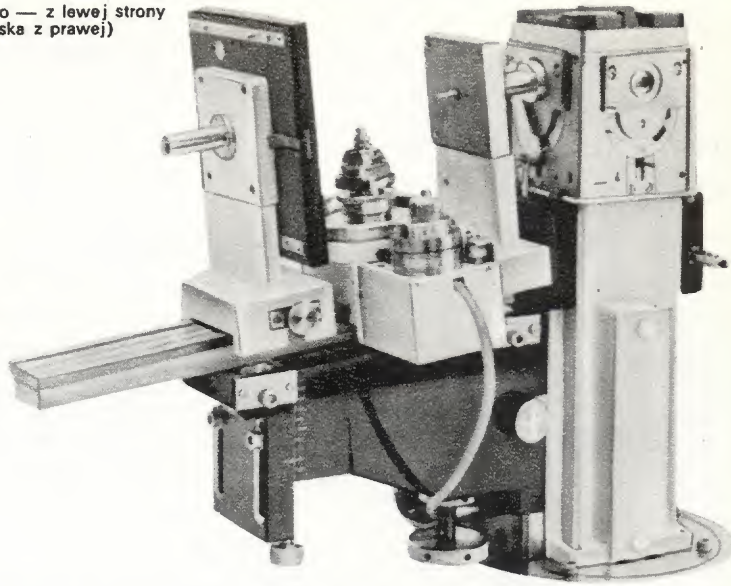


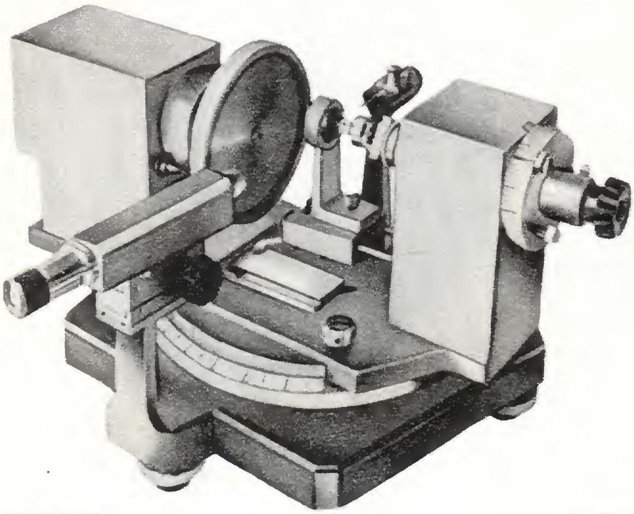
71. Rentgenogramy polimerów wykonane metodą Debye'a–Scherrera–Hulla: a) polimetylen nierozciągnięty, b) polimetylen rozciągnięty do 500% w temperaturze 96°C, c) rentgenogram tekstury soli sodowej kwasu dezoksyrybonukleinowego (Na-DNA)



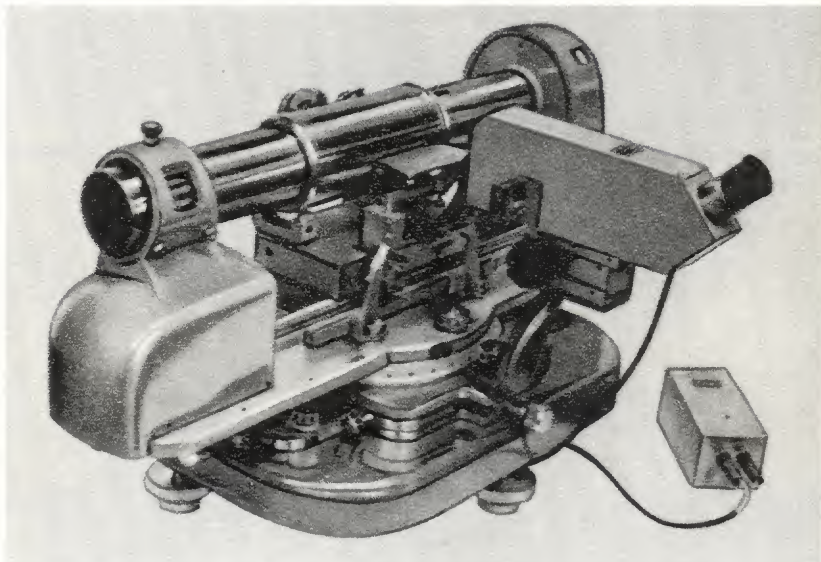
72. Retigram kryształu $CuSO_4 \cdot 3CO(NH)_2$, układ rombowy, oś obrotu (010); warstwica pierwsza brak wygaszeń systematycznych. W jednej ćwiartce pokazano sposób wskaźnikowania refleksów promieniowanie CuK_α

73. Kamera Lauego — z lewej strony (lampa rentgenowska z prawej)

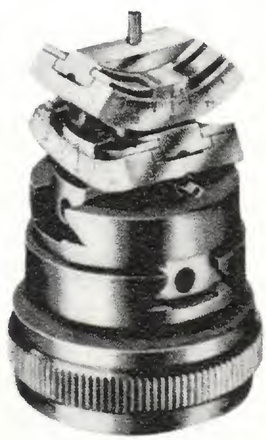




74. Retigraf

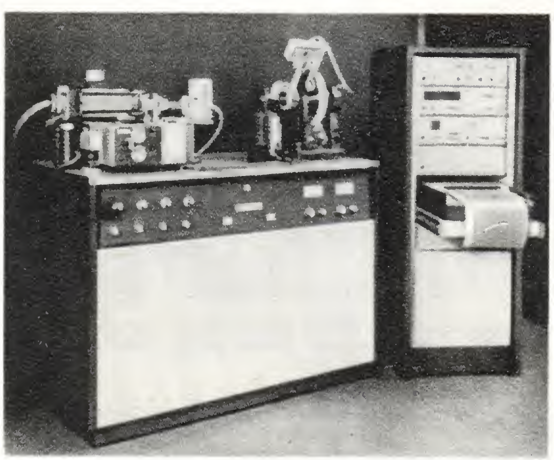


75. Główka goniometryczna

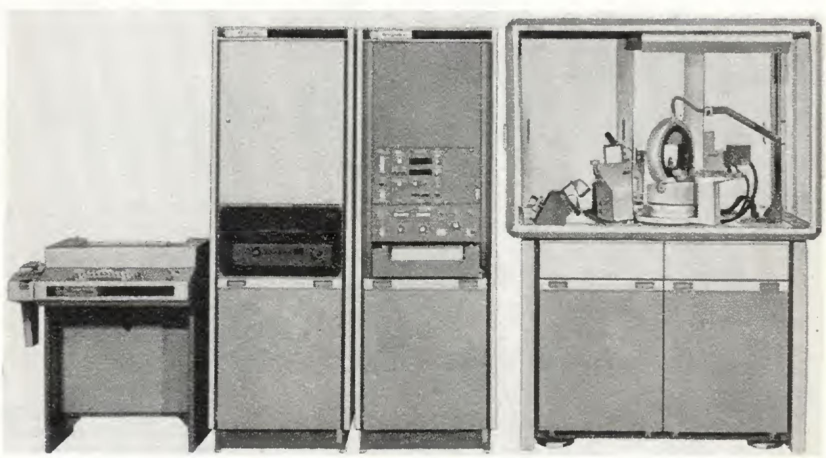


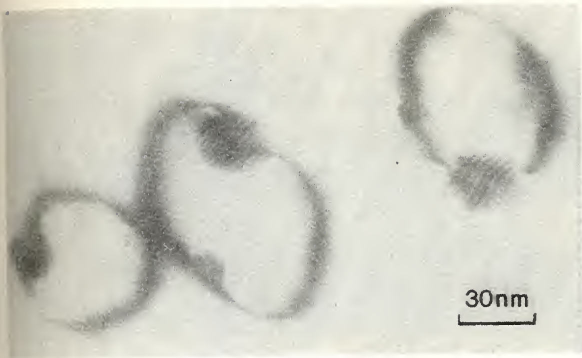
76. Kamera Weissenberga

77. Ogólny widok dyfraktometru rentgenowskiego (firmy Philips); na stole dyfraktometru są ustawione dwa goniometry (urządzenia zmieniające położenie preparatu i licznika), z lewej — poziomy, z prawej — pionowy; z prawej strony stołu — „szafa" z układami elektronicznymi i rejestratorem



78. Czterokołowy dyfraktometr automatyczny do monokryształów (firmy Rigaku)



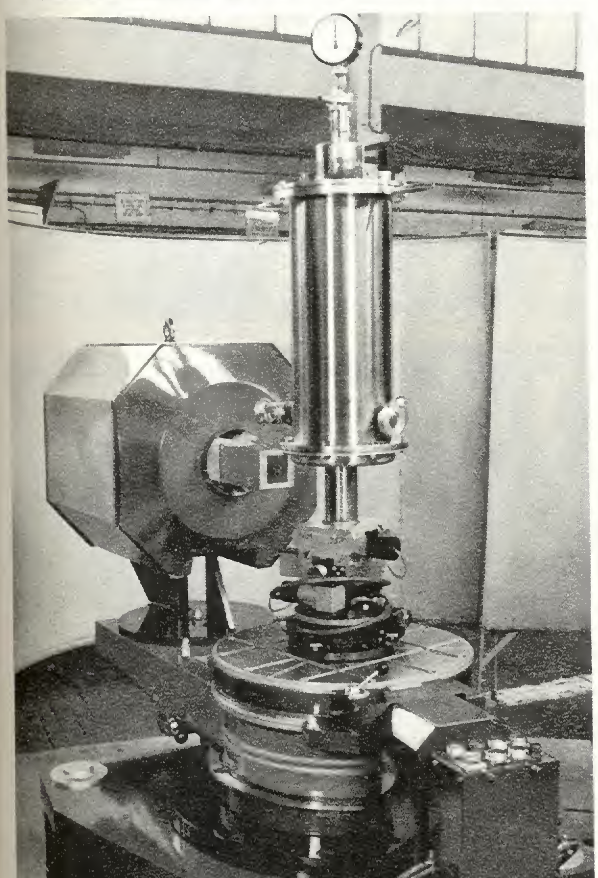


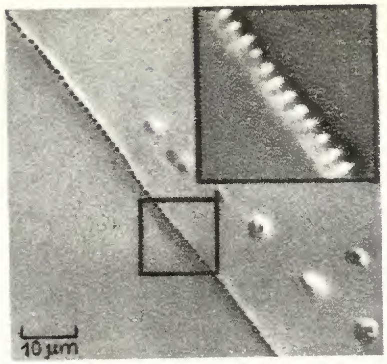
80. Pętle dyslokacyjne w ZnTe generowane przez wydzielenie Ag_2Te



82. Granica małokątowa w ZnTe (pow. 32 000 ×)

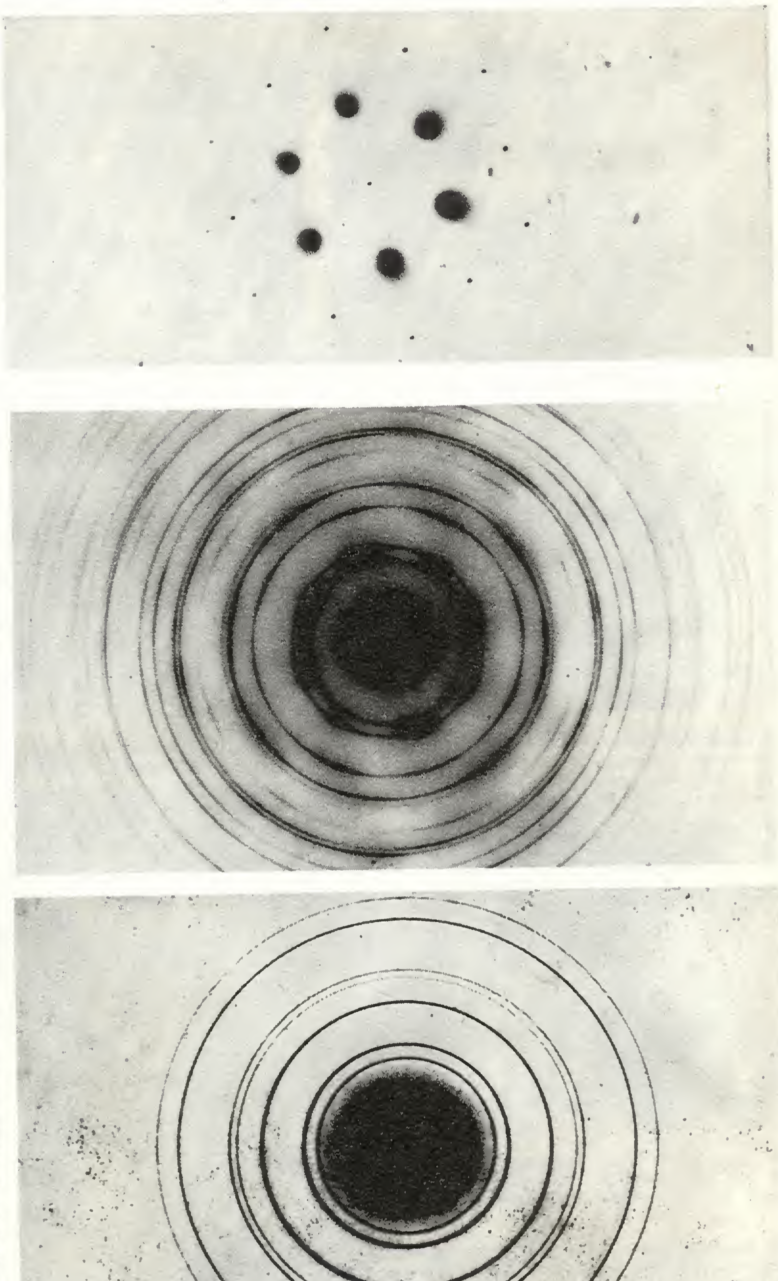
83. Dyfraktometr neutronów typu DN-520 przy reaktorze EWA w Świerku





79. Jamki trawienia ujawniające dyslokacje na granicy ziaren w germanie

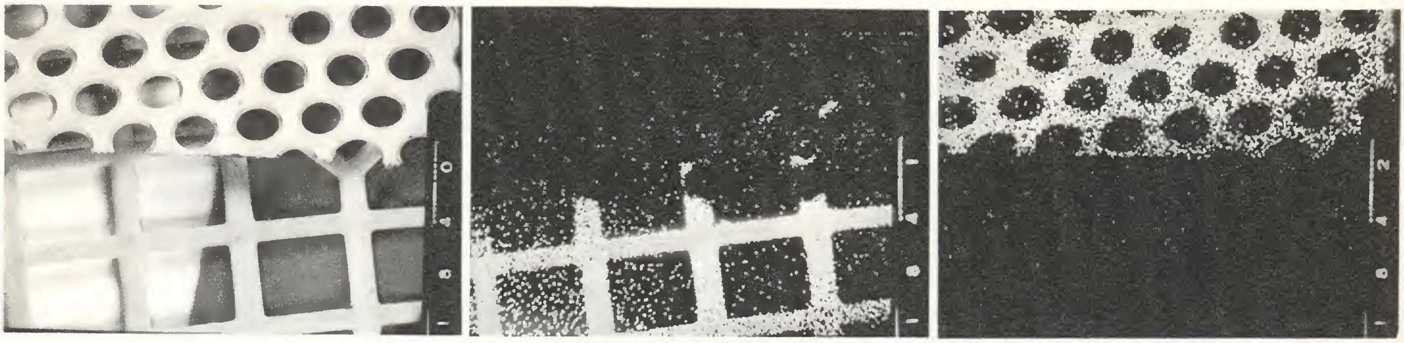
81. Elektronogramy: a) cienkiej folii HgTe (monokryształ); b) drobnopolikrystalicznej folii Au (tekstura); c) próbki polikrystalicznej NaCl



a

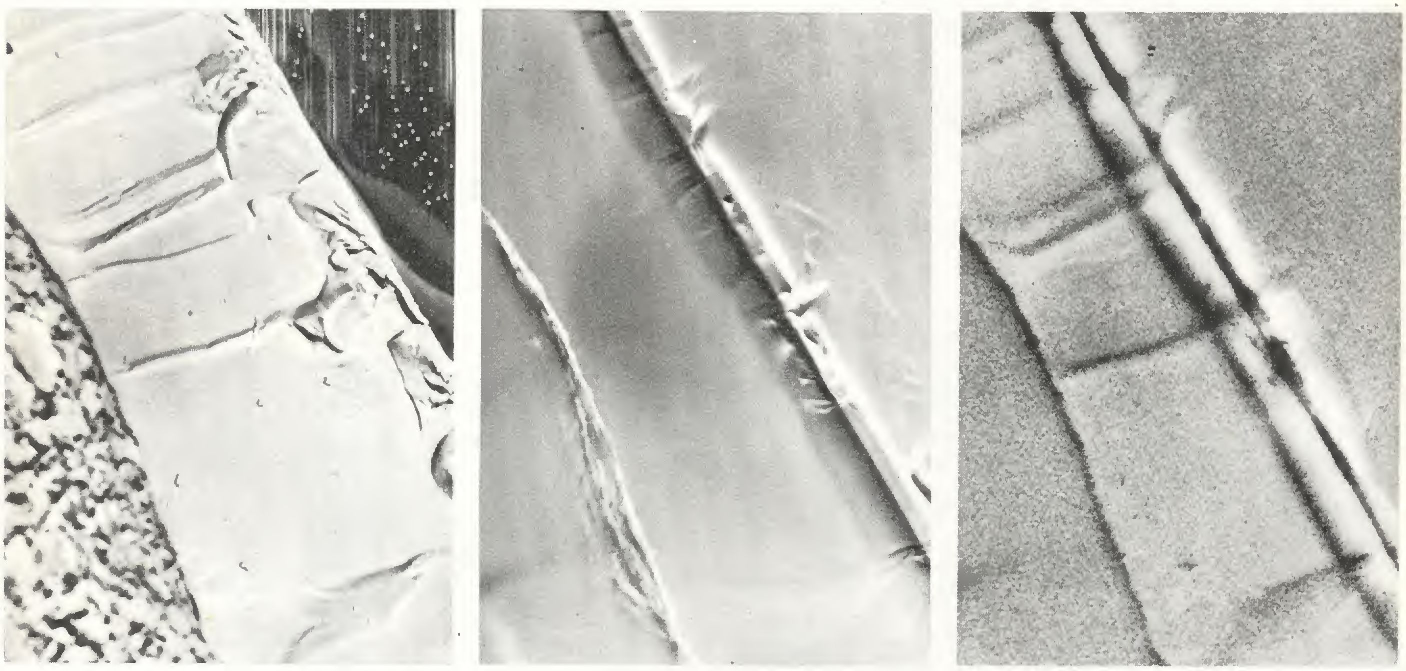
b

c



a
b
c

84. Dwie siatki — molibdenowa i miedziana — na próbce krzemu pokrytej w połowie glinem; obraz uzyskany w skaningowym mikroanalizatorze elektronowym typu JXA-50A (pow. 80 ×): a) w świetle elektronów odbitych (obszar pokryty glinem jest ciemniejszy), b) w świetle charakterystycznego promieniowania rentgenowskiego — linia CuK_α (widoczna siatka miedziana); c) w świetle charakterystycznego promieniowania rentgenowskiego — linia MoL_α (widoczna siatka molibdenowa)



a
b
c

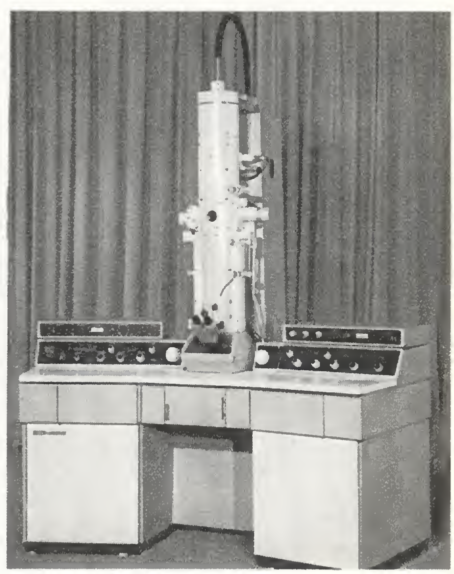
85. Przekrój poprzeczny diody z GaAs; obraz uzyskany w skaningowym mikroskopie elektronowym typu JXA-50A (pow. 80 ×): a) w świetle elektronów odbitych, b) obraz z kontrastem pochodzącym od siły elektromotorycznej — widoczne defekty struktury złącza *p-n*, c) obraz katodoluminescencji w bliskiej podczerwieni (800–1200 μm) w obszarze złącza



86. Powierzchnia krzemu ze złączem *p-n*; obraz uzyskany w zwierciadlanym mikroskopie elektronowym (pow. 100 ×); w części ciemniejszej obszar typu *p*

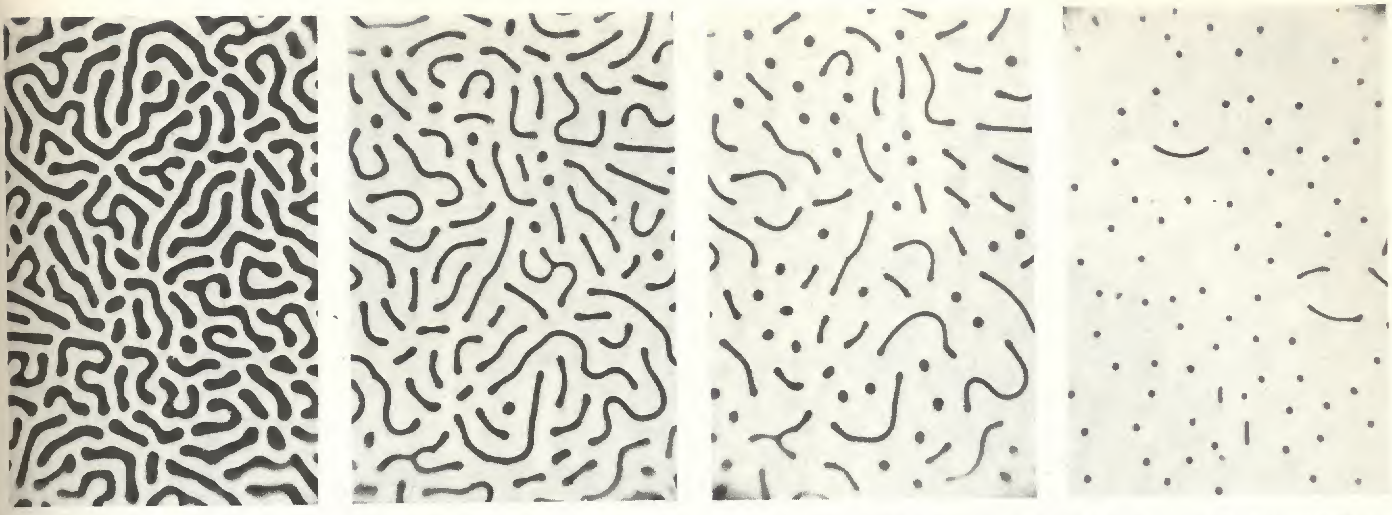


87. Dyslokacja helikoidalna w ZnTe:Ag zawierającym drobne wydzielenia Ag_2Te; obraz uzyskany w transmisyjnym mikroskopie elektronowym typu JEM-200A (jasne pole, napięcie 200 kV)



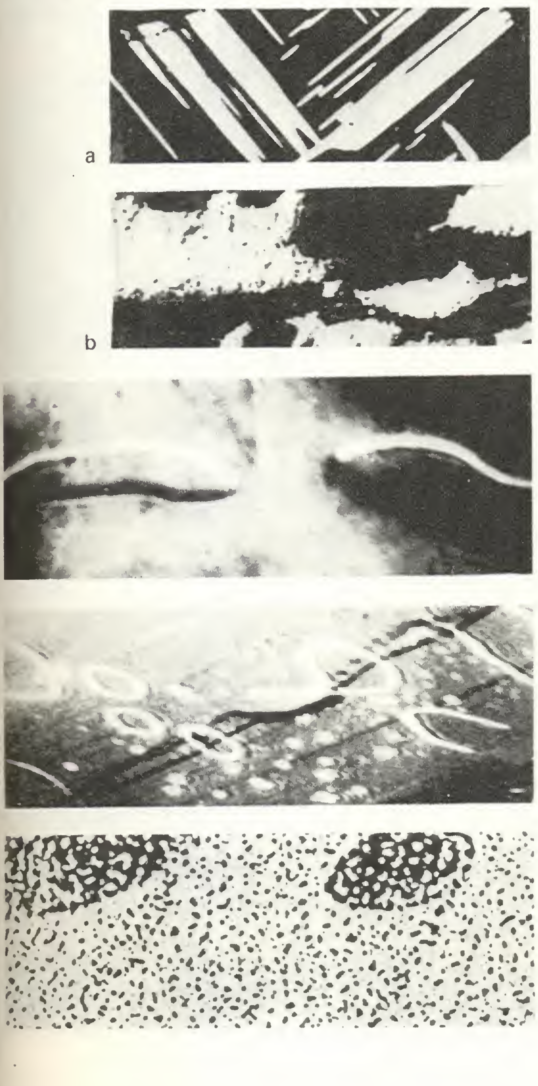
88. Transmisyjny mikroskop elektronowy o dużej rozdzielczości (typ JEM-200C)

a
b
c
d



89. Struktura domenowa cienkich warstw granatu $Y_{1,6}Gd_{0,6}Bi_{0,8}Fe_{3,8}Ga_{1,2}O_{12}$. Zdjęcia uzyskane metodą Faradaya: a) struktura przy H = 0; b), c) i d) struktura odpowiednio przy H = 2170 A/m, 2900 A/m, 3790 A/m (pole skierowane prostopadle do warstwy)

93. Struktury domenowe ferroelektryków: a) tytanianu baru, obraz otrzymany metodą optyczną (w mikroskopie polaryzacyjnym); b) siarczanu trójglicyny, obraz otrzymany metodą proszkową; c) siarczanu trójglicyny, obraz otrzymany metodą trawienia; d) siarczanu trójglicyny, obraz otrzymany za pomocą elektronowego mikroskopu skaningowego; e) siarczanu trójglicyny, obraz otrzymany za pomocą mikroskopu elektronowego z zastosowaniem metody dekoracji

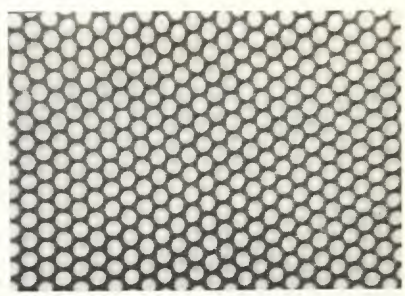


a
b
c
d
e



a
b

90. Struktura domenowa ferrytu barowego $BaFe_{12}O_{19}$. Zdjęcia uzyskane: a) metodą proszkową, b) metodą Faradaya



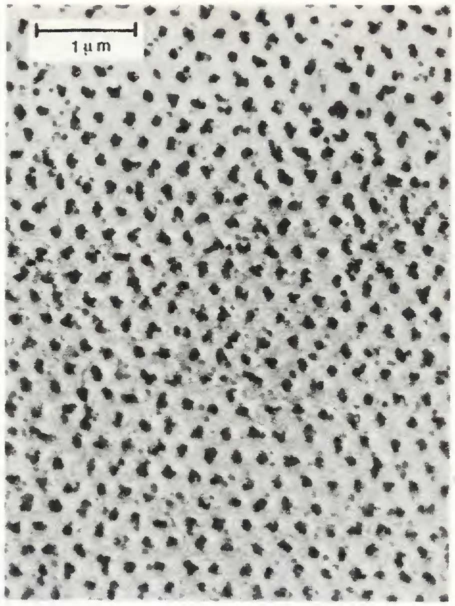
91. Siatka domen cylindrycznych w granacie $Y_{2,5}G_{0,5}Fe_4GaO_{12}$



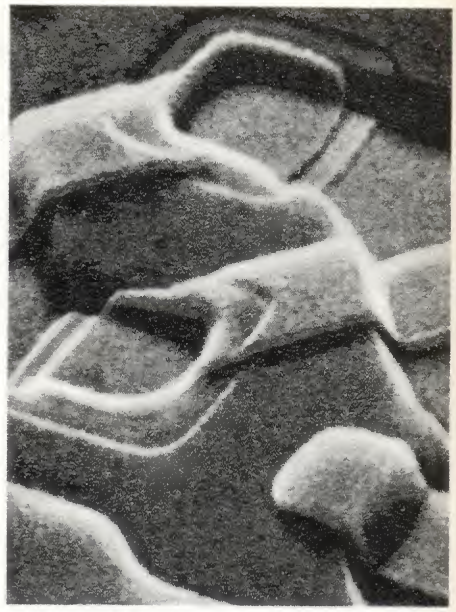
92. Struktura domenowa cienkich warstw amorficznych (stop Gd-Co)



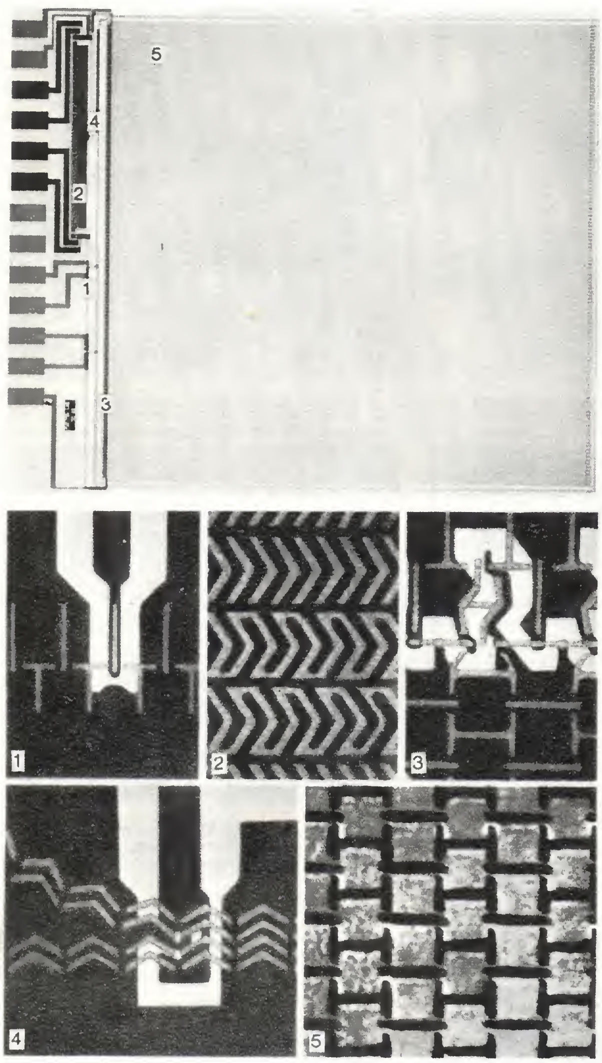
94. Figury proszkowe na powierzchni nadprzewodnika w stanie pośrednim (Faber, 1958 r.)

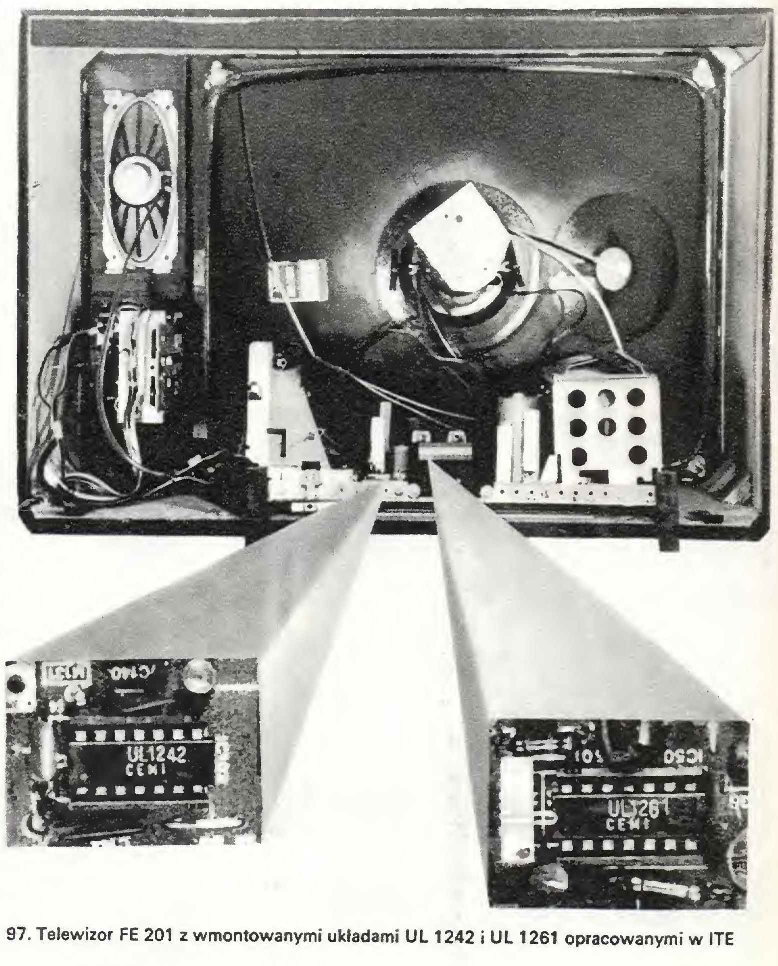


95. Figury proszkowe na powierzchni nadprzewodnika typu II w stanie mieszanym, powierzchnia zdjęcia jest prostopadła do linii sił pola (Essman i Träuble, 1967 r.)



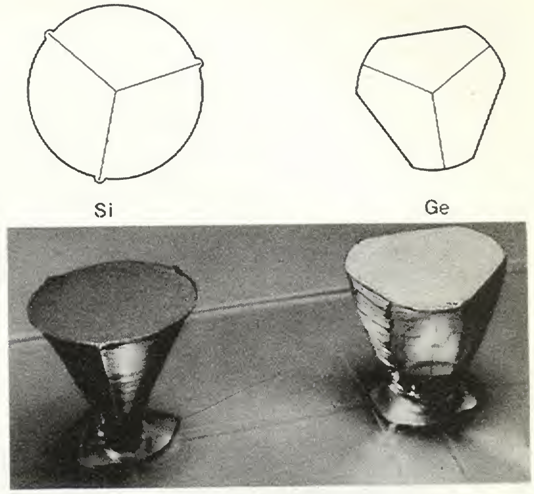
96. Fragment struktury układu scalonego otrzymanego metodą dyfuzji i fotolitografii. Widoczne są paski aluminium stanowiące połączenia elektryczne między elementami układu



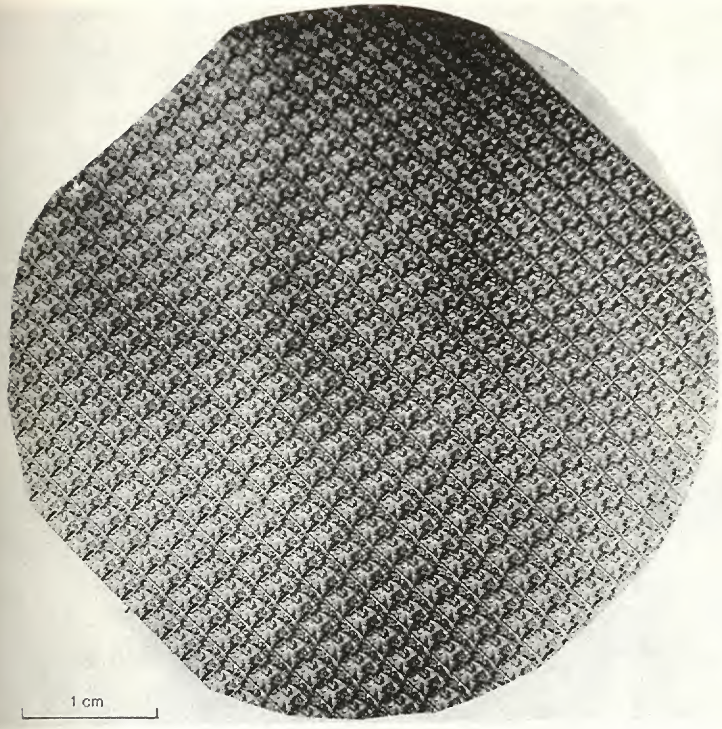


97. Telewizor FE 201 z wmontowanymi układami UL 1242 i UL 1261 opracowanymi w ITE

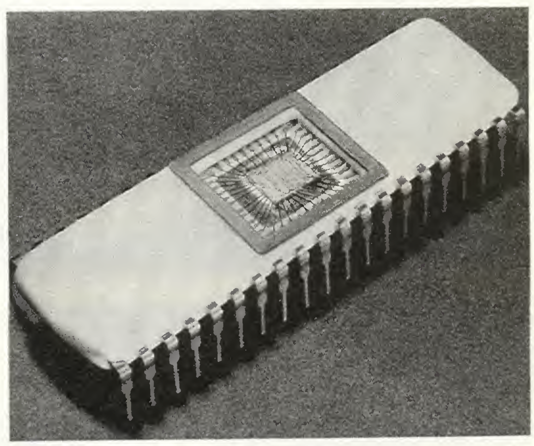
98. Moduł pamięci na domenach cylindrycznych. U dołu — powiększenia fragmentów części układu oznaczonych *1–5*. Domeny widoczne na dolnej fotografii (jasne plamki) mają średnicę ok. 6 µm



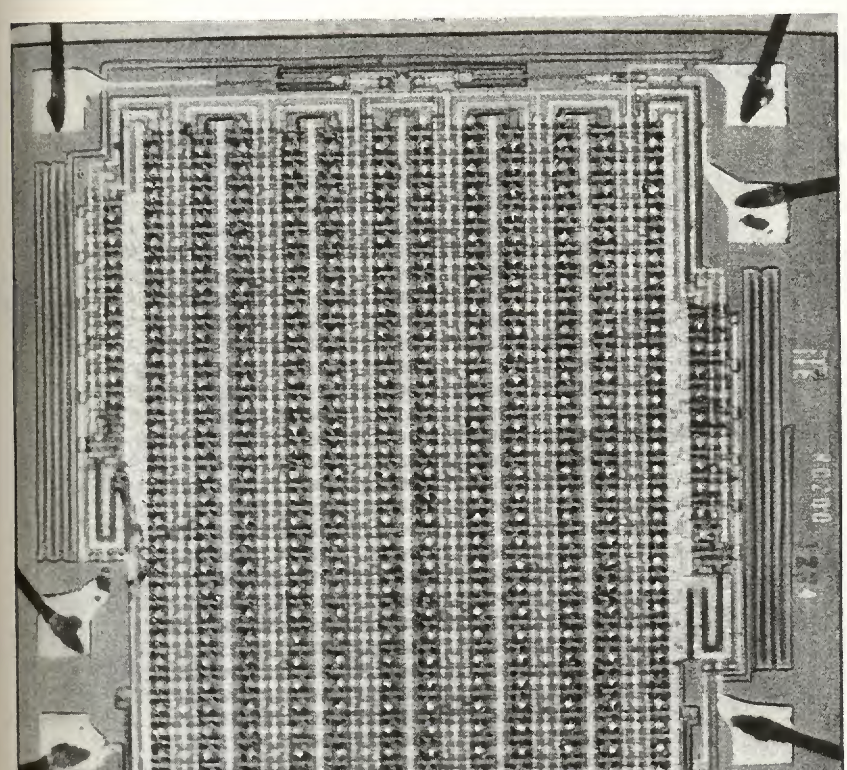
99. Kryształy monokryształów krzemu i germanu z widoczną symetrią trójkątną w kierunku (111)



101. Płytka krzemowa z gotowymi układami scalonymi



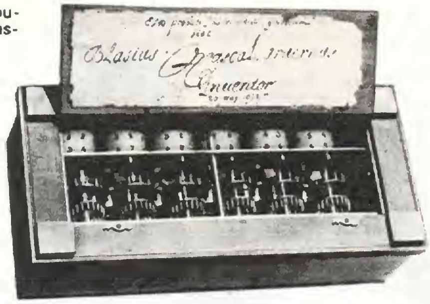
100. Układ scalony kalkulatorowy (4 działania) MCX 74204 J. Ukazany jest sposób połączenia struktury układu scalonego z nóżkami (opracowany w ITE)

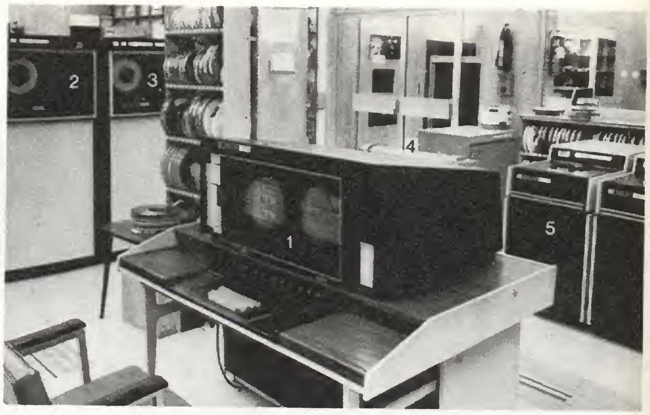




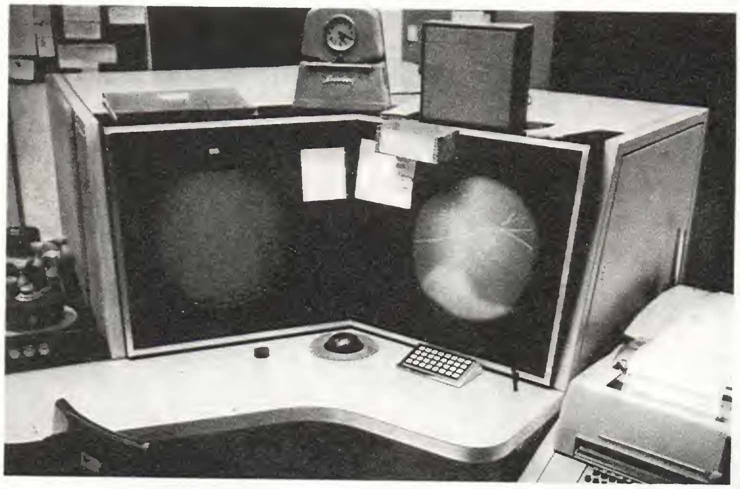
102. Rejestr 2×100 bitów — MCY 7506 L — pierwszy całkowicie polski unipolarny układ scalony wielkiej skali integracji. Stosowany jako cyfrowa linia opóźniająca lub pamięć szeregowa. a) Mikrofotografia struktury układu, b) wygląd zewnętrzny, c) z odsłoniętym układem scalonym; (opracowany w ITE)

103. Maszyna matematyczna zbudowana w 1652 r. przez B. Pascala

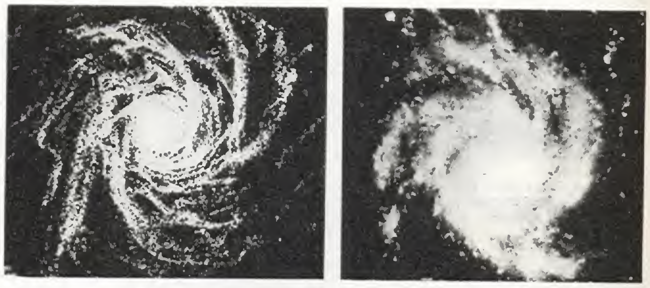




104. Widok sali z zainstalowanym komputerem Cyber 73-16; *1* konsola operatorska z monitorem ekranowym, *2* stacje taśm magnetycznych, *3* podręczny stojak z taśmami, *4* ploteks, *5* stacje dysków magnetycznych (IBJ, Świerk)

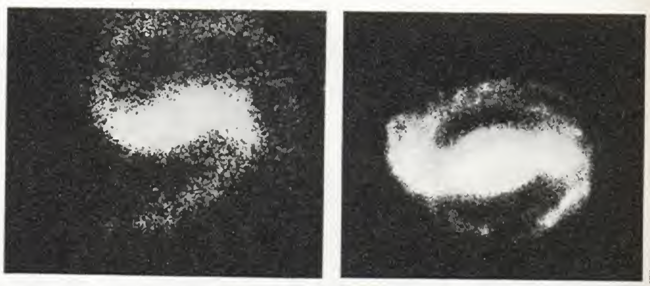


105. Konsola operatorska automatu pomiarowego Polly. Prawy monitor ukazuje obraz, lewy natomiast służy do wprowadzania informacji przez operatora na żądanie układu



a
b

106. Porównanie wyniku modelowania (a) z obserwowaną galaktyką WGC-175 (b)

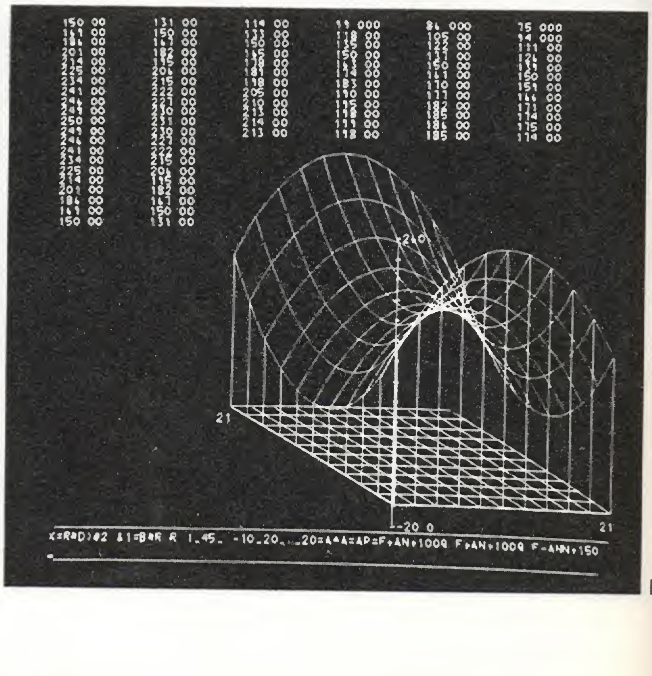


a
b

107. Porównanie wyniku modelowania (a) z obserwowaną galaktyką spiralną *M*-101 (b)



108. Interakcyjny tryb pracy: użytkownik wprowadza dane za pomocą „pióra świetlnego"; obok obraz wyników na monitorze ekranowym



a
b



109. Przenikanie się warstw w grubej próbce ciekłego kryształu cholesterolowego. Fotografia wykonana przy użyciu mikroskopu polaryzacyjnego (pow. 300 ×)

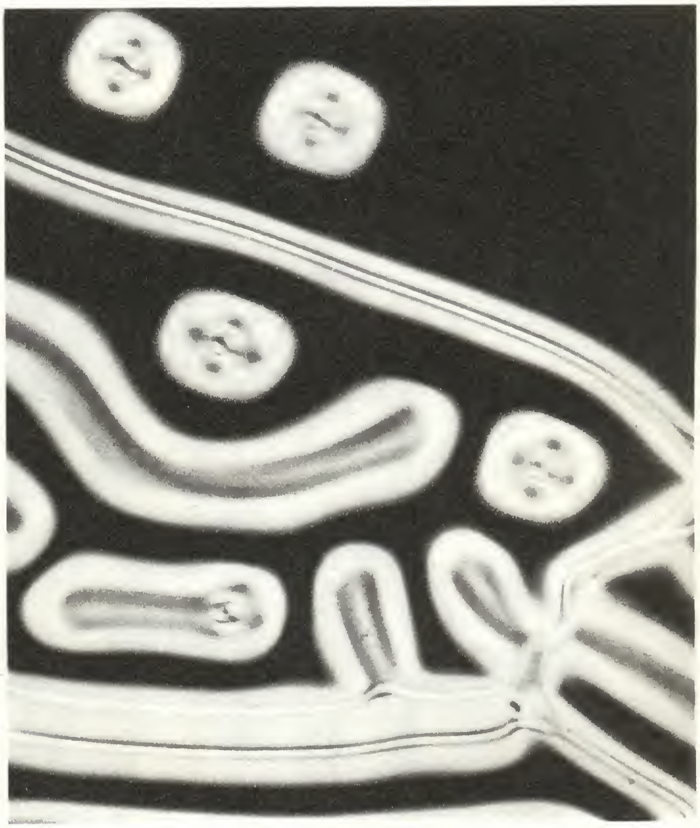


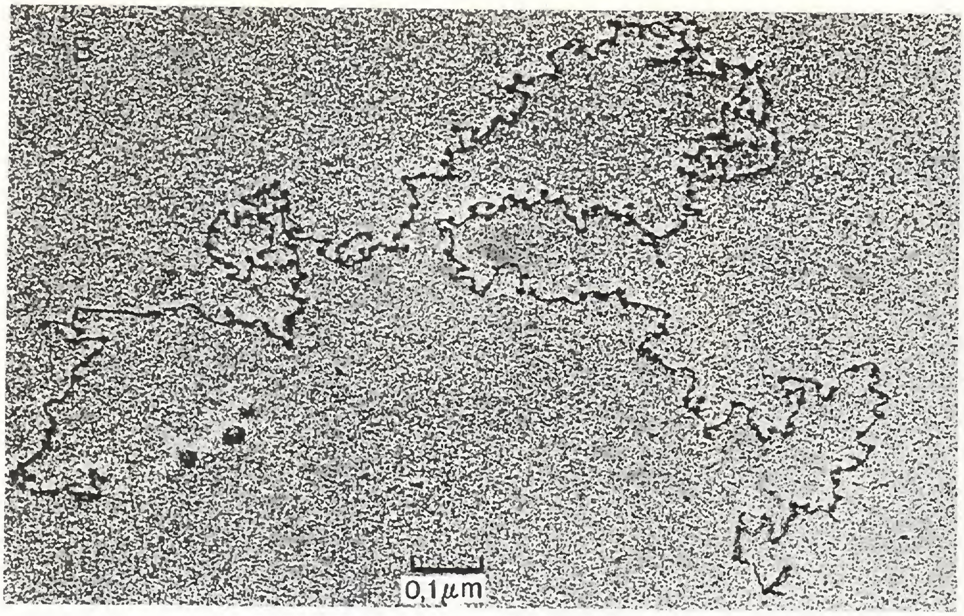
110. Upakowanie warstw cholesterolowych w cienkiej próbce (mikroskop polaryzacyjny, pow. 300 ×)

111. Uwarstwienie kropelek ciekłych kryształów cholesterolowych (mikroskop polaryzacyjny, pow. 300 ×)

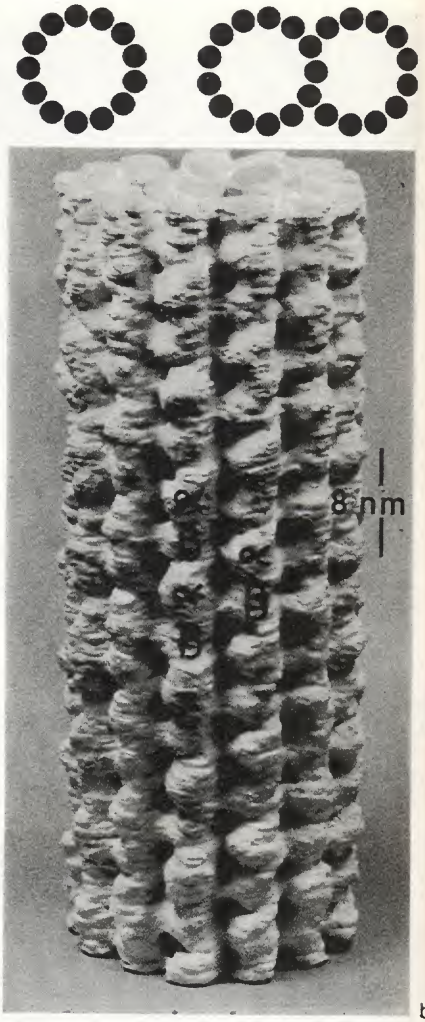


112. Modelowanie tworów mielinowych za pomocą ciekłego kryształu cholesterolowego (mikroskop polaryzacyjny, pow. 300 ×)

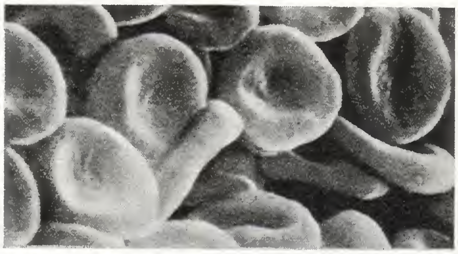




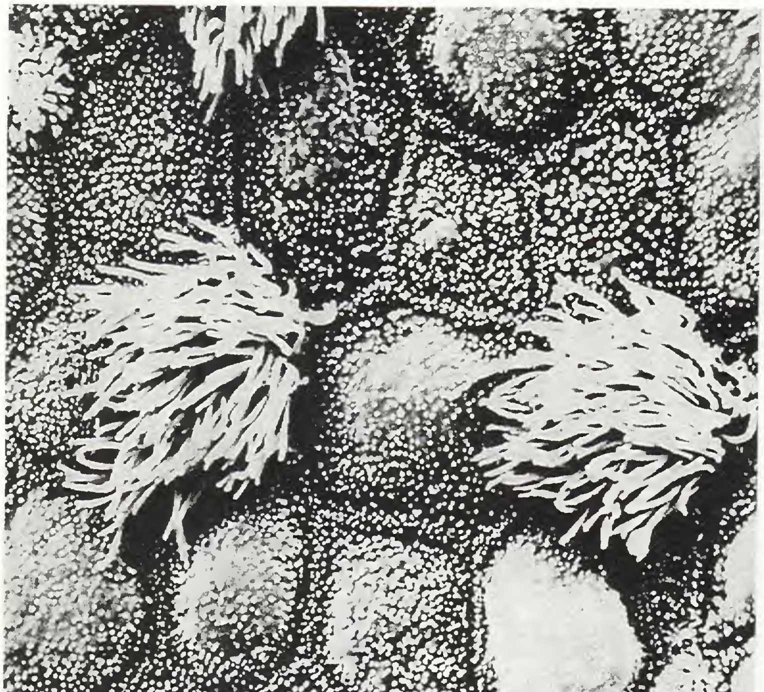
113. Obraz chromatyny otrzymany w mikroskopie elektronowym

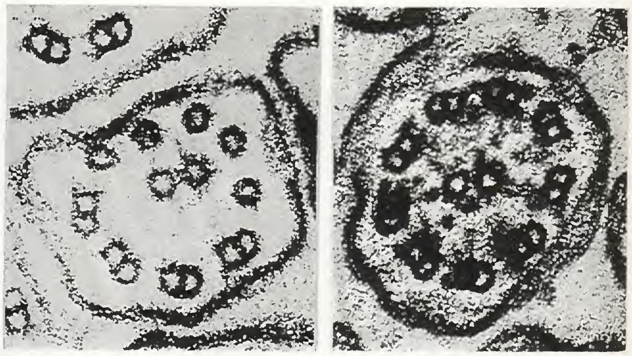


114. Mikrotubula — wewnętrzny białkowy element strukturalny komórki: a) przekrój poprzeczny mikrotubuli pojedynczej i podwójnej; b) model mikrotubuli pojedynczej, skonstruowanej na podstawie badań krystalograficznych i za pomocą mikroskopu elektronowego. Ściana mikrotubuli zbudowana jest z helikalnie ułożonych cząsteczek tubuliny i białka dimerycznego (podjednostki α i β)



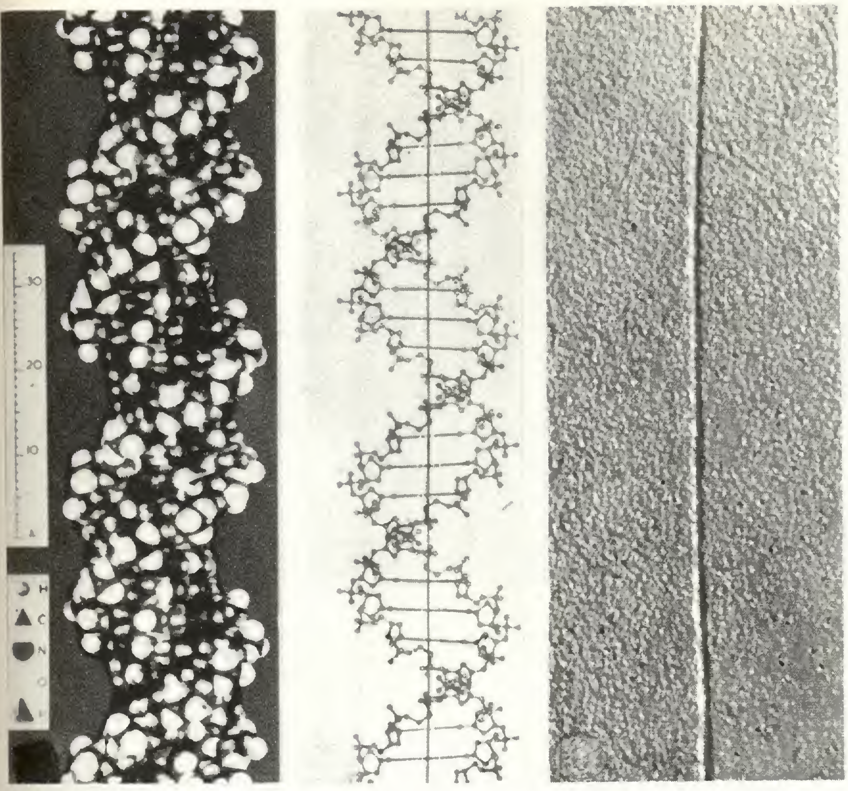
115. Obraz czerwonych komórek krwi ludzkiej otrzymany w skaningowym mikroskopie elektronowym (T. Hayes)



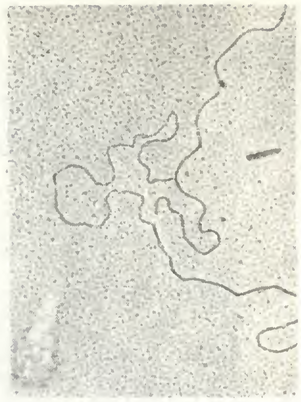




116. Rzęski w nabłonku tchawicy (a) oraz przekroje przez rzęski (b–d) wskazujące, jak jedna połowa podwójnej mikrotubuli przesuwa się wzdłuż drugiej w czasie ruchu rzęski

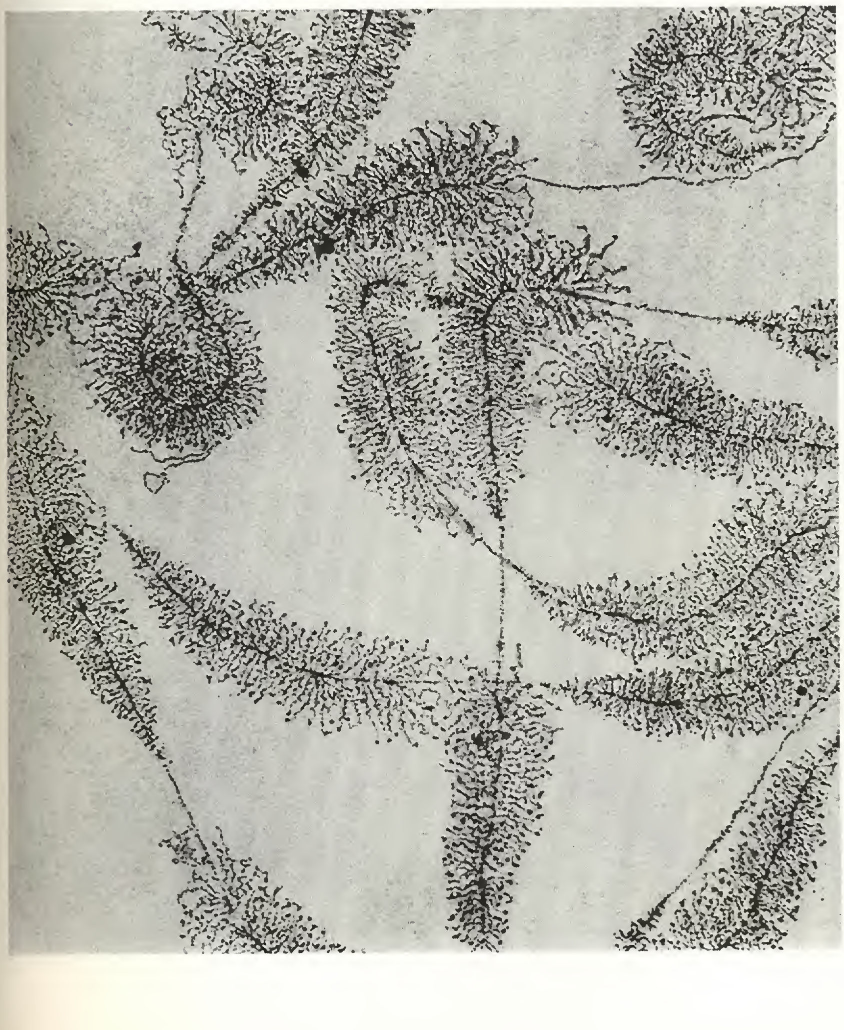


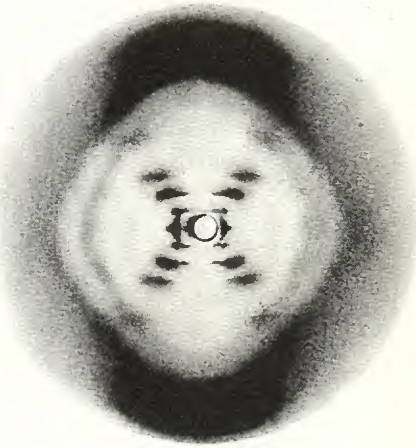
a
b
c



d

117. Makrocząsteczka DNA: a) Model heliksu DNA zbudowany z kulkowych modeli atomów. b) Model heliksu DNA wykonany z drucikowych modeli atomów. Płaszczyzny zasad są prostopadle do osi heliksu i dlatego widoczne są jako odcinki linii prostej. Skala ta sama jak a). c) Obraz wycinka makrocząsteczki DNA bakteriofaga λ otrzymany w mikroskopie elektronowym powiększającym 82100 ×. d) Obraz wycinka makrocząsteczki DNA bakteriofaga λ; widoczne miejsce, w którym dwie nici DNA zostały rozdzielone

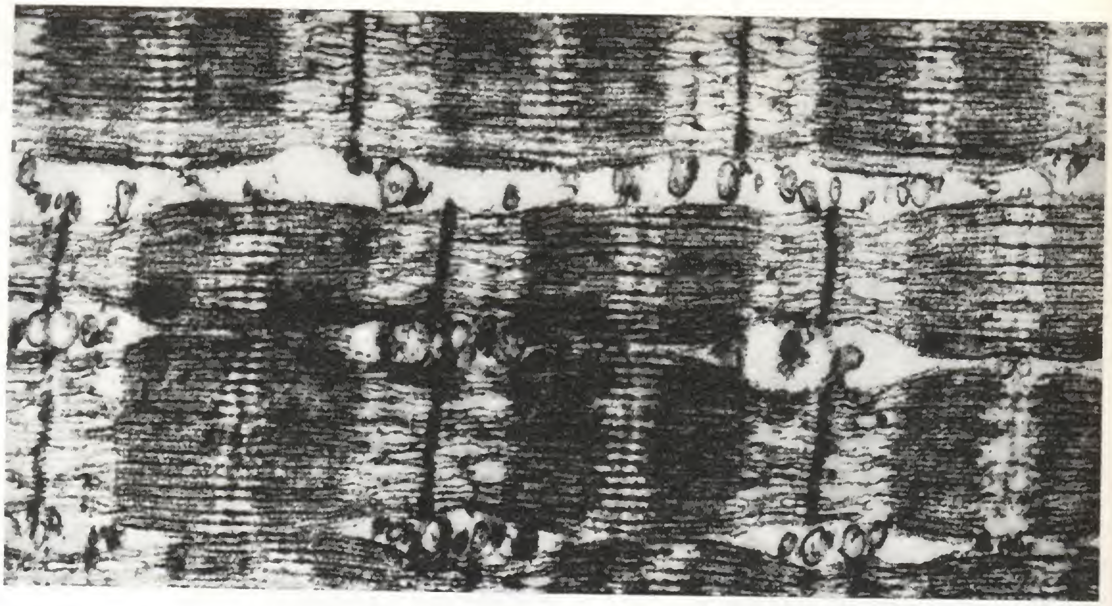




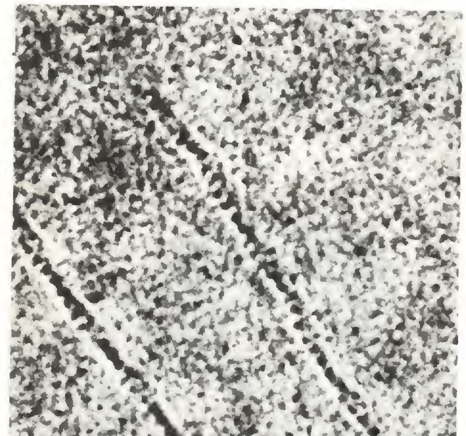
118. Rentgenogram włókien DNA. Ciemne plamy układające się w kształt krzyża świadczą o spiralnej strukturze cząsteczki DNA, a ich umiejscowienie związane jest z parametrami spirali. Dwa łukowe zaciemnienia (u góry i u dołu) świadczą o wzajemnym równoległym ułożeniu zasad oraz o ich prostopadłości do osi heliksu (R. Franklin, 1953)

119. DNA z oocytów żaby (obraz otrzymany w mikroskopie elektronowym, pow. 26000 ×) w trakcie procesu transkrypcji rybosomalnego RNA(rRNA). Wzdłuż tej samej cząsteczki widać fragmenty DNA, które nie są transkrybowane. Na aktywnym odcinku DNA zachodzi jednoczesna transkrypcja wielu cząsteczek rRNA (poprzeczne niteczki tworzące „choinkę"), początek transkrypcji przypada na czubku „choinki", u podstawy „choinki" transkrypcja się kończy i rRNA odłącza się od matrycy DNA. Przy większym powiększeniu można rozróżnić wzdłuż nici DNA globule polimerazy RNA

120. Fragment kilku włókien mięśnia poprzecznie prążkowanego; obraz otrzymany w mikroskopie elektronowym, powiększenie 53 000 × (H. E. Huxley)



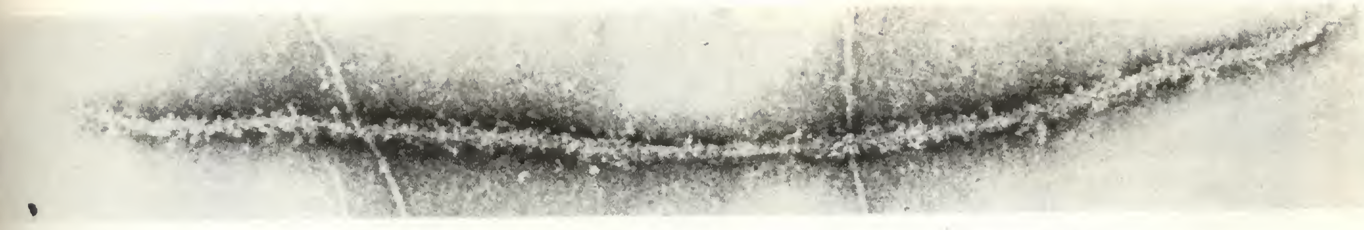
121. Polimer aktyny, obraz otrzymany w mikroskopie elektronowym. Preparat kontrastowany negatywowo octanem uranylu, powiększenie 455 000 × (J. Hanson, J. Lowy)



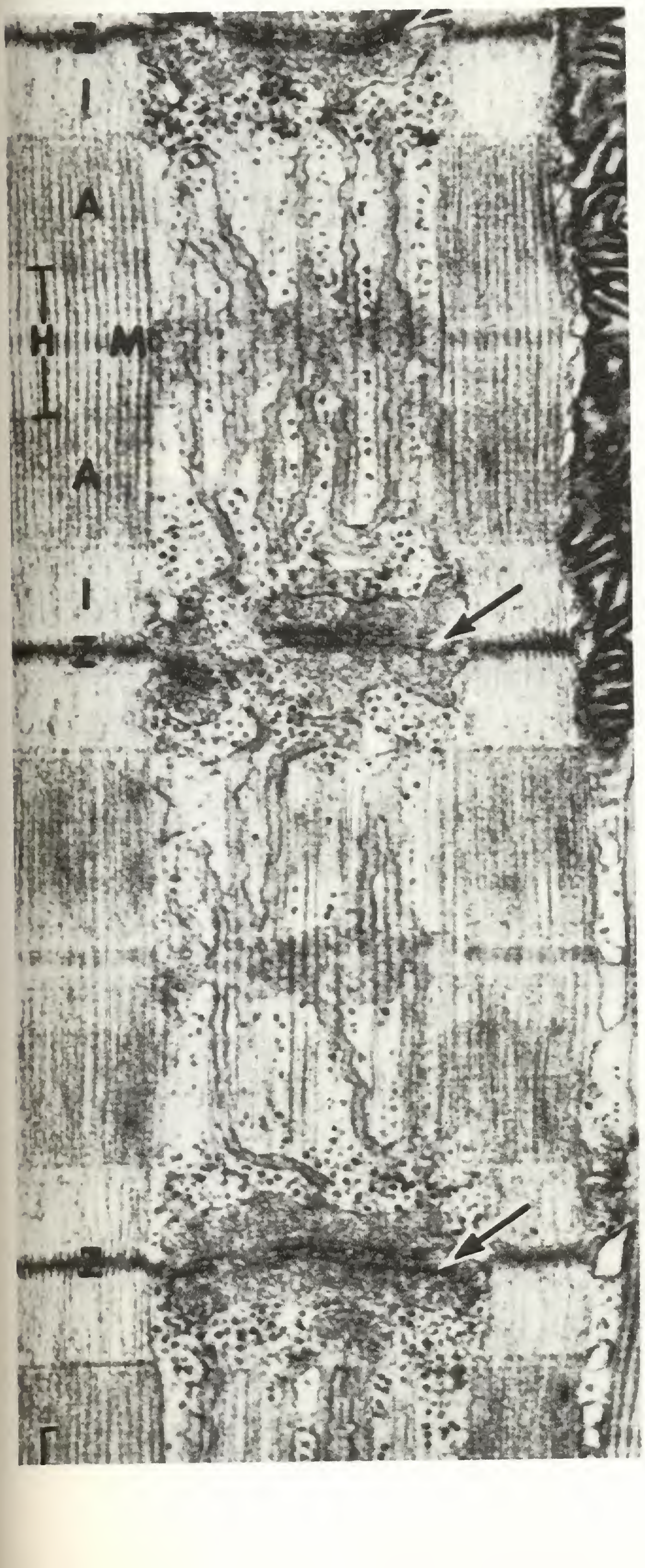


a
b

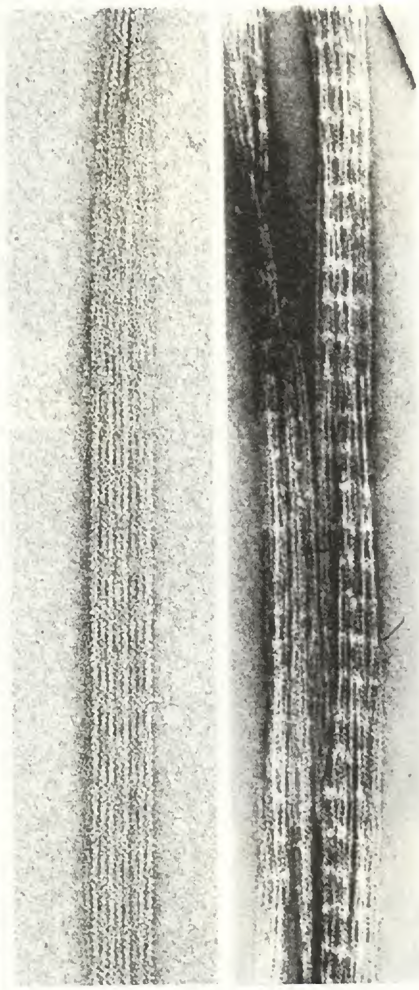
122. Struktury typu „grotów strzał" utworzone w wyniku połączenia cząsteczek subfragmentu *S1* miozyny z filamentami aktynowymi: a) Obraz otrzymany w mikroskopie elektronowym. Preparat kontrastowany negatywowo octanem uranylu. Powiększenie 180 000 ×. b) Model (P. B. Moore i in.)



123. Filament miozynowy; obraz otrzymany w mikroskopie elektronowym. Preparat kontrastowany negatywowo octanem uranylu (H. E. Huxley)

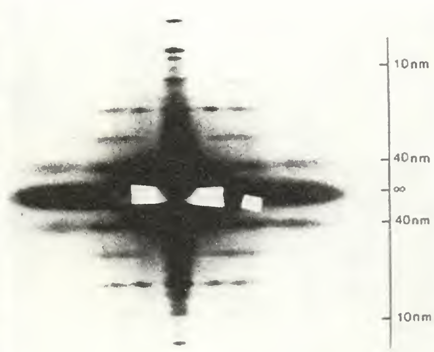


124. Obraz podłużnego przekroju przez mięsień sartorius żaby w mikroskopie elektronowym. Strzałki wskazują kanaliki poprzeczne układu *T* (wg D.S. Smitha)

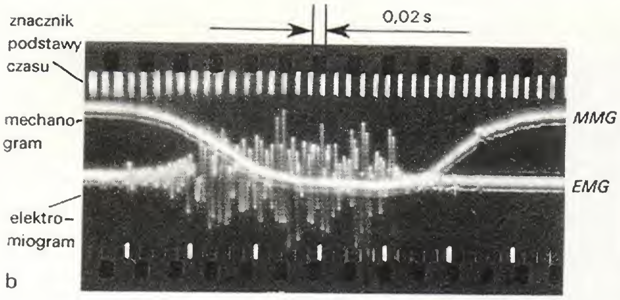


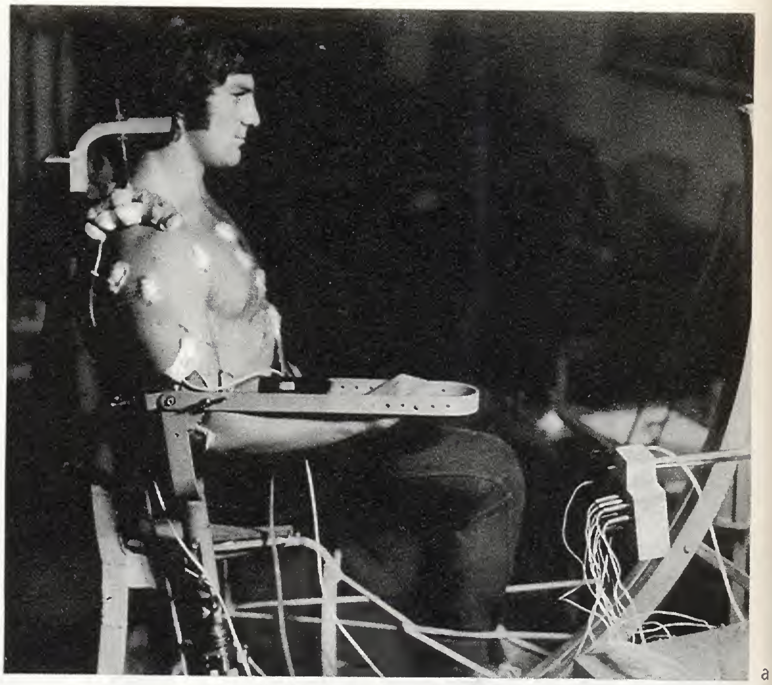
a
b

125. Obrazy otrzymane w mikroskopie elektronowym: a) agregatów filamentów czystej aktyny, powiększenie 144 000 ×, b) kompleksu aktyny z tropomiozyną i troponiną, pow. 130 000 ×; preparaty kontrastowane negatywowo octanem uranylu (A. H. Strzelecka-Gołaszewska 1972)



126. Dyfrakcja niskokątowa promieniowania rentgenowskiego na żywym mięśniu w stanie spoczynkowym. Dłuższa oś mięśnia przebiega pionowo. Widoczne są linie warstwowe i refleksy południkowe (dla uwidocznienia refleksów równikowych stosuje się krótszy czas ekspozycji) (H.E. Huxley, W. Brown)



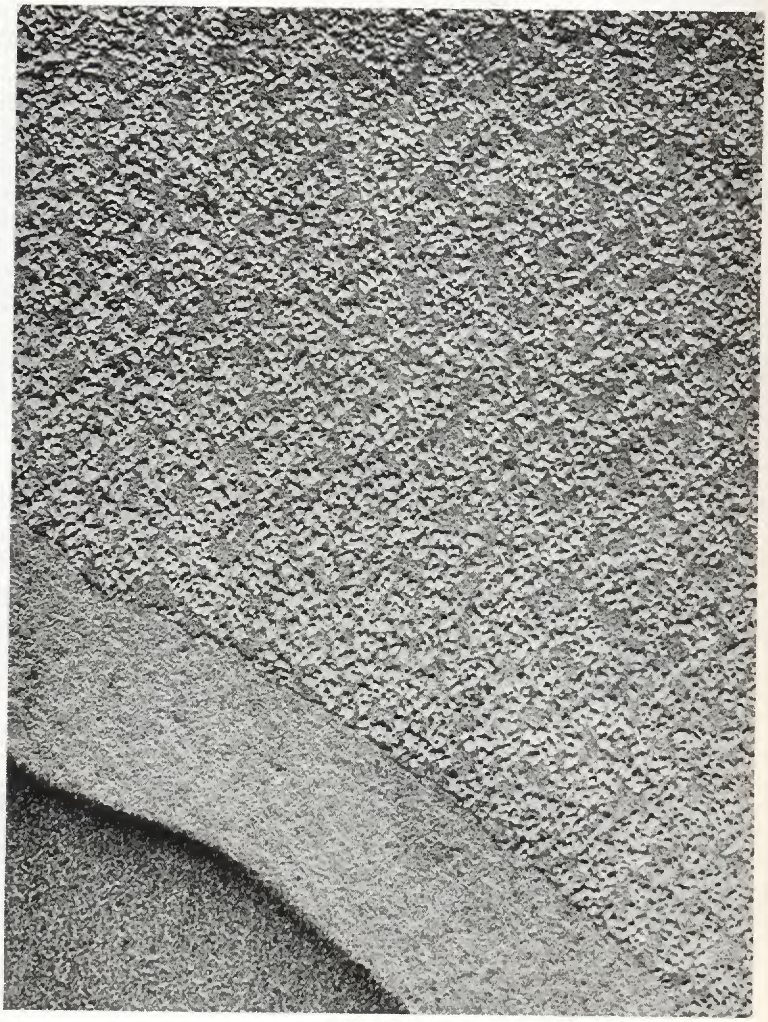


127. Badanie mięśni obsługujących staw ramienny człowieka: a) stanowisko pomiarowe; b) mechanogram i elektromiogram mięśnia dwugłowego ramienia — zapis oscylograficzny uzyskany podczas badań w warunkach statycznych

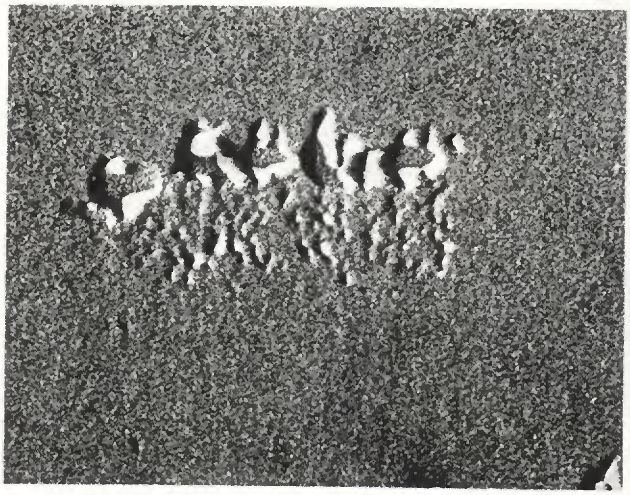
128. Obraz dojrzałego erytrocytu otrzymany w mikroskopie elektronowym; widoczna jest struktura trójwarstwowa

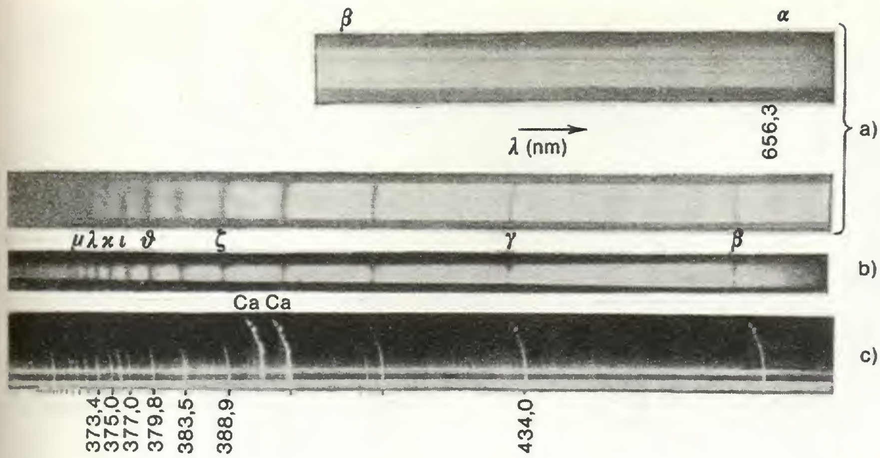


129. Błona erytrocytu, obraz otrzymany w mikroskopie elektronowym po wytrawieniu zamrożeniowym (pow. 100 000 ×)



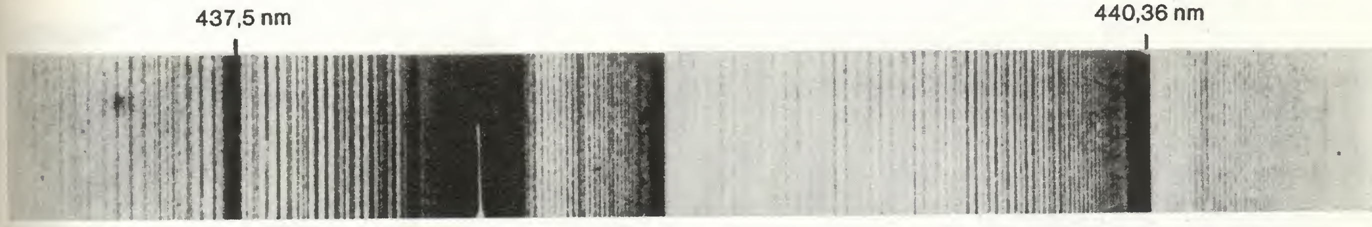
130. Dwuwarstwowa błona lipidowa, obraz otrzymany w mikroskopie elektronowym po wytrawieniu zamrożeniowym (pow. 100 000 ×)

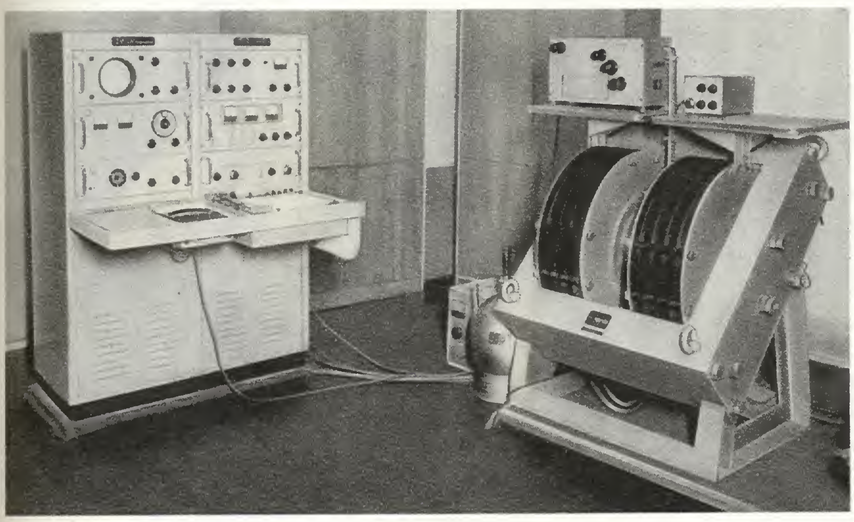




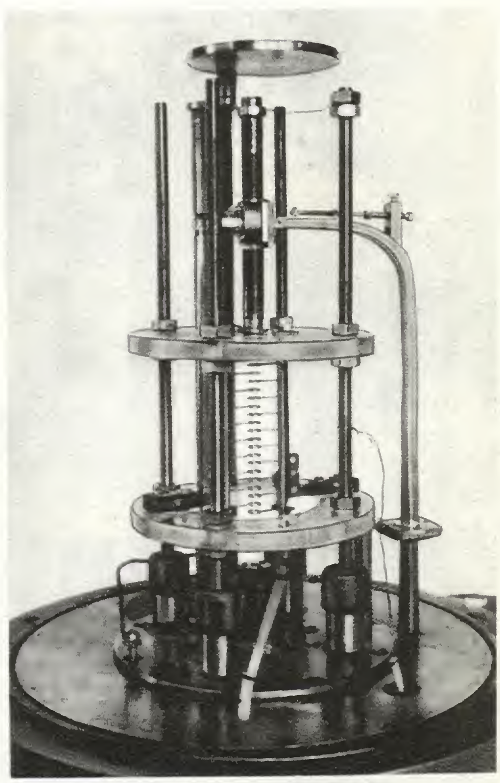
131. Seria Balmera wodoru w widmach różnych obiektów gwiezdnych: a) widmo absorpcyjne z α Pegaza, b) widmo absorpcyjne z α Liry, c) widmo emisyjne z chromosfery Słońca

132. Wycinek widma elektronowego cząsteczki NCO

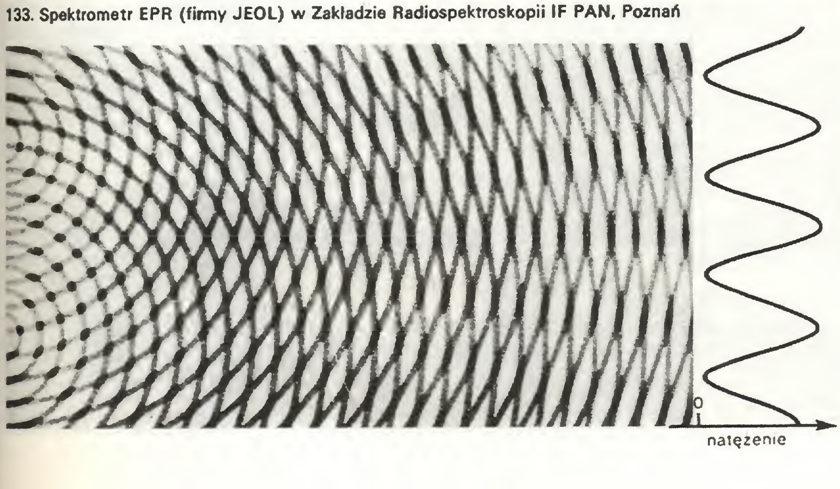




133. Spektrometr EPR (firmy JEOL) w Zakładzie Radiospektroskopii IF PAN, Poznań

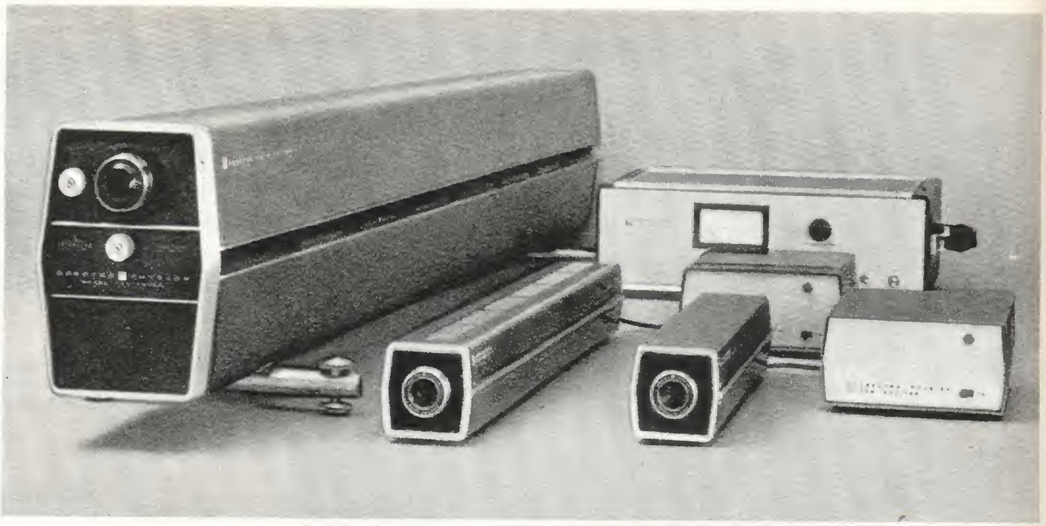


134. Maser amoniakalny skonstruowany w Instytucie Fizyki PAN

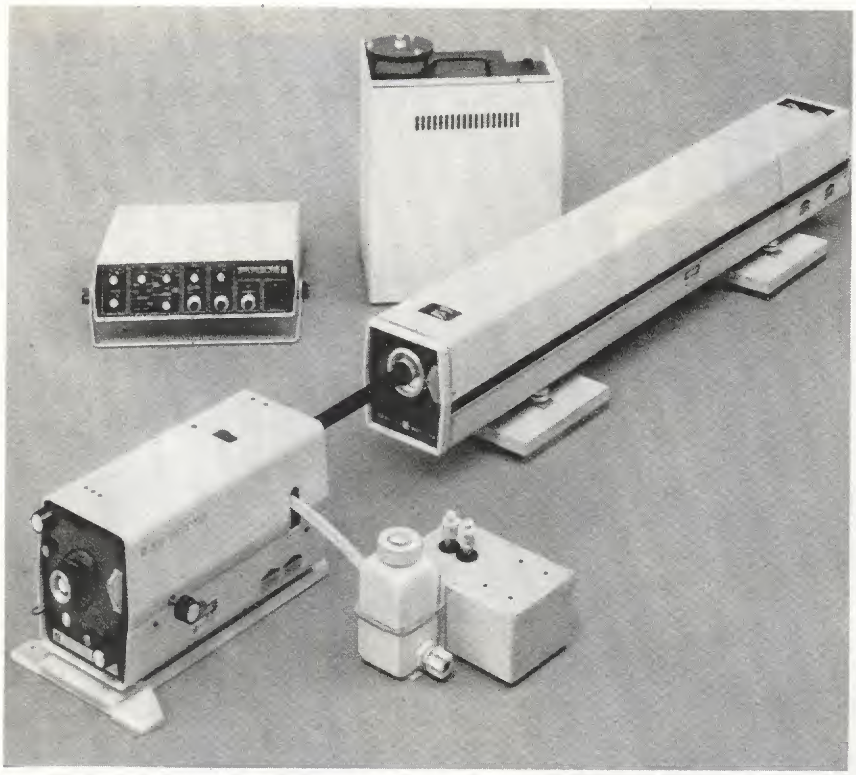


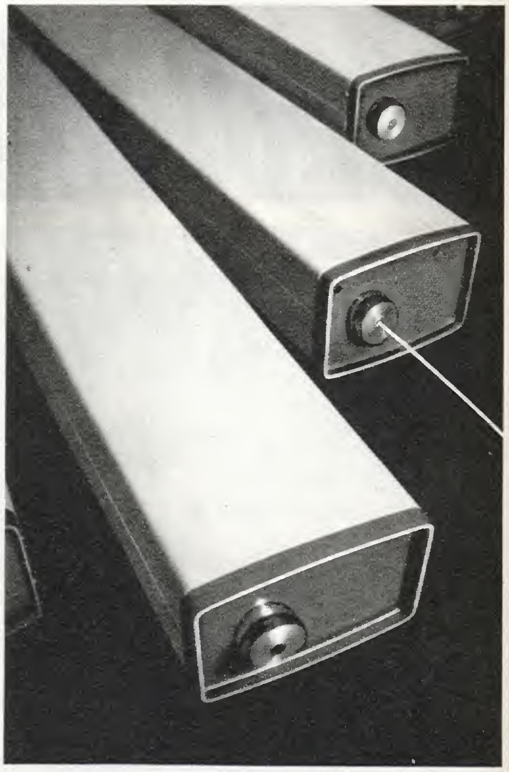
135. Fale rozchodzące się na powierzchni wody mogą tworzyć trwały układ miejsc wzmocnień i wygaszeń

136. Różne typy laserów: a) lasery helowo-neonowe o mocy 50, 15 i 5 W (typy 125 A, 124 A i 120 firmy Spectra-Physics); b) laser barwnikowy, za nim laser jonowy, pompujący (układ 580 A firmy Spectra-Physics); c) laser jonowy (firmy Coherent Radiation) i d) jego wnętrze

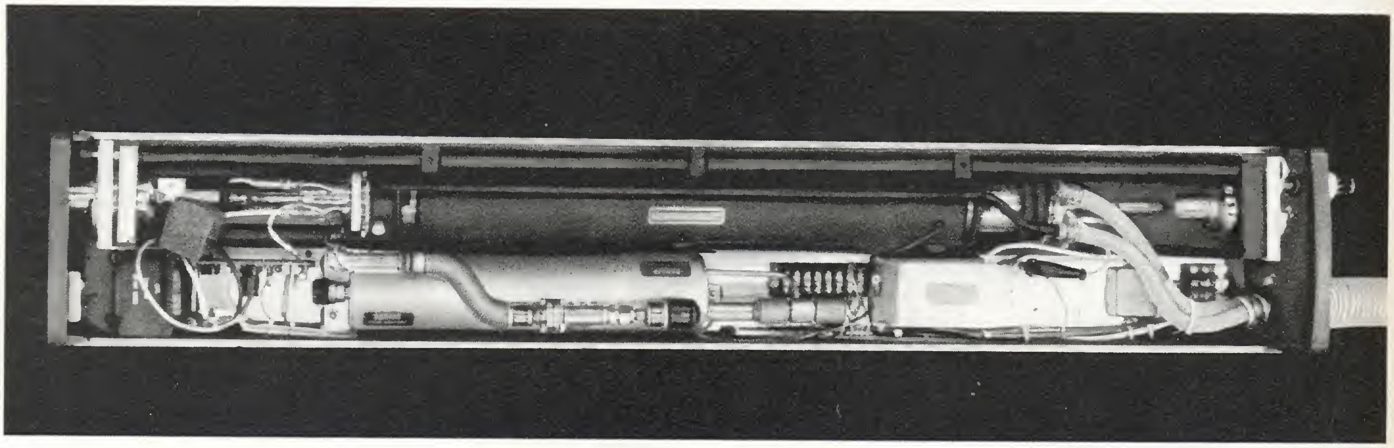


a



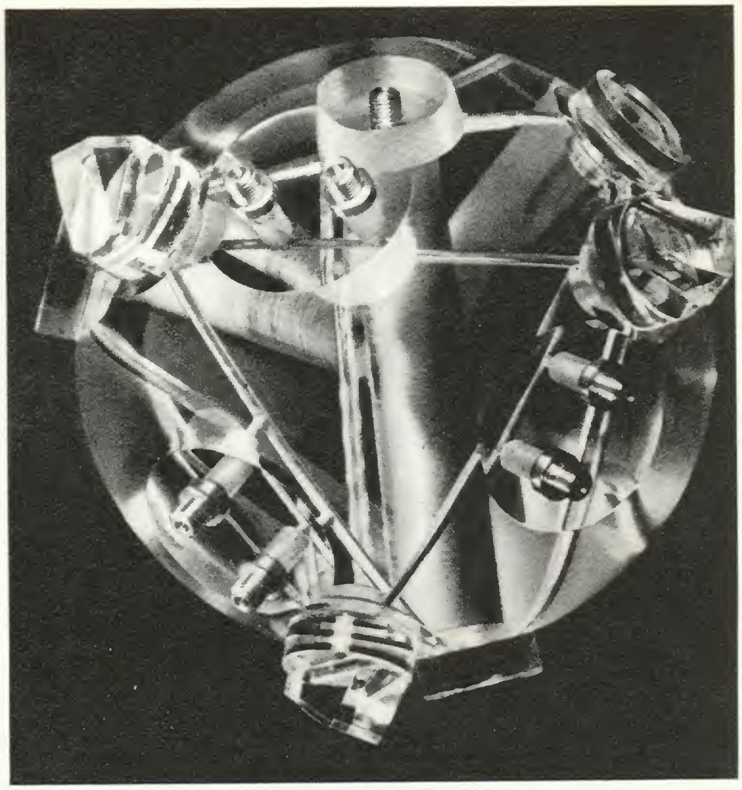


b
c



d

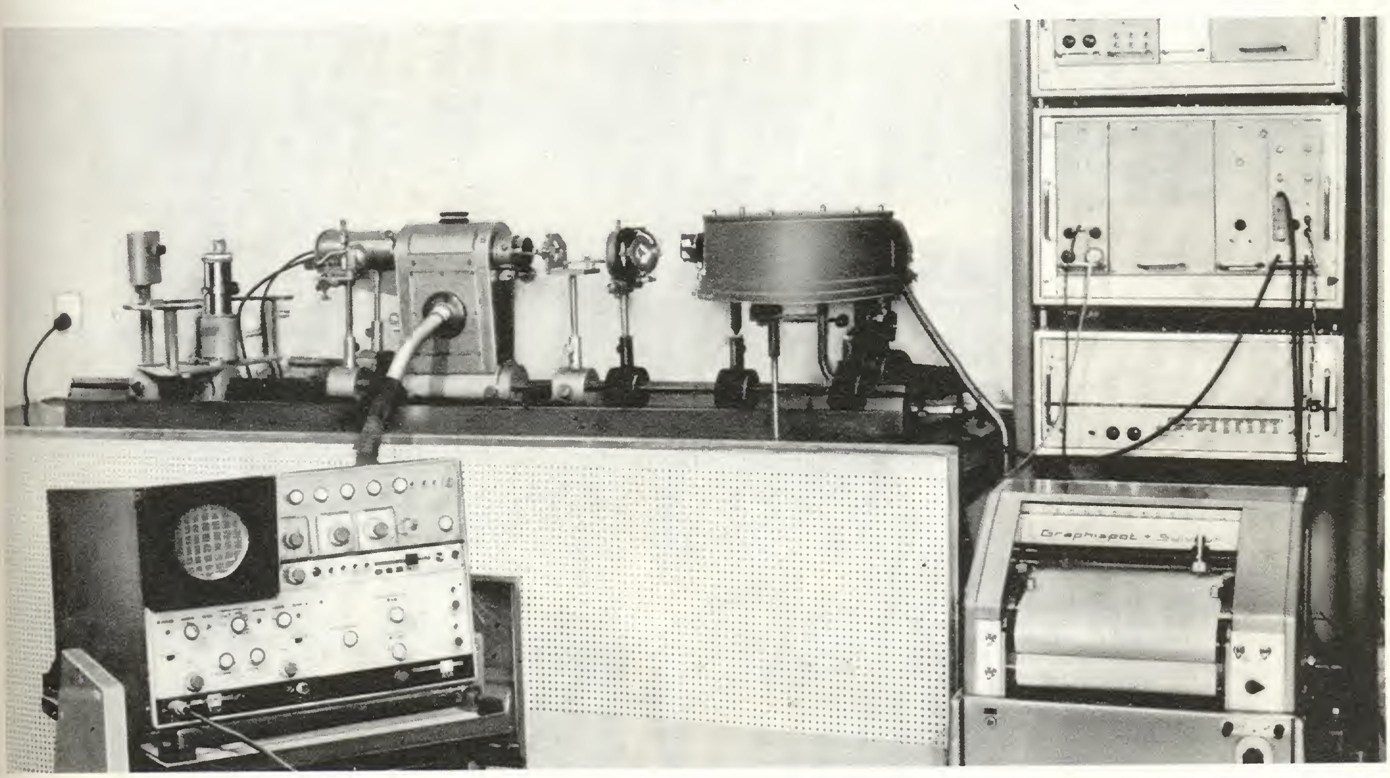


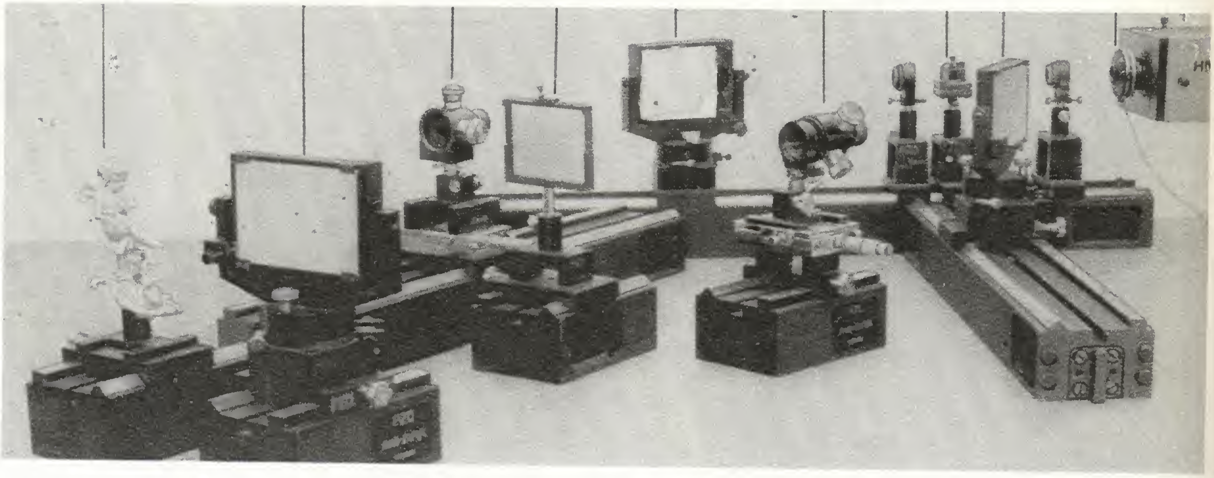


137. Głowica żyroskopu laserowego (firma Honeywell, Inc.)

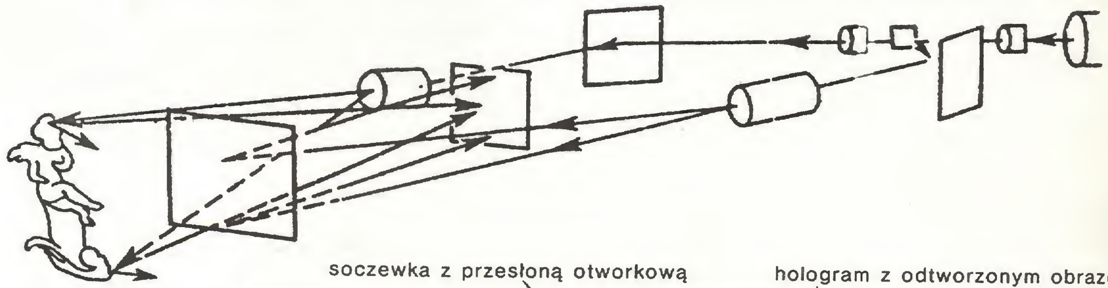
138. Uniwersalny przyrząd geodezyjny (firma ESSA) z laserem He-Cd

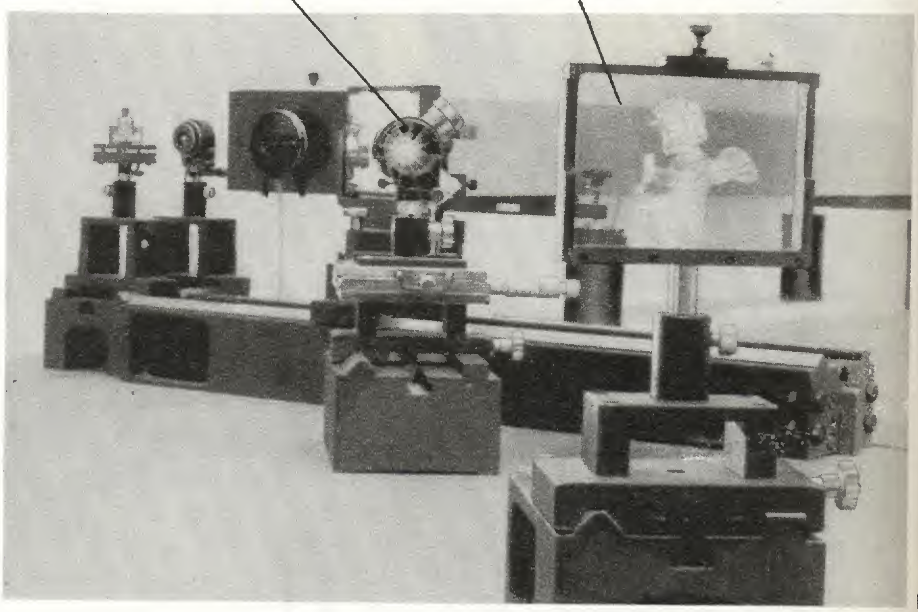
139. Aparatura do badania generacji drugiej harmonicznej światła w sproszkowanych kryształach





a

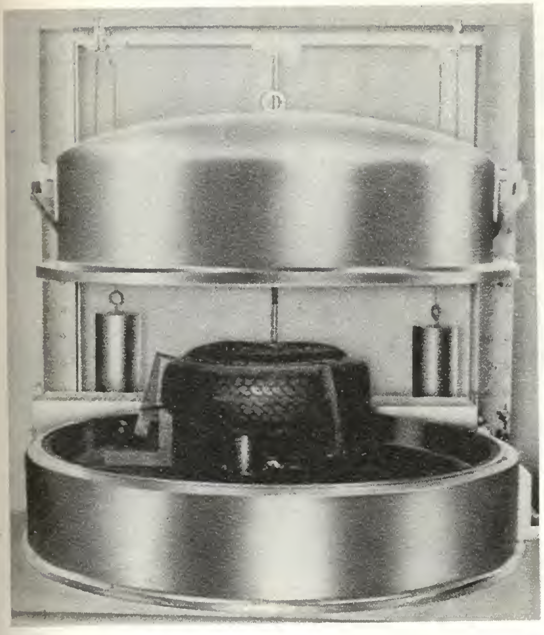




b



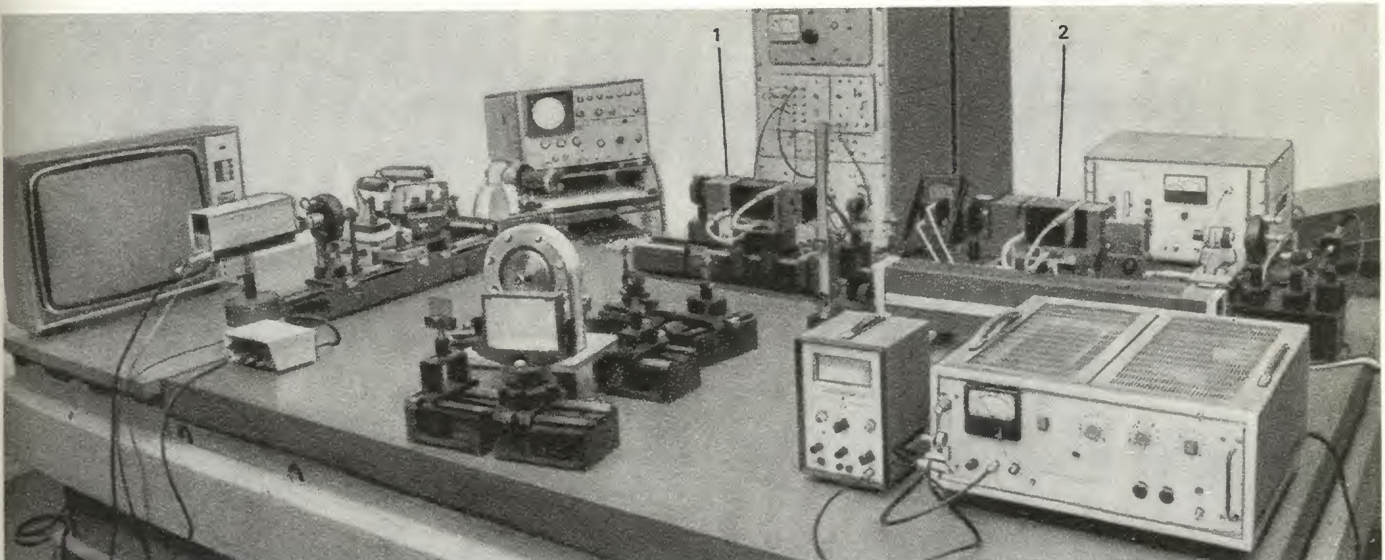
140. Holografowanie przedmiotów nieprzezroczystych: a) Układ zestawiony według schematu przedstawionego na rys. 8, str. 384 (poszczególne elementy układu są mocowane do ciężkich podstawek spoczywających na masywnym stole). b) Układ odtwarzający; przez hologram widoczny jest obraz odtworzony



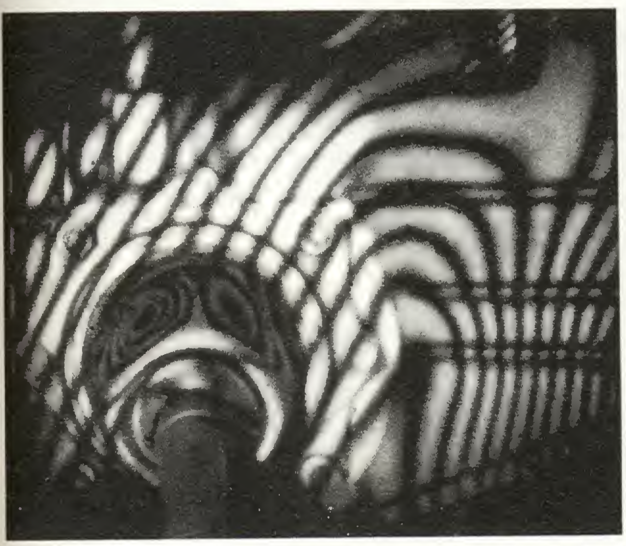


a
b

141. Holograficzne badanie opon metodą dwukrotnej ekspozycji: a) ogólny wygląd urządzenia; b) interferogram — obszar z gęstymi owalnymi prążkami świadczy o rozwarstwieniu opony



a

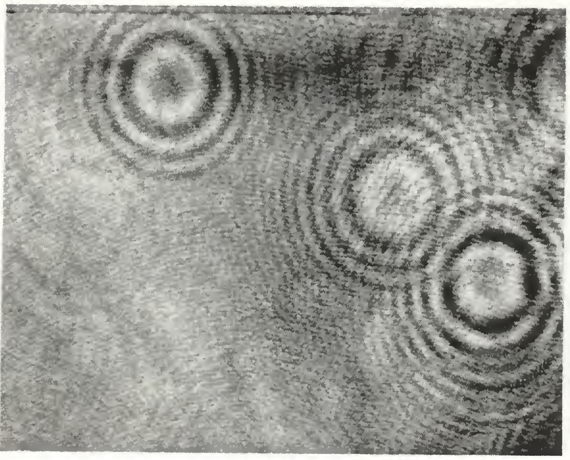


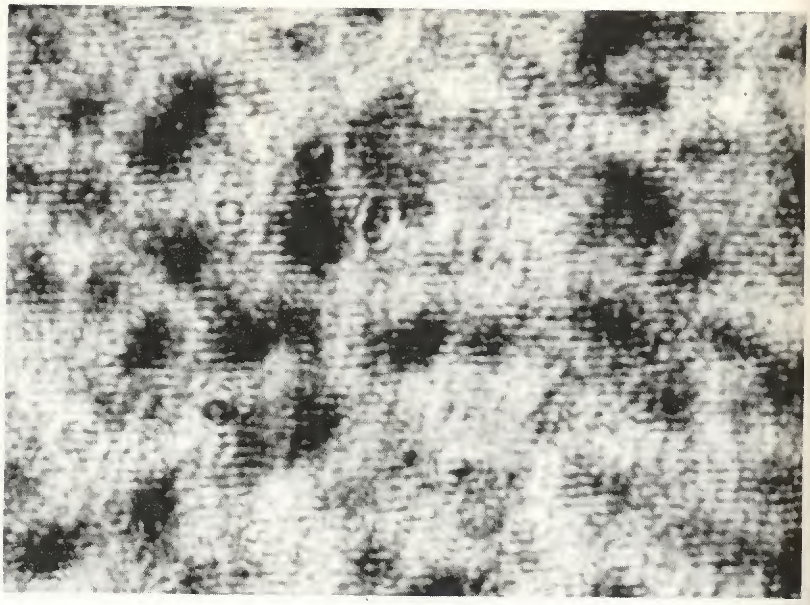


b
c

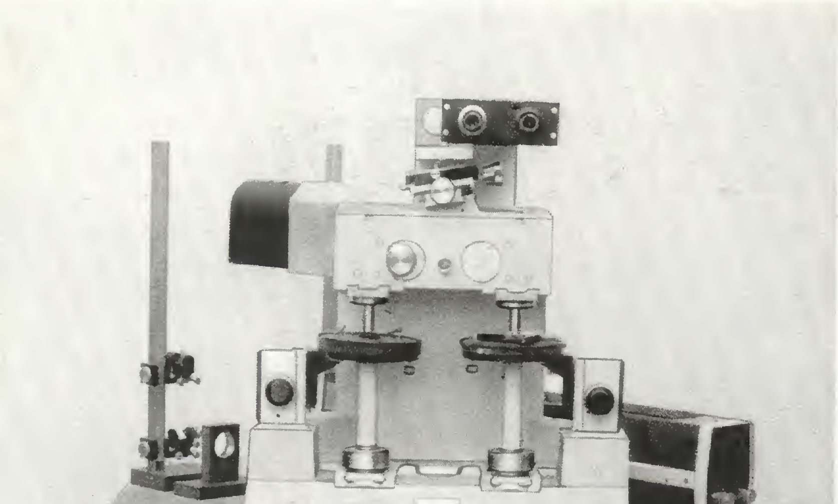
142. Holograficzne badania z użyciem lasera rubinowego: a) laboratoryjne stanowisko holograficzne z laserami rubinowymi (generator *1* i wzmacniacz *2*) do badań obiektów dynamicznych; b) interferogram drgań samochodowego zespołu napędowego; c) interferogram drgań gardła

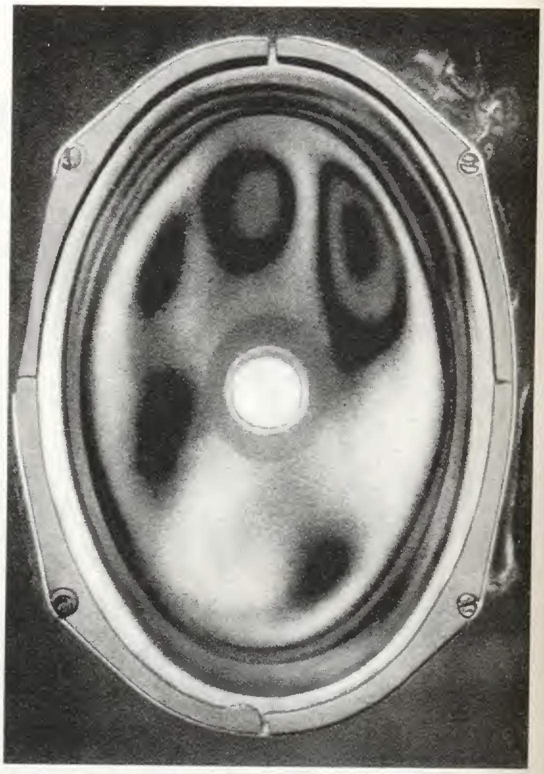
143. Hologram (z prawej strony) i jego fragment powiększony ok. 1000 razy (koncentryczne prążki wywołane są wadami elementów optycznych układu holograficznego)





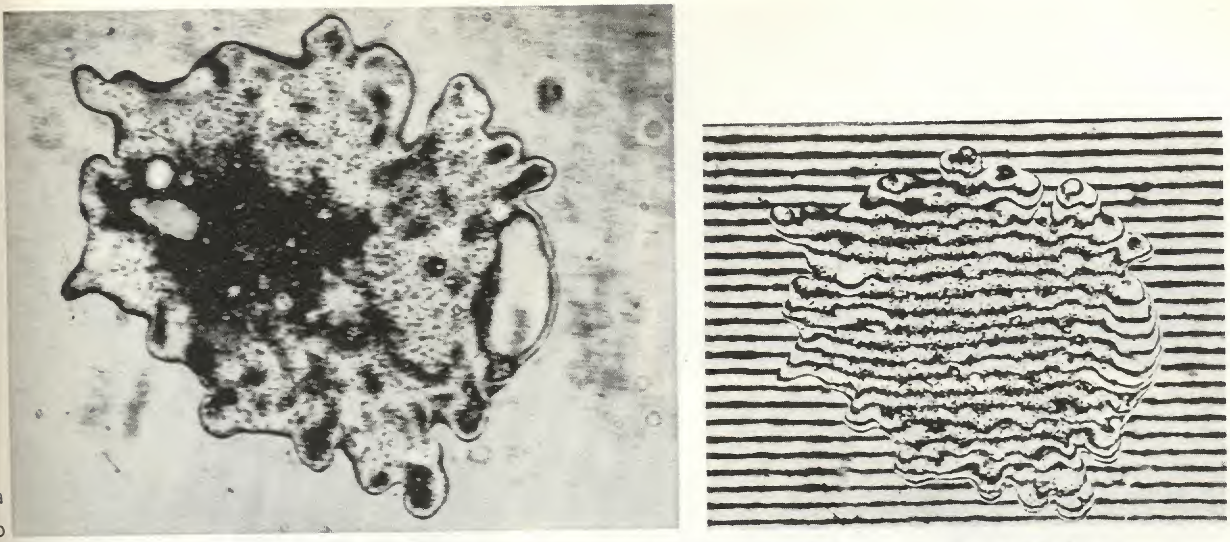
a
b





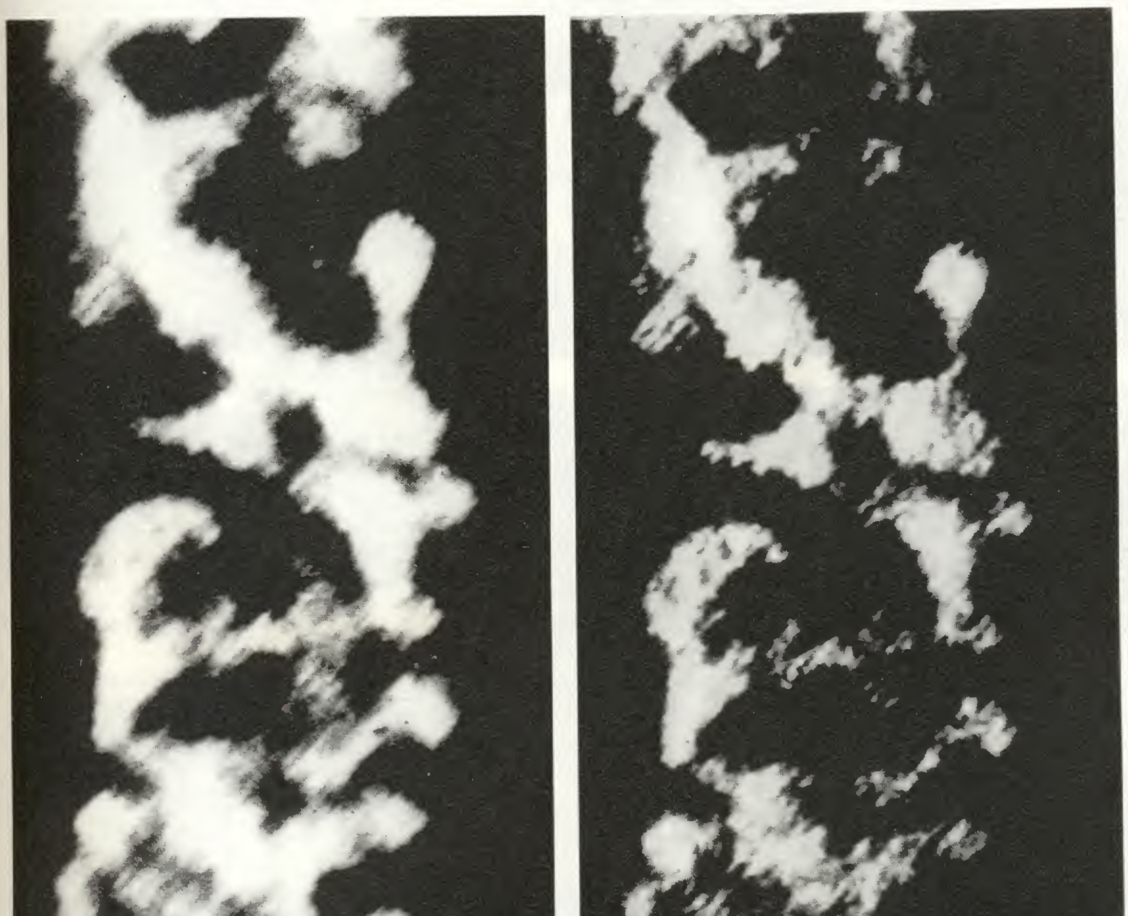
144. Interferogram uszkodzonej membrany głośnika drgającej z częstością 1000 Hz

145. Mikrointerferometr holograficzny



a
b

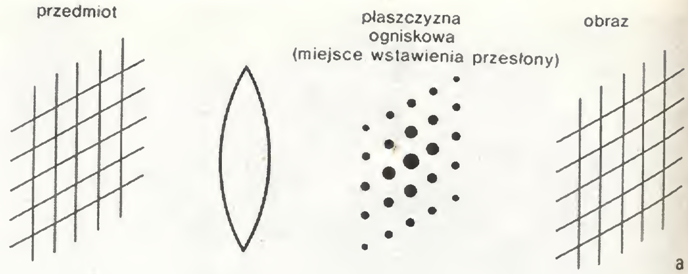
146. Mikrofotografie interferogramów ameby uzyskane w mikrointerferometrze holograficznym metodą czasu rzeczywistego przy różnych warunkach pracy urządzenia: a) w polu jednorodnym, b) w polu prążkowym



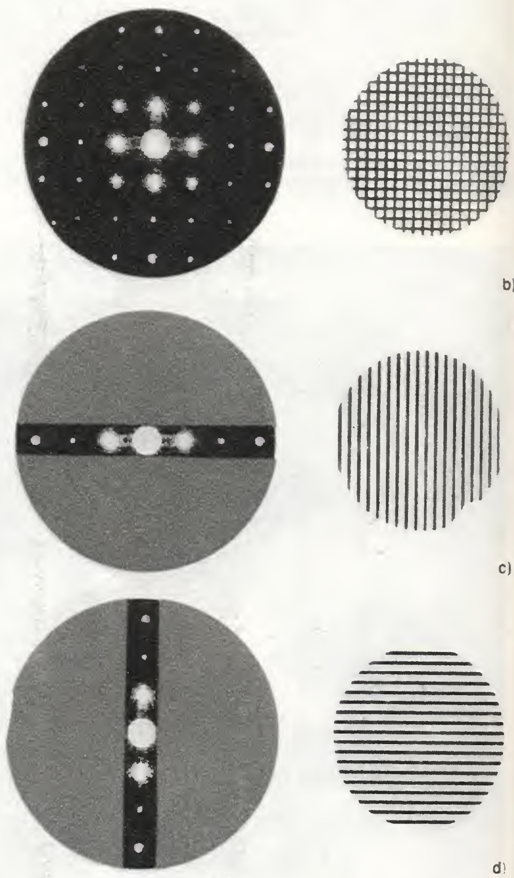
a
b

147. Zdjęcia wirusa fd uzyskane za pomocą mikroskopu elektronowego, ukazujące strukturę podwójnego heliksu wirusa: a) przed zastosowaniem holograficznej metody wyostrzenia, b) po jej zastosowaniu (G. W. Stroke i in. 1972)

148. Doświadczenie E. Abbego i A. Portera — pierwszy eksperyment z filtracją częstości przestrzennych sygnałów optycznych: a) schemat układu optycznego, b) obraz Fouriera przedmiotu-siatki i obraz wyjściowy, c) przefiltrowany przez szczelinę poziomą obraz Fouriera siatki i obraz wyjściowy, d) przefiltrowany przez szczelinę pionową obraz Fouriera siatki i obraz wyjściowy



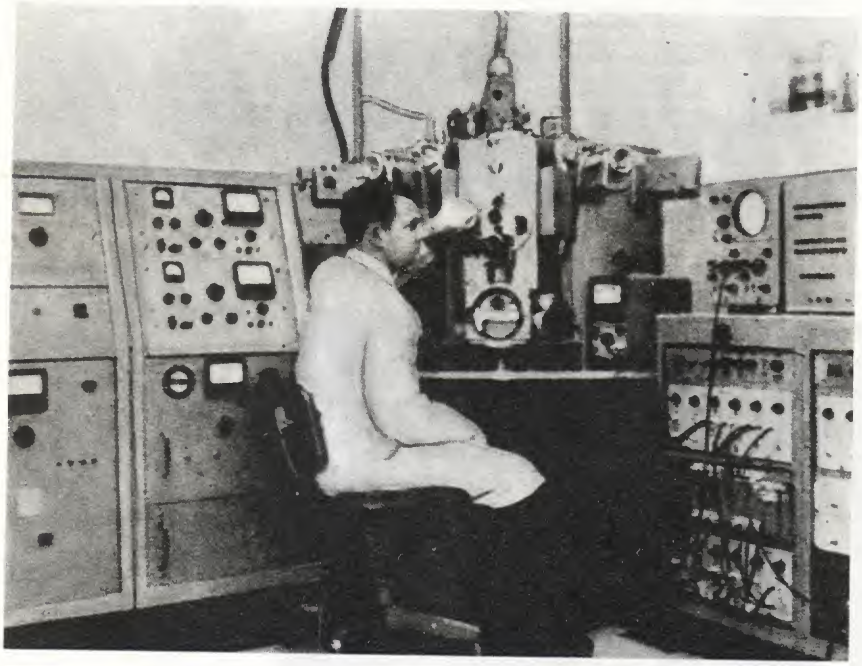
a



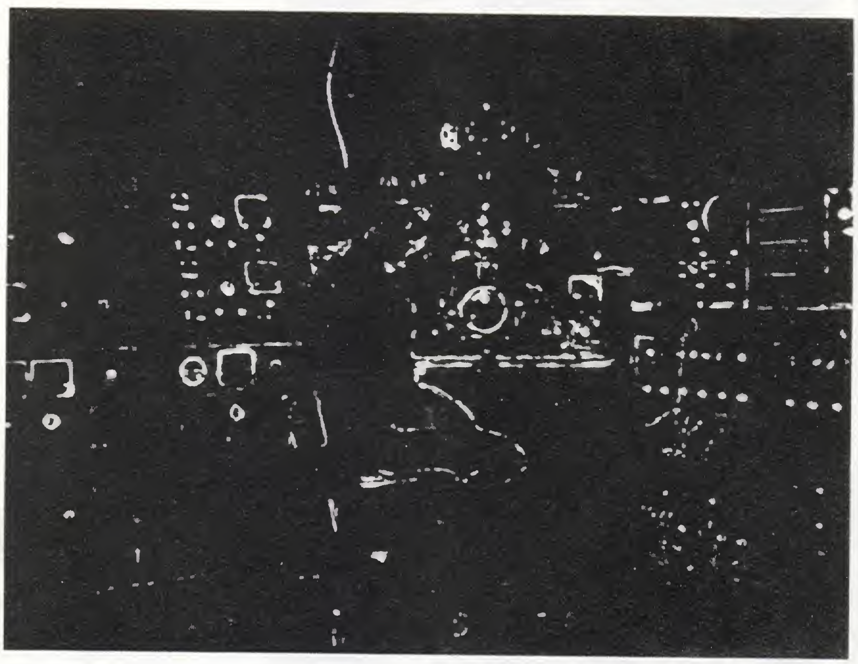
b)

c)

d)



a



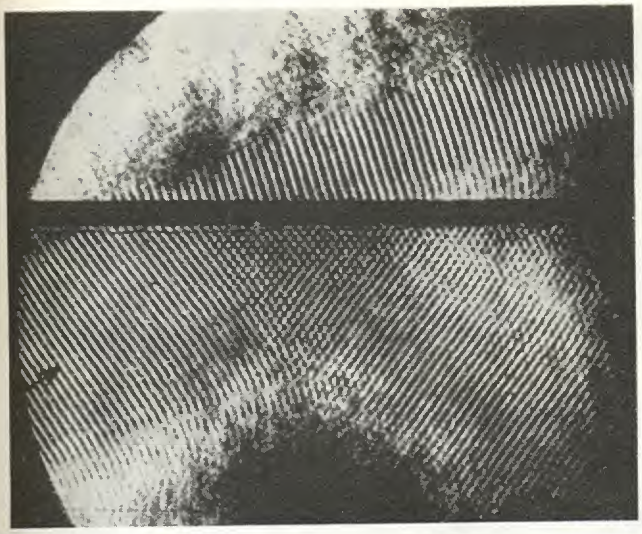
b

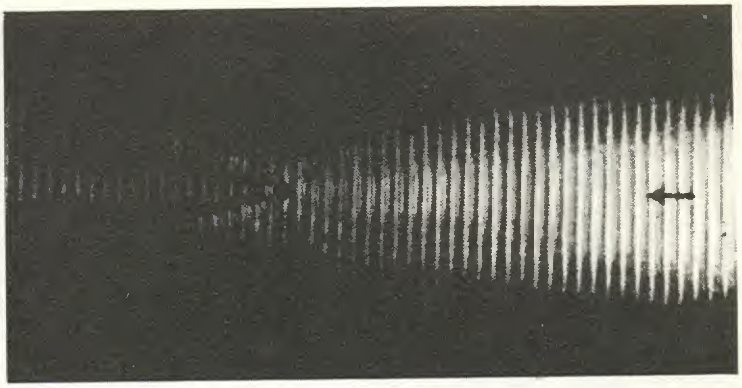
149. Filtracja przestrzenna: a) obraz przed filtracją, b) obraz po użyciu górnoprzepustowego filtru przestrzennego



a b

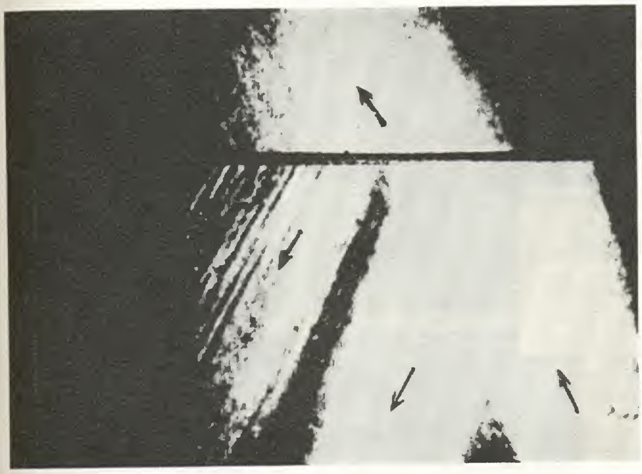
150. Rastrowy portret K. E. Bethego, złożony z biało-czarnych plamek (a), został zamieniony w obraz tonowany (b) dzięki filtracji przestrzennej usuwającej najwyższe częstości (R. A. Philips, 1969)

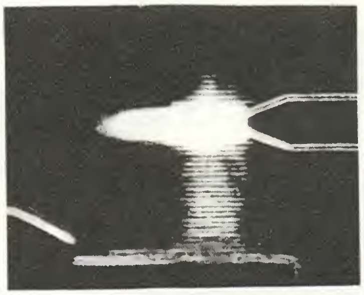




151. Odbicie i załamanie fali ultradźwiękowej padającej na granicę rozdziału oleju i roztworu soli kuchennej. Strzałka oznacza kierunek fali padającej (wg W. A. Krasilnikowa)

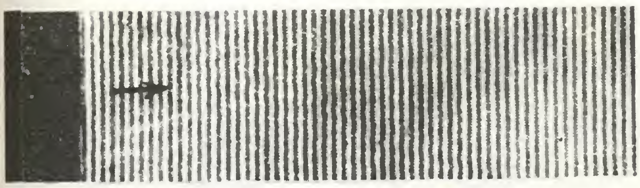
152. Dyfrakcja płaskiej fali ultradźwiękowej o długości λ = 1,6 mm na walcu o średnicy 1,8 mm (wg W. A. Krasilnikowa)



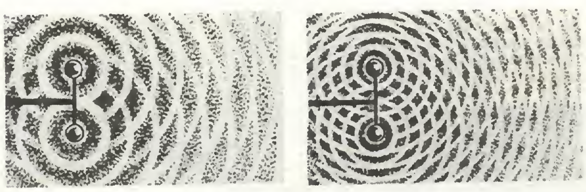


153. Obraz cieniowy fali ultradźwiękowej generowanej przez strumień powietrza wypływający z dyszy (wg L. Bergmana)

154. Odbicie i wtórne wypromieniowanie fali ultradźwiękowej z płytki mosiężnej, w której powstaje fala Lamba

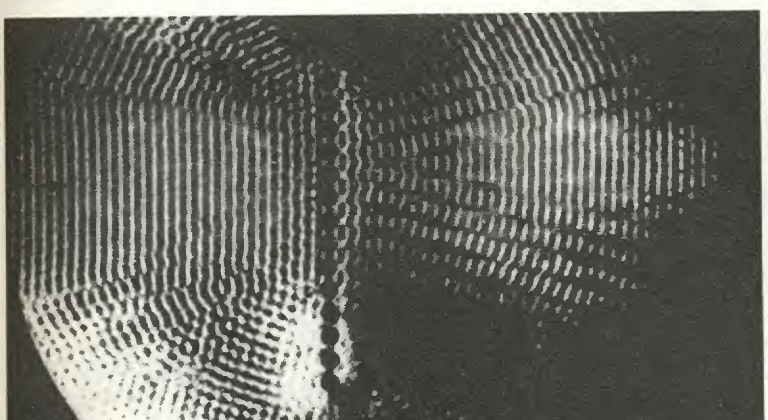


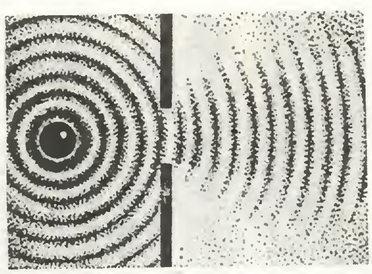
155. Fotografia cieniowa płaskiej fali ultradźwiękowej o częstości około 1,5 MHz w cieczy



a
b

156. Interferencja dwóch fal na powierzchni wody: a) fale o częstości drgań mniejszej, b) — większej



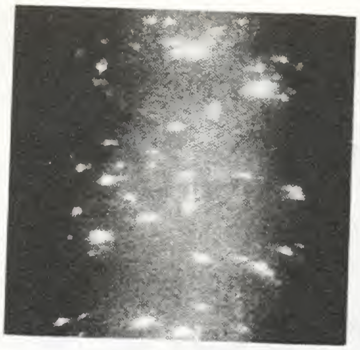


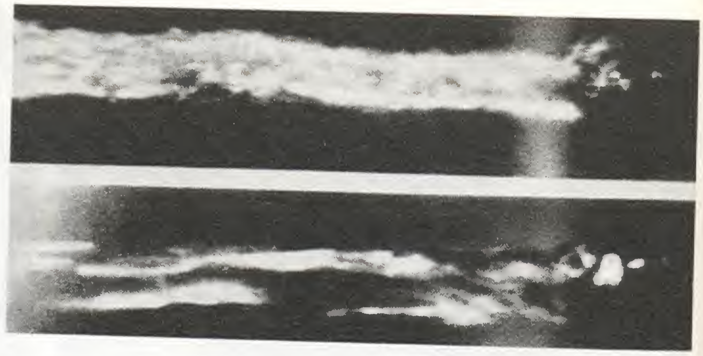
a
b

157. Dyfrakcja fali akustycznej: a) na układzie szczelin (siatka dyfrakcyjna), b) na pojedynczej szczelinie

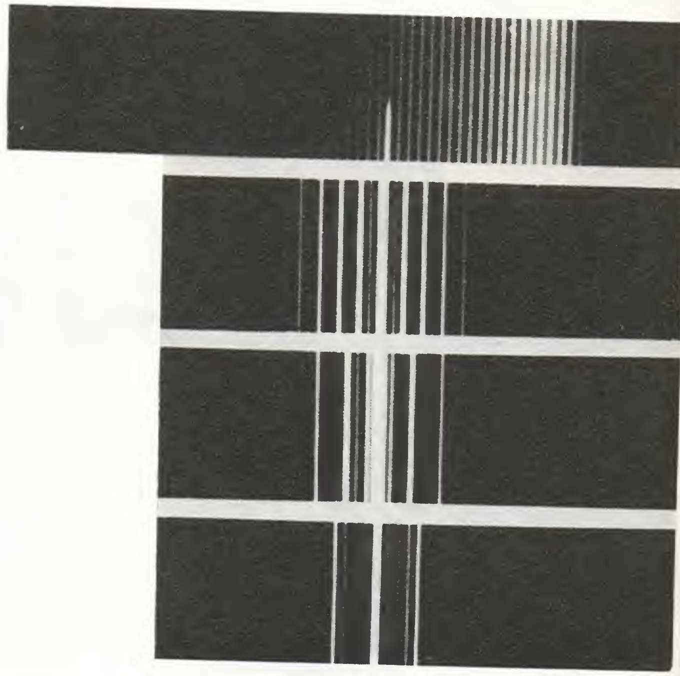
158. Pęcherzyki kawitacyjne w polu ultradźwiękowym w wodzie po przekroczeniu progu natężenia

159. Sonoluminescencja w fontannie ultradźwiękowej w cieczy (wg Rosenberga)





160. Ugięcie światła na fali ultradźwiękowej zniekształconej nieliniowo (o dużym natężeniu 15,1 W/cm², o częstości $f = 583$ kHz: a) fala zawiera wyższe harmoniczne, b) z fali ultradźwiękowej wydzielono przy użyciu filtra akustycznego drugą harmoniczną, c) trzecią harmoniczną, d) czwartą harmoniczną (wg Michajłowa i Szutiłowa)



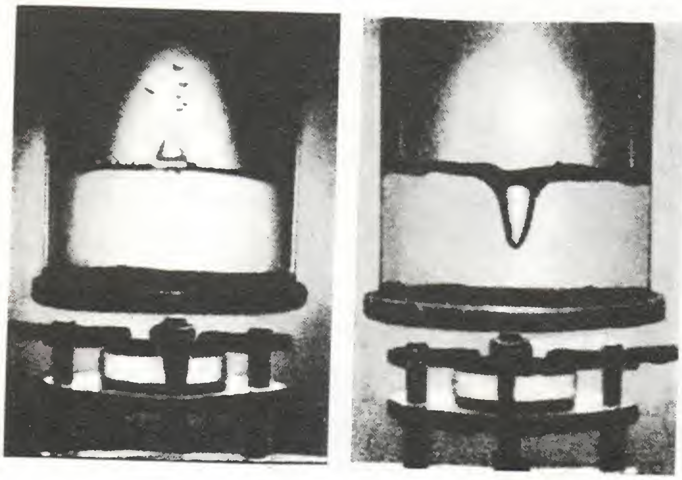
a

b

c

d

161. Fontanna wywołana ciśnieniem promieniowania fali ultradźwiękowej na granicy dwóch nie mieszających się cieczy: a) woda nad czterochlorkiem węgla (prędkość dźwięku odpowiednio $c_1 = 1500$ m/s i $c_2 = 940$ m/s), b) woda nad aniliną ($c_1 = 1500$ m/s i $c_2 = 1660$ m/s)



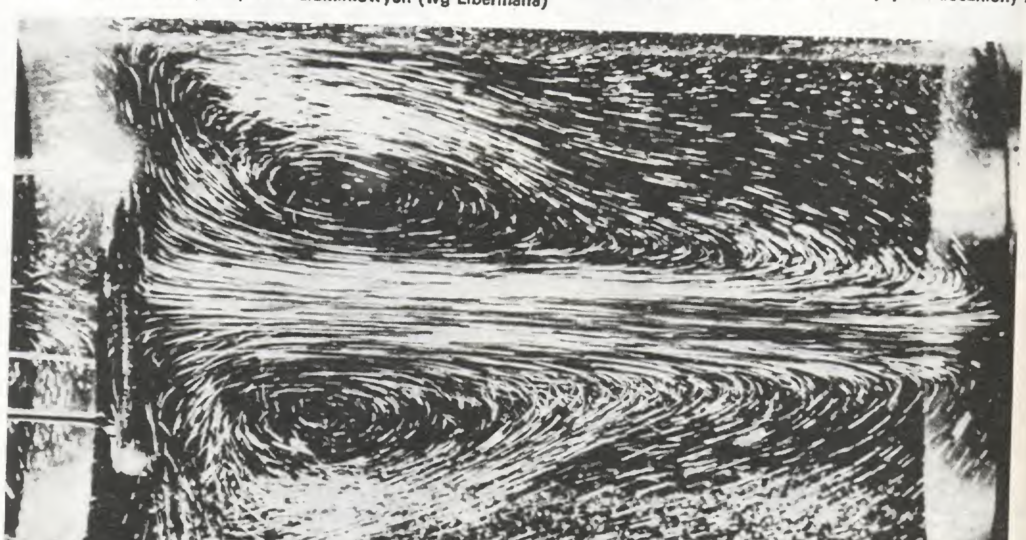
a

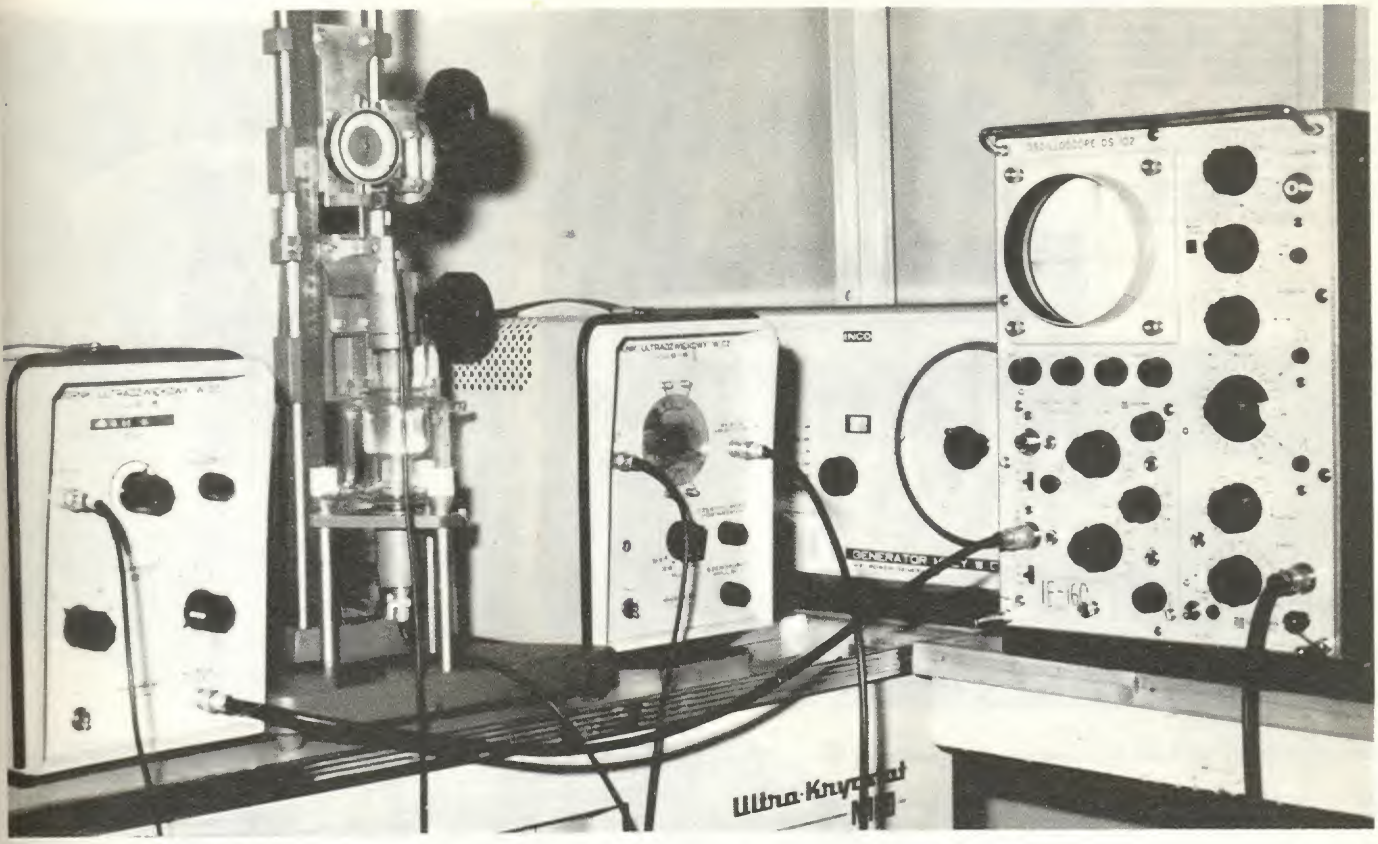
b

162. Fontanna oleju parafinowego wyrzucona przez ciśnienie promieniowania fali ultradźwiękowej promieniowanej przez przetwornik kwarcowy (wg Bergmana)

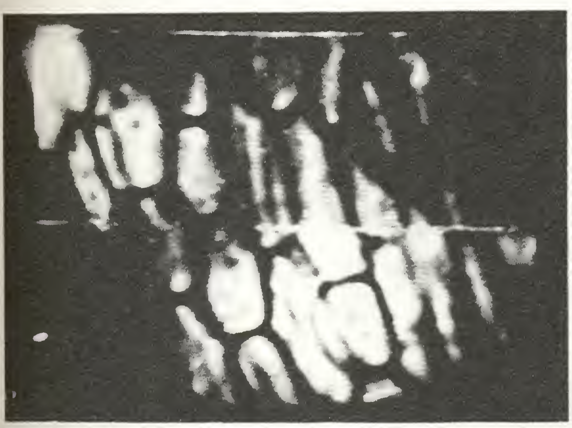


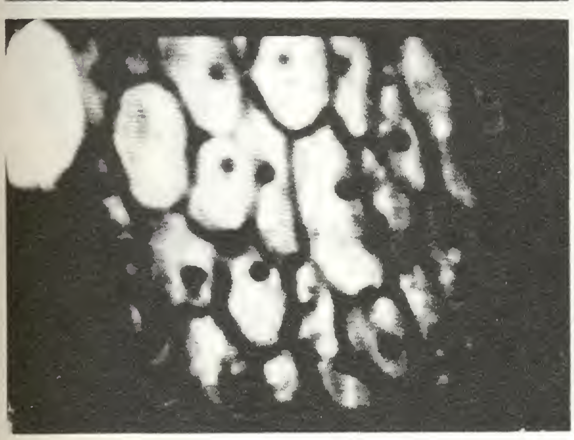
163. Przepływ akustyczny wytwarzany przez falę ultradźwiękową o dużym natężeniu („wiatr akustyczny") uwidoczniony za pomocą bardzo drobnych opiłków aluminiowych (wg Libermana)



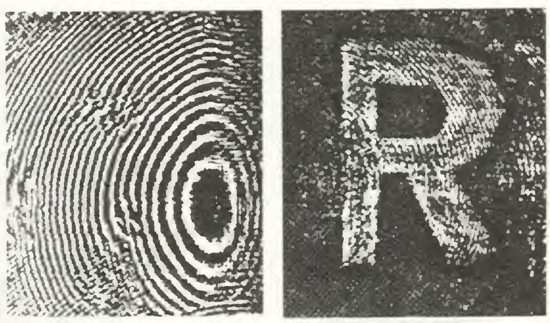


164. Układ do pomiarów prędkości i tłumienia fal ultradźwiękowych w cieczach metodą interferometrii impulsowo-fazowej (IPPT-PAN, Warszawa)

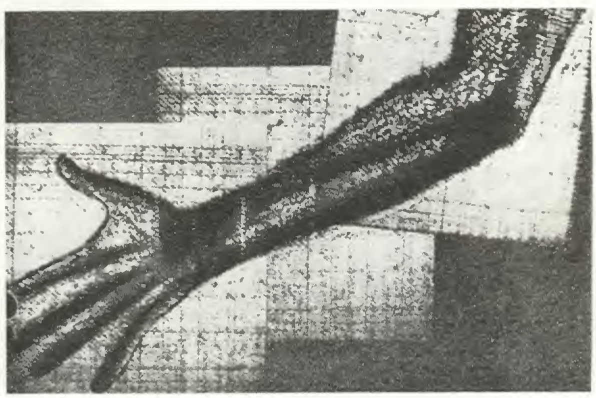

a


b

165. Obrazy mikroskopowe odcinka płatka cebuli: a) otrzymane za pomocą mikroskopu świetlnego, b) otrzymany za pomocą mikroskopu ultradźwiękowego (częstość ultradźwięków 220 MHz, poziomy wymiar pola widzenia 750 μm)



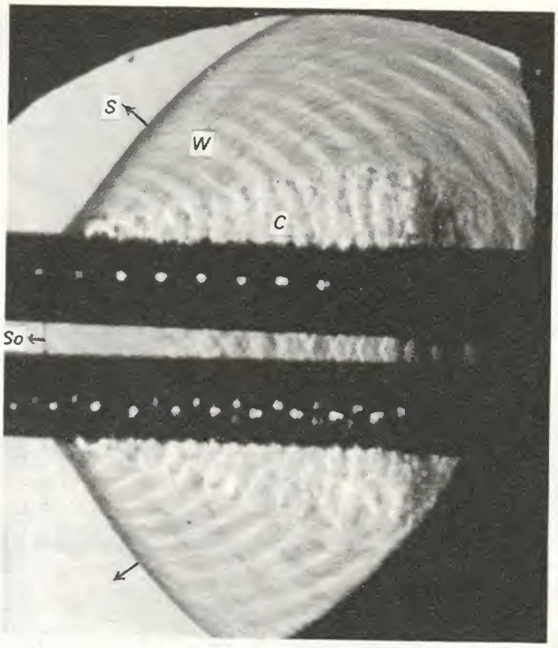
166. Hologram i odtworzony obraz płytki metalowej w kształcie litery R

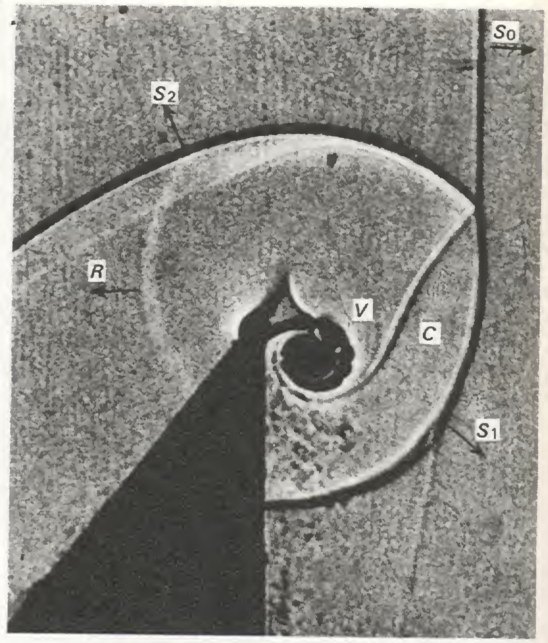


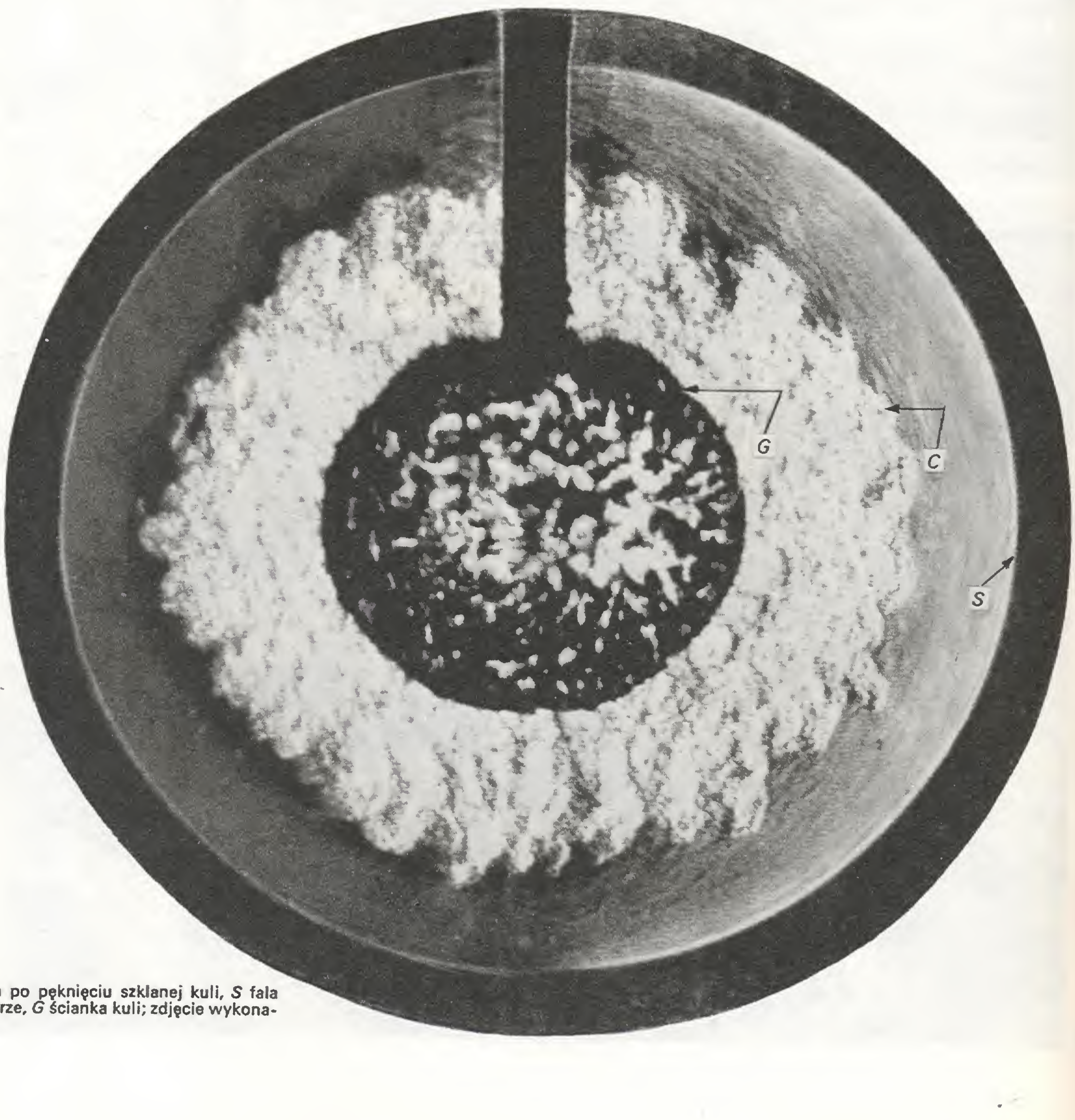
167. Obraz ręki ludzkiej uzyskany z hologramu otrzymanego i odtworzonego metodą elektronową

168. Płaska fala uderzeniowa S_0 biegnąca perforowanym przewodem; S zakrzywiona fala uderzeniowa; C front gazu, W zaburzenie akustyczne; zdjęcie wykonane metodą smug

169. Odbicie i dyfrakcja płaskiej fali uderzeniowej; S_0 fala płaska, S_1 fala zakrzywiona, S_2 fala odbita, C powierzchnia nieciągłości, V wir, R fala rozrzedzeniowa; zdjęcie wykonane metodą cieni





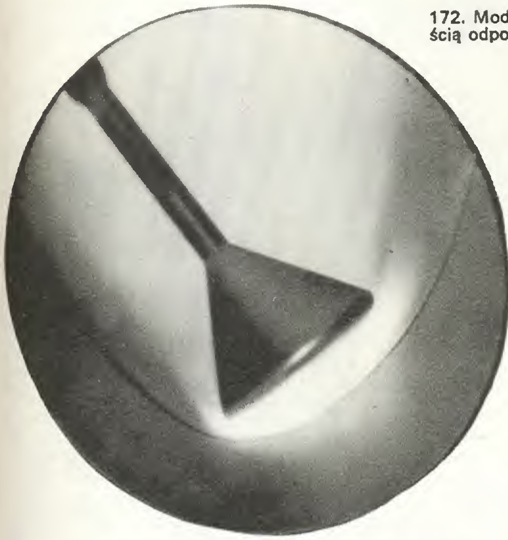


170. Struktura przepływu powietrza po pęknięciu szklanej kuli, S fala uderzeniowa, C wypływające powietrze, G ścianka kuli; zdjęcie wykonane metodą smug

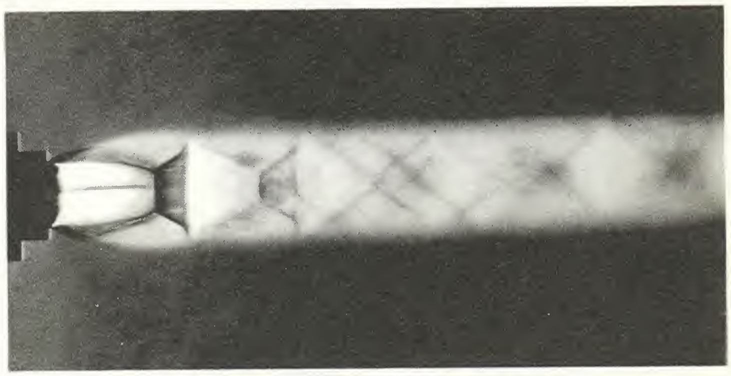


171. Skośne fale uderzeniowe w kanale; zdjęcie wykonane przy użyciu interferometru Macha–Zehndera

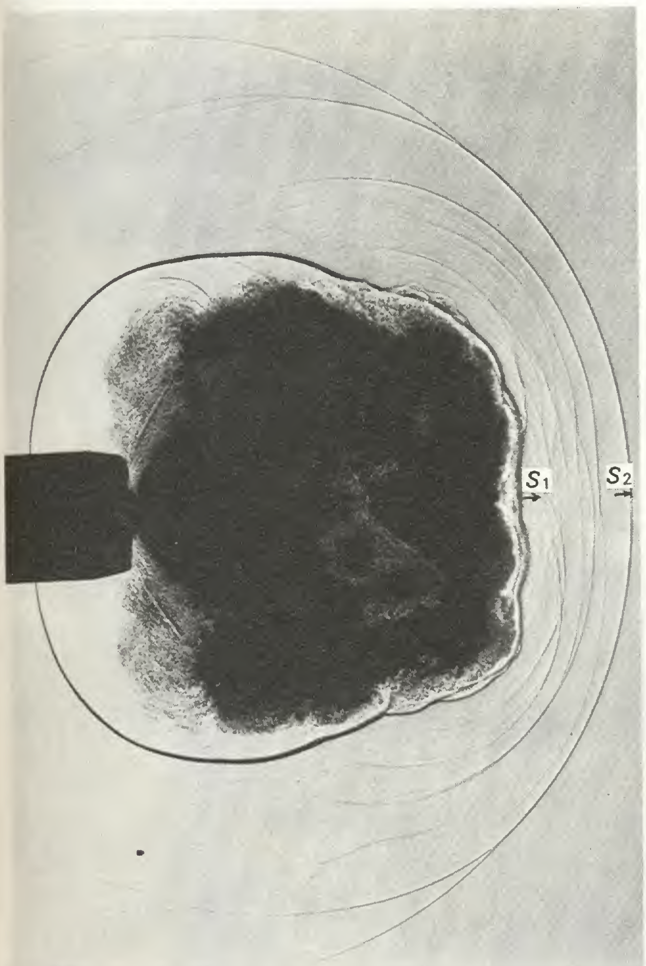
172. Modelowe badanie lotu kabiny kosmicznej w atmosferze z prędkością odpowiadającą liczbie Macha 10,2 (metoda smug, światło laserowe)



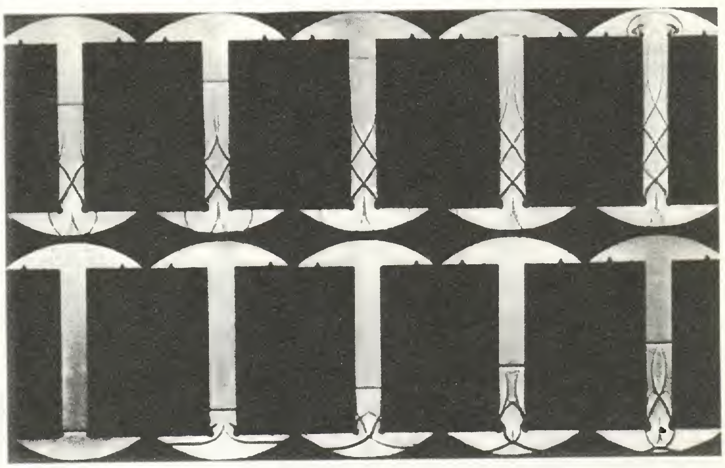
173. Płaska fala uderzeniowa zwana dyskiem Macha w strumieniu gazu wypływającym z dyszy; zdjęcie wykonane metodą smug



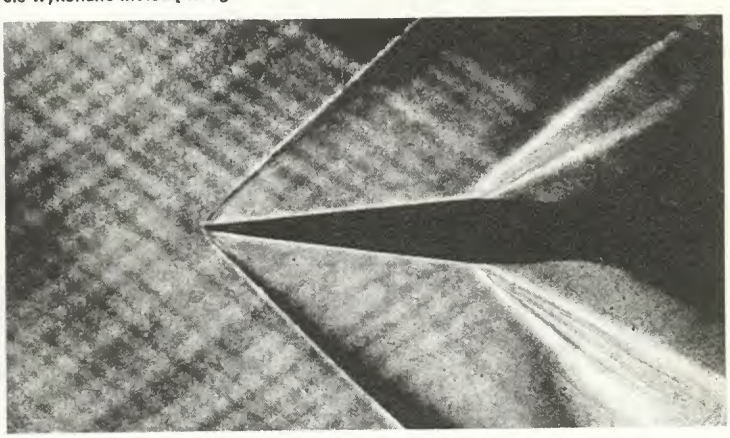
174. Wypływ gazów prochowych z lufy; S_1 fala uderzeniowa wychodząca z lufy przed pociskiem, S_2 fala uderzeniowa wywołana wypływem gazów prochowych; zdjęcie wykonane metodą smug

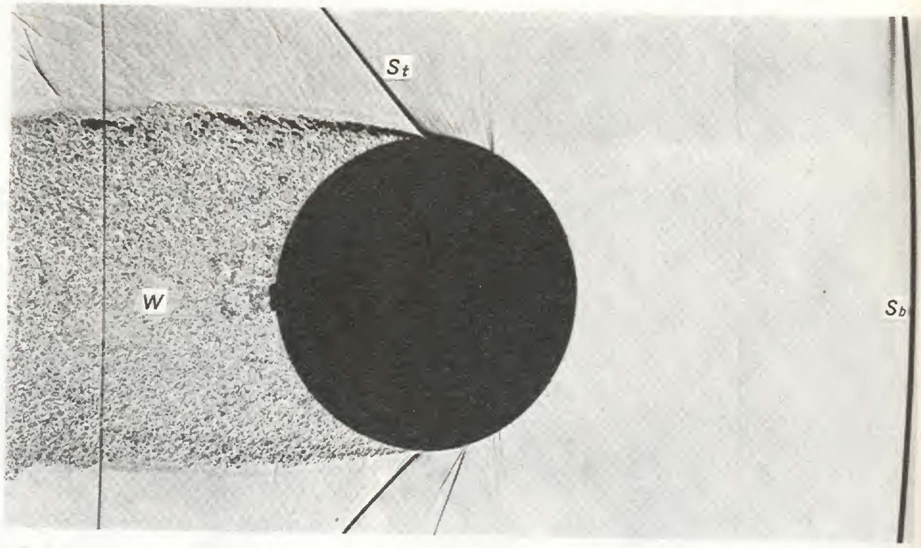


175. Zmiany struktury przepływu spowodowanego przejściem fali uderzeniowej przez kanał



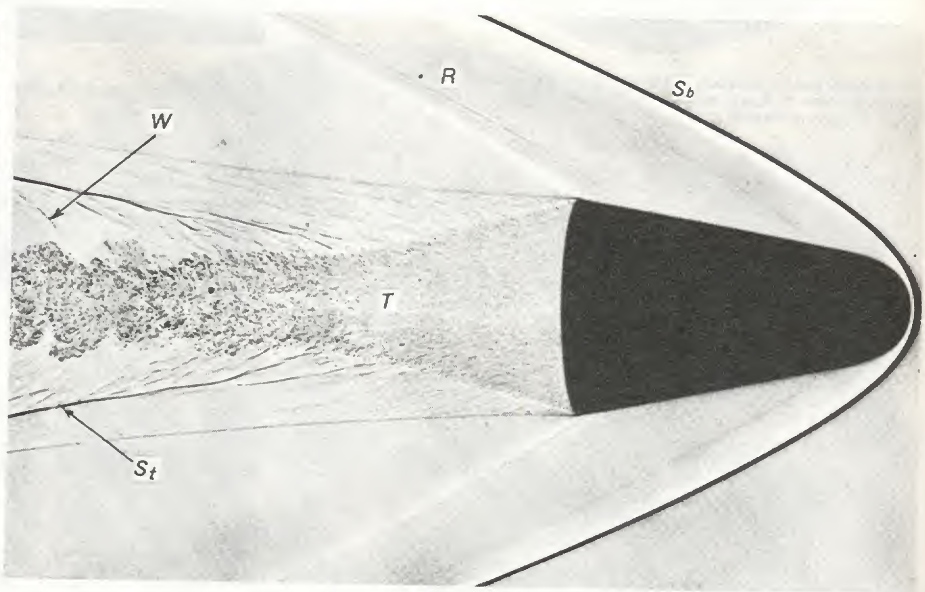
176. Opływ romboidalnego profilu z prędkością odpowiadającą liczbie Macha 1,7; zdjęcie wykonane metodą smug

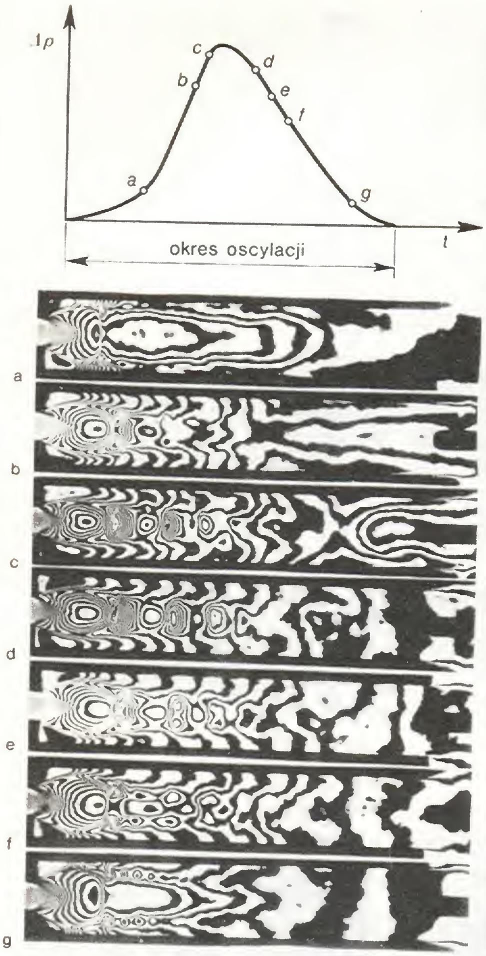




177. Kula poruszająca się w powietrzu z prędkością odpowiadającą liczbie Macha 1,05: S_b odsunięta fala uderzeniowa, S_t stożkowa fala uderzeniowa; *W* burzliwy obszar zastoju; zdjęcie wykonane metodą cieni

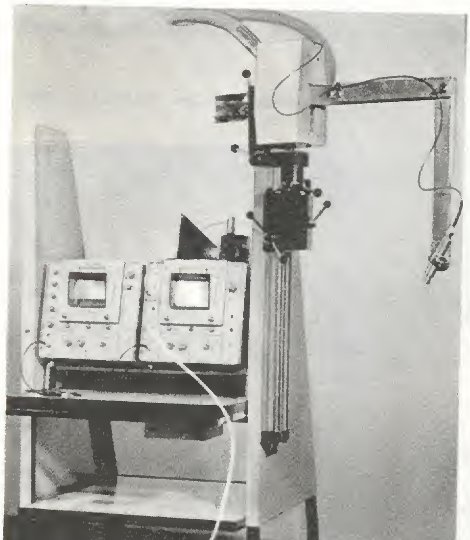
178. Zaokrąglony stożek poruszający się w powietrzu z prędkością odpowiadającą liczbie Macha 3,4; S_b odsunięta fala uderzeniowa, *R* fala rozrzedzeniowa, *T* obszar zastoju, S_t stożkowa fala uderzeniowa, *W* fale akustyczne; zdjęcie wykonane metodą cieni



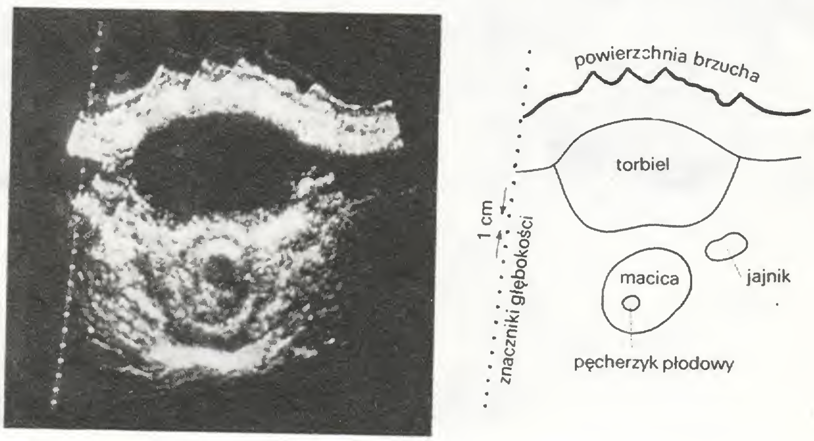


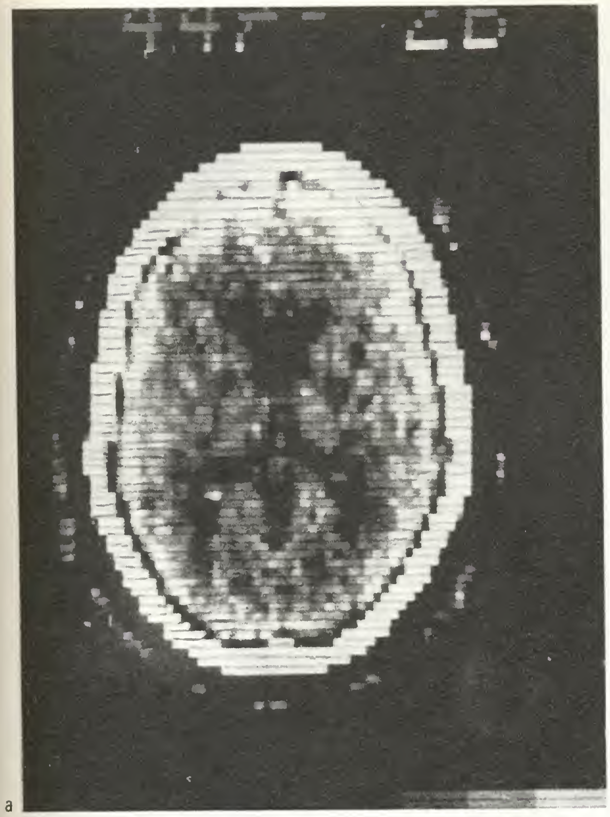
179. Przebieg ciśnienia w czasie (u góry) w obszarze zastoju za dyszą i odpowiadające mu struktury przepływu pulsującego w kanale; zdjęcie wykonane przy użyciu interferometru Macha–Zehndera

180. Oftalmograf ultradźwiękowy konstrukcji L. Filipczyńskiego i współpracowników (IPPT PAN Warszawa)



181. Ultradźwiękowy obraz dziesięciotygodniowej ciąży z torbielą w prezentacji *B* (przekrój poprzeczny nad spojeniem łonowym)



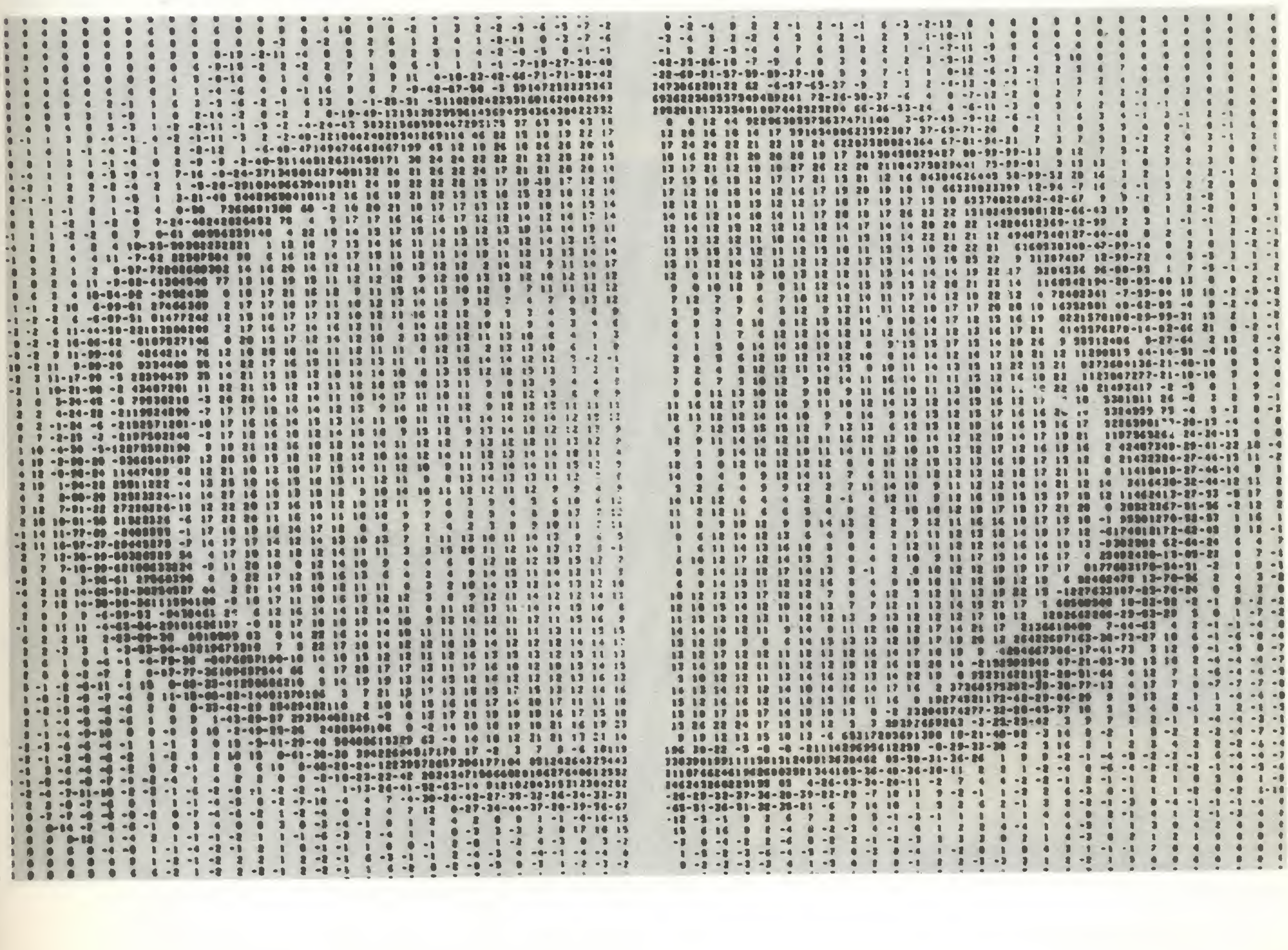


a



182. Tomograf komputerowy EMI-CT 5000 (fragament). Pacjent jest ułożony w bramce skaningowej

183. Jeden z przekrojów mózgu u zdrowego człowieka: a) Zdjęcie wykonane za pomocą skenera rentgenowskiego. Widać (bez wprowadzania środka cieniującego) obie komory boczne oraz szarą i białą substancję. b) To samo zdjęcie po obróbce komputerowej (wymiar poziomy jest rozciągnięty w stosunku do zdjęcia a). Liczby odpowiadają wartościom współczynnika pochłaniania promieni *X* przez tkankę w miejscach, przez które przechodzą promienie

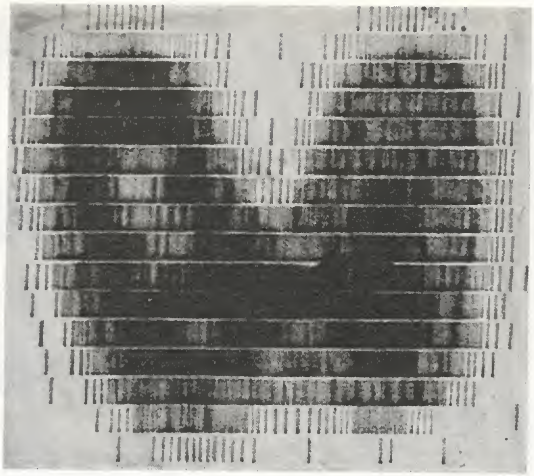


b

a

b

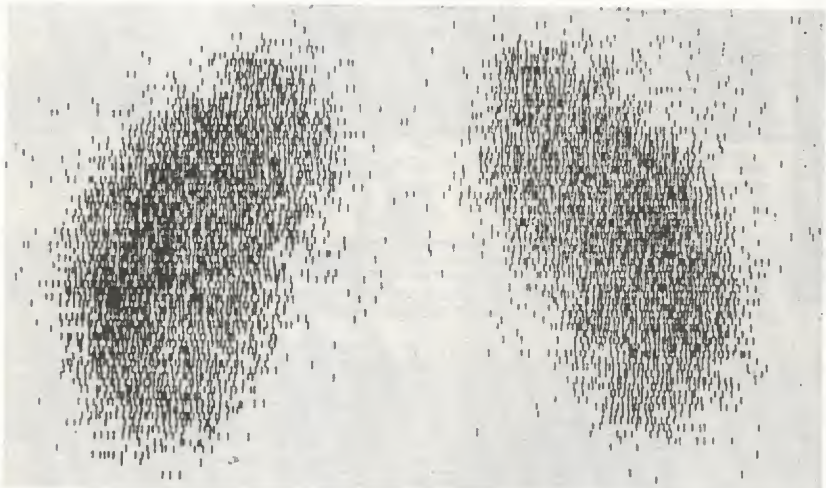




184. Scyntygramy tarczycy: a) Wykonany scyntykamerą po podaniu ^{99m}Tc. Widać guz „gorący" w tarczycy oraz zarys dolnej części głowy, szyi i ramion, b) Wykonany scyntygrafem po podaniu Na ^{131}J. Widać miejsca większej i mniejszej aktywności wewnątrz gruczołu

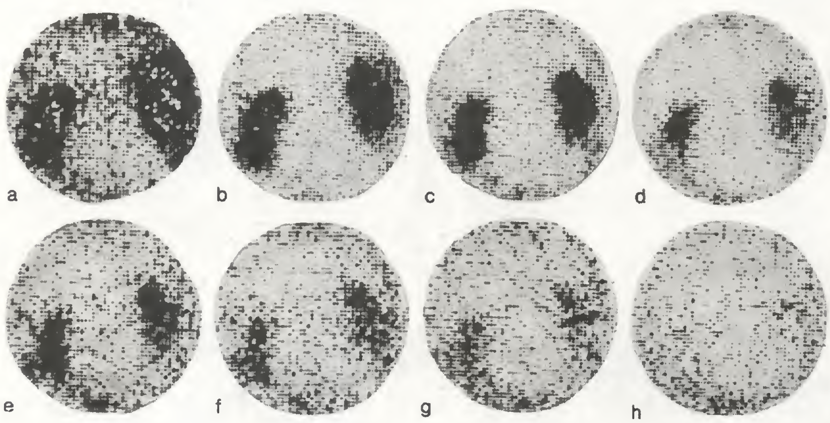


185. Licznik scyntylacyjny do pomiaru in vivo jodochwytności tarczycy



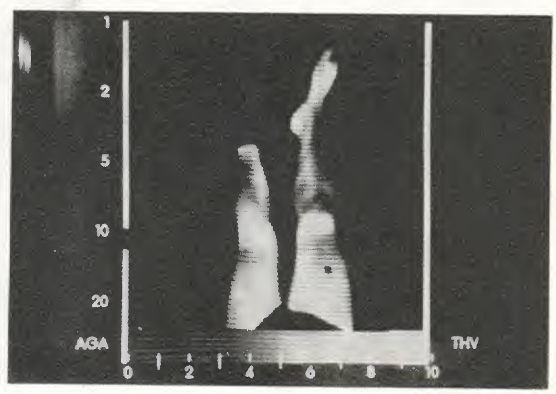
187. Scyntygram kreskowy otrzymany za pomocą scyntygrafu. Przedstawia prawidłowe rozmieszczenie w nerkach wprowadzonego związku (hipuranu) znakowanego ^{131}J

188. Scyntygramy otrzymane za pomocą scyntykamery na ekranie oscyloskopu. Przedstawiają kolejne obrazy (a–h) rozmieszczenia w nerkach hipuranu znakowanego ^{131}J, rejestrowane co 3 min. od jego podania. Widać nierównomierne wypełnianie się nerek

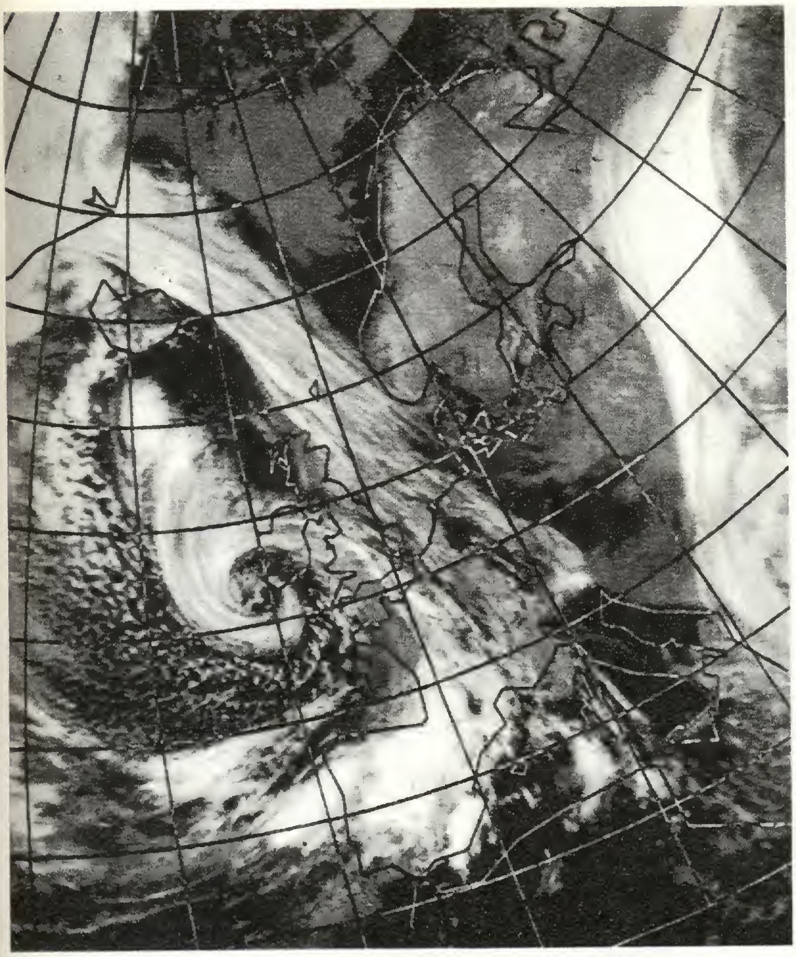




186. Efekt oziębiania się palców przy paleniu papierosów u pacjenta z chorobą Raynauda: a) przed paleniem palce są ciepłe, b) po zapaleniu jednego papierosa niektóre palce są zimne (brak zarysu palców)



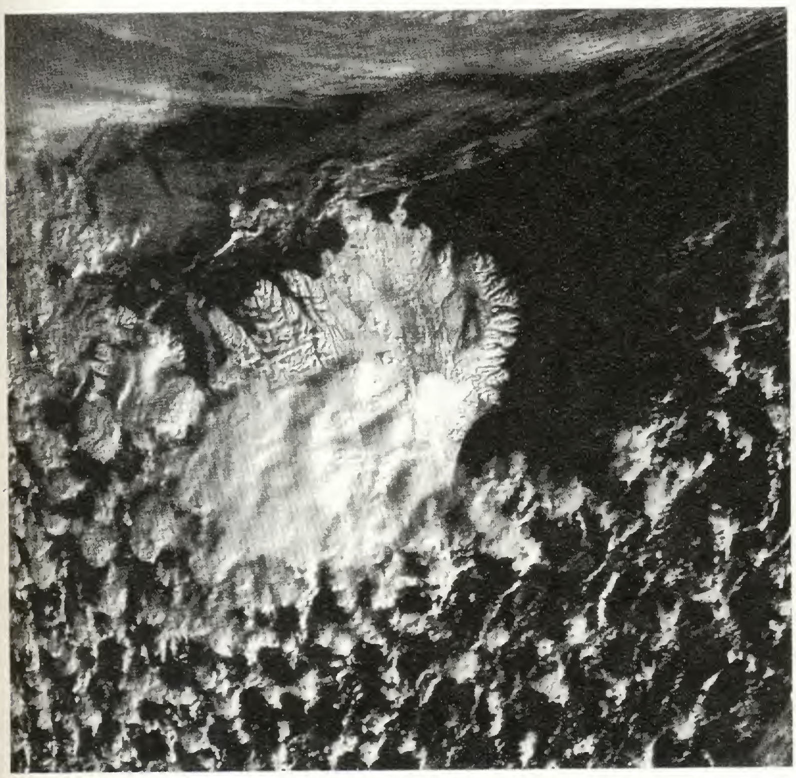
189. Termogram nóg chorego na chorobę Bürgera. Niedokrwienie części jednej nogi ujawnia się jako jej brak na zdjęciu



190. Fotografia satelitarna rozwiniętego układu cyklonalnego. Spiralne struktury chmurowe związane są z położeniem frontów

191. Radiosonda produkcji fińskiej firmy „Vaisala" (wg Vaisala News) — jeden z kilkunastu typów radiosond używanych przez służby meteorologiczne całego świata. Unoszona przez balon na wysokości sięgającej 35 km przekazuje drogą radiową informacje o ciśnieniu, temperaturze i wilgotności na trasie swojego lotu. Śledząc jej położenie można uzyskać dane o rozkładzie wiatrów na różnych wysokościach. Pomiary położenia radiosondy wykonuje się optycznie, radiolokacyjnie lub (w najnowszych rozwiązaniach) metodami radionawigacyjnymi. Do zapewnienia właściwych informacji dla nowoczesnej numerycznej prognozy pogody stacje radiosondażowe powinny być rozmieszczone co 300–400 km. Jak dotąd tylko niektóre obszary świata mają sieć aerologiczną o takiej gęstości. Czynione są wysiłki by informacje uzyskiwane dotychczas przez radiosondy otrzymywać z satelitarnych pomiarów promieniowania w różnych wycinkach widmowych





192. Zdjęcie satelitarne mezoskalowego układu chmur. Choć przypomina ono chmury kłębiasto-warstwowe, rozmiary poszczególnych „kłębków" są rzędu kilkudziesięciu kilometrów. Każdy kłębek zawiera wiele chmur kłębiastych (*cumulus*), kłębiasto-warstwowych (*stratocumulus*)



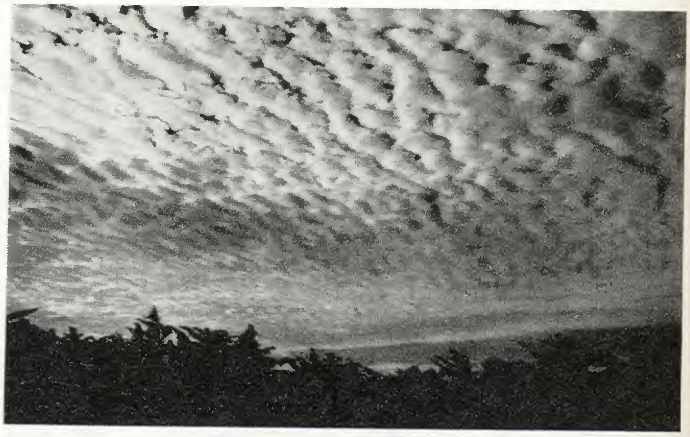
a



b



c



d



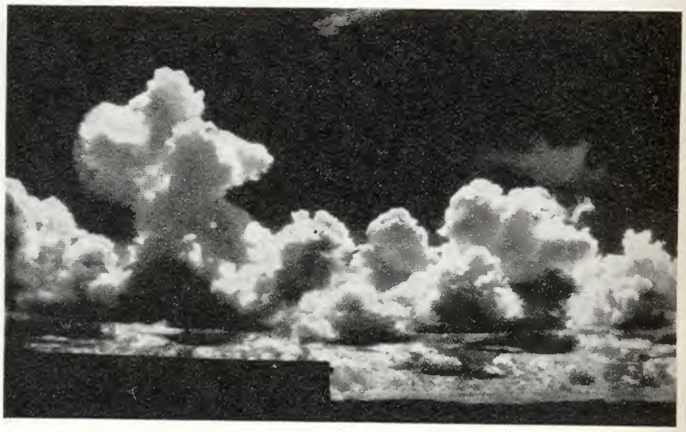
e



f



g



h

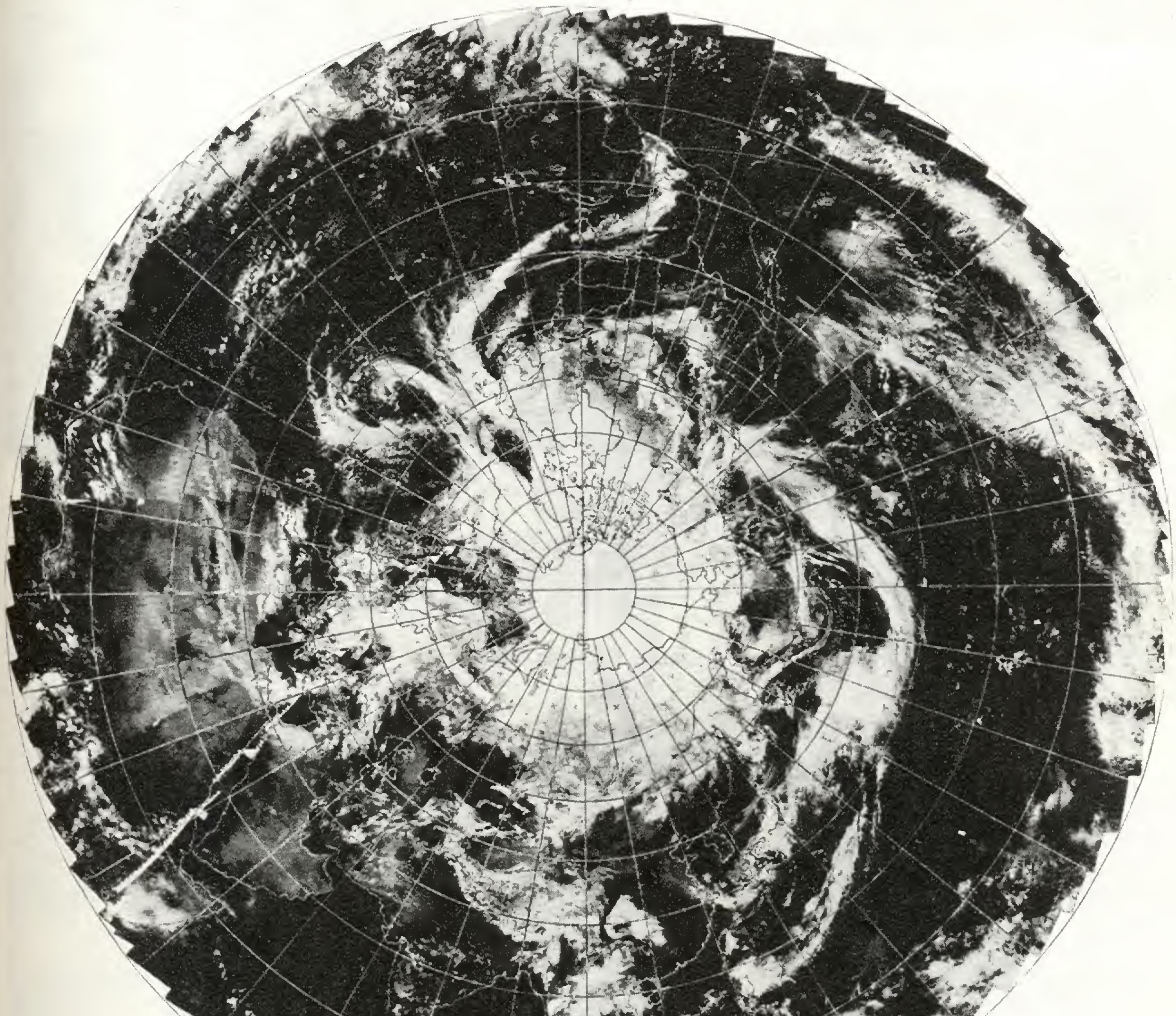
i



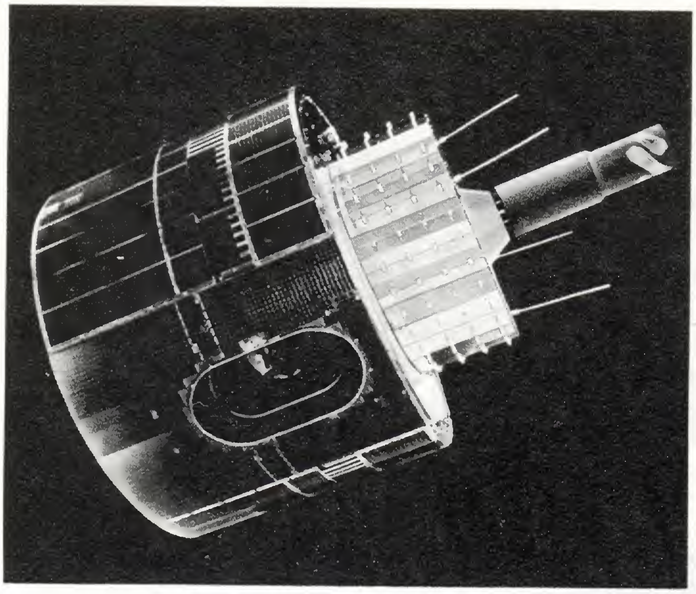
j



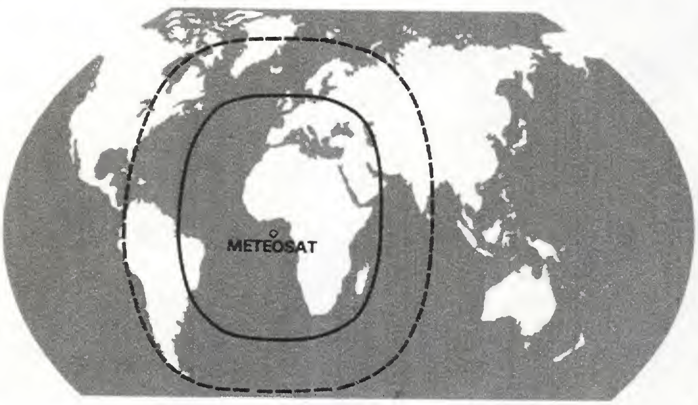
193. Główne rodzaje chmur: a) cirrus, b) cirrocumulus, c) cirrostratus nebulosus, d) altocumulus, e) niskie strzępy chmur „złej pogody" na tle chmur altostratus, f) stratocumulus, g) stratus, h) i i) cumulusy, j) cumulonimbus



194. Mozaika zestawiona ze zdjęć wykonanych przez satelitę Essa 9 poruszającego się na orbicie polarnej w dniu 3 IV 1970



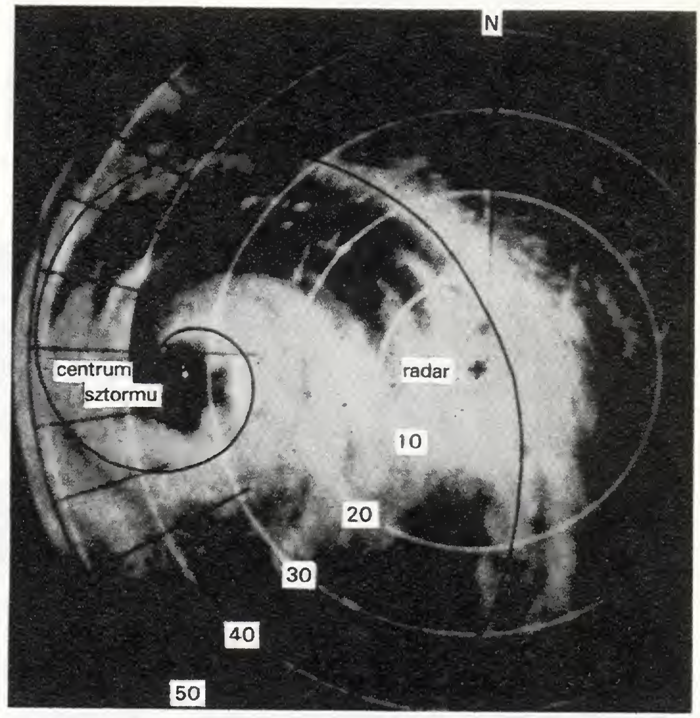
a



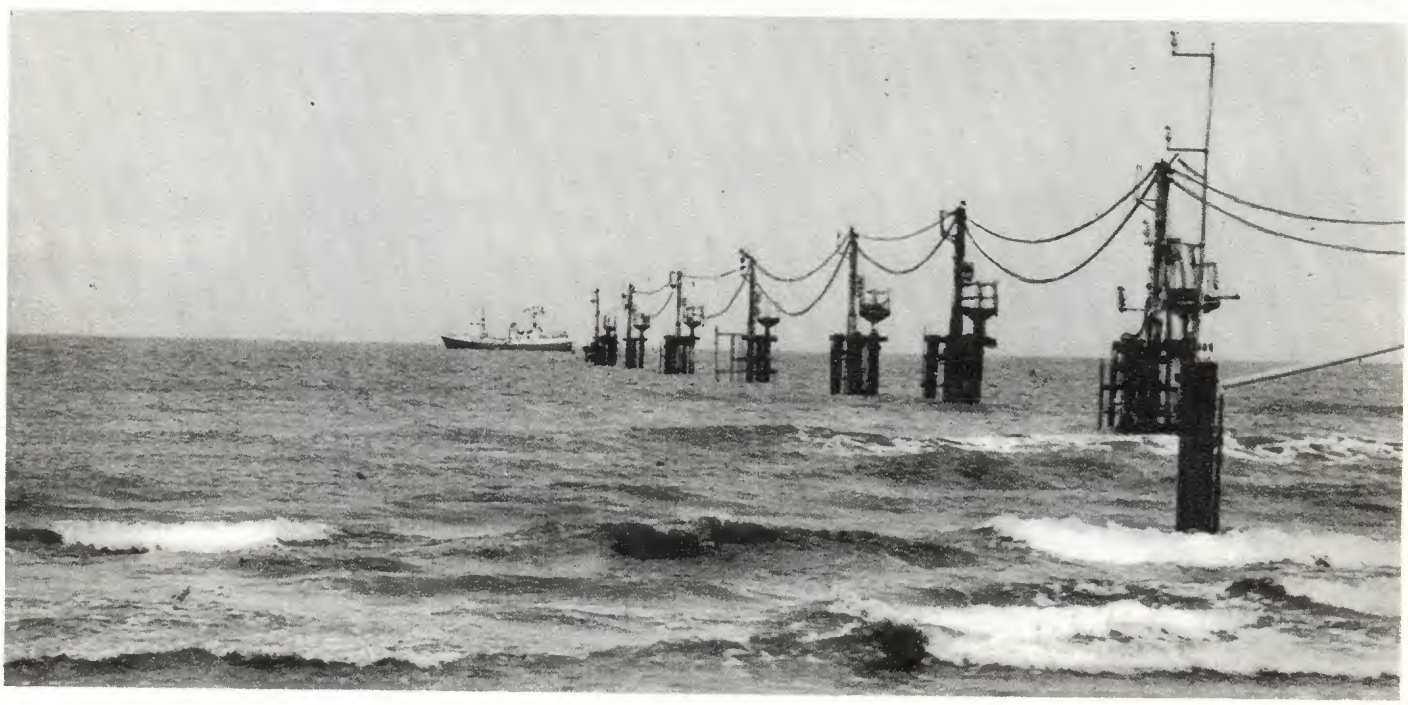
b

195. Geostacjonarny satelita meteorologiczny Meteosat wykonany w wyniku współpracy państw Europy Zachodniej. Mapka ukazuje zasięg jego obserwacji (linia ciągła) i zasięg łączności (linia przerywana). Wyposażenie satelity umożliwia rozróżnianie obiektów o średnicy 2,5 km w świetle widzialnym i o średnicy 5 km w podczerwieni (pasma 5,7–7,1 μm i 10–12,5 μm). Obrazy satelitarne są przetwarzane przez centralną stację odbiorczą i rozsyłane użytkownikom. W uproszczonej postaci mogą być odbierane bezpośrednio z satelity przez wszystkie służby meteorologiczne dysponujące nieskomplikowanymi stacjami odbiorczymi w zasięgu nadajników satelity. Satelita jest także przystosowany do zbierania informacji z naziemnych automatycznych stacji meteorologicznych

196. Fotografia układu chmur cyklonu tropikalnego na wskaźniku panoramicznym. Koncentryczne kręgi — linie jednakowej odległości poziomej podane w milach morskich

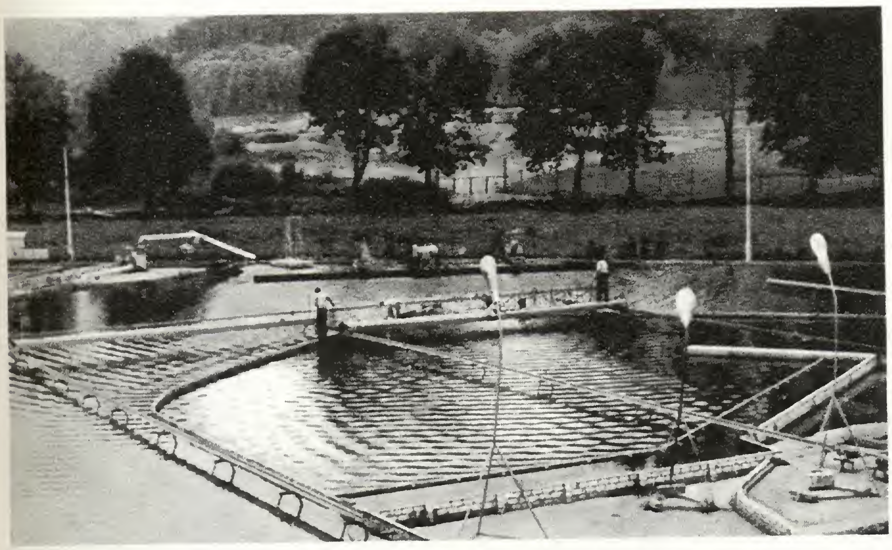


197. Morska baza doświadczalna IBW-PAN w Lubiatowie





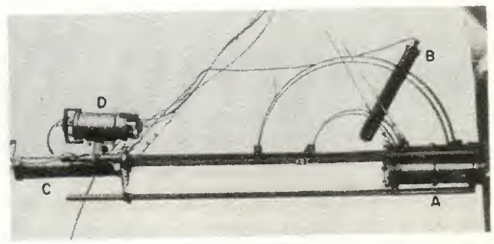
198. Falowanie wiatrowe



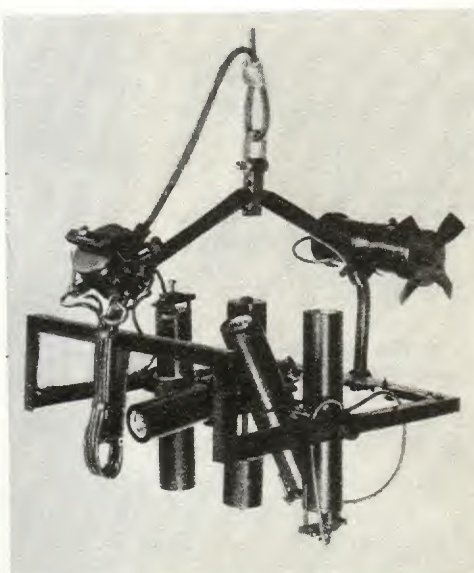
199. Hydrauliczne modelowanie falowania wiatrowego w bazie badawczej IBW-PAN w Oliwie

200. Falowanie sztormowe w strefie przybrzeżnej





201. Sonda miernika funkcji rozpraszania światła w morzu, w której źródłem światła jest laser (G. Kullenberg 1968 r.); *A* laser, *B* ruchomy odbiornik światła rozproszonego pod dużymi kątami, *C* odbiornik światła rozproszonego pod małymi kątami (1–3,5°), *D* złącza przewodów

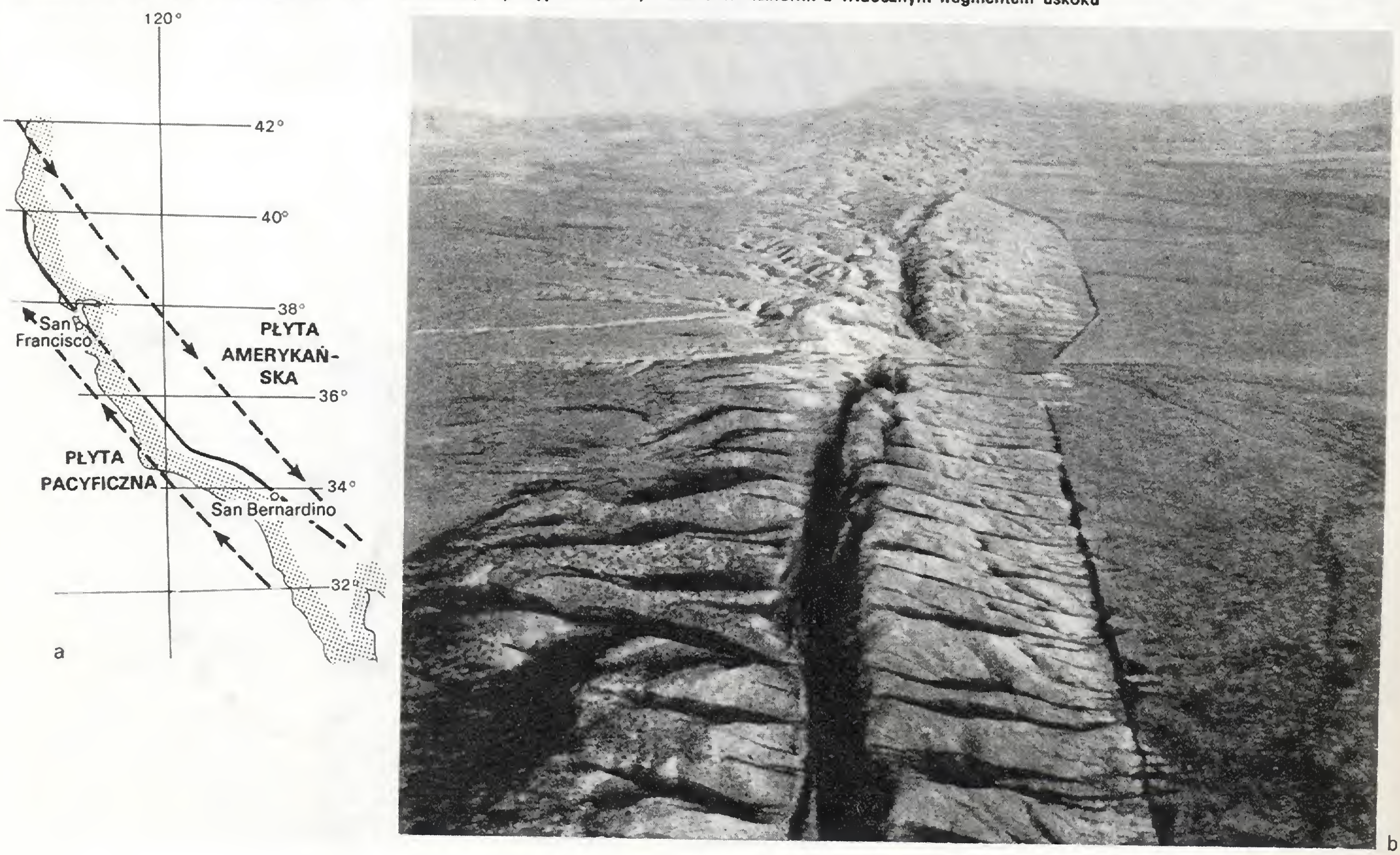


202. Sonda morska z zestawem mierników radiacji, przeznaczona do pomiarów strumieni światła przychodzącego z różnych kierunków w morzu. Na zdjęciu widoczne są: 4 mierniki radiacji ustawione w różnych kierunkach, motor z turbiną do obracania układu w przestrzeni wodnej, pojemnik na połączenia przewodów, konstrukcja nośna oraz zawieszenie na kablu doprowadzającym wyposażonym jednocześnie w linę nośną (N.G. Jerlov 1968 r.)

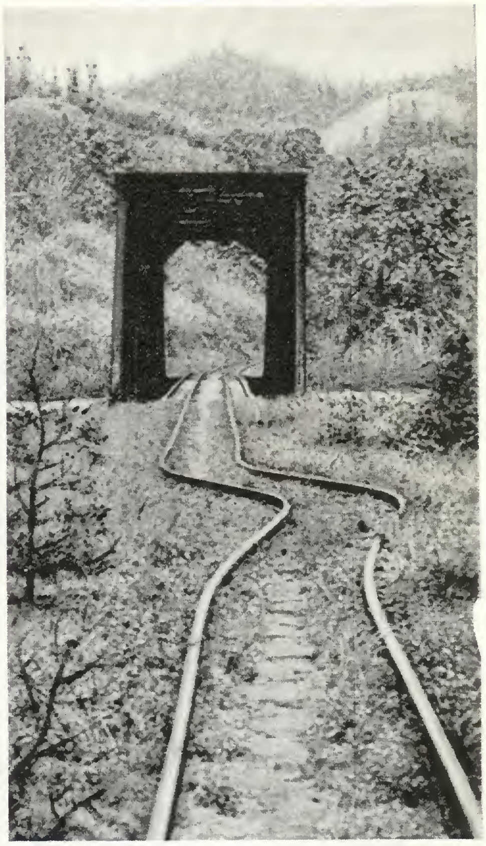
203. Pęknięcie uskokowe w Mongolii, powstałe w wyniku trzęsienia ziemi 4 grudnia 1957 r.



204. Uskok San Andreas ciągnący się wzdłuż granicy między Płytą Pacyficzną i Płytą Amerykańską: a) mapka Kalifornii, na której zaznaczony jest uskok i kierunek przesuwania się Płyty Amerykańskiej i Płyty Pacyficznej; b) zdjęcie równiny Carizzo w Kalifornii z widocznym fragmentem uskoku







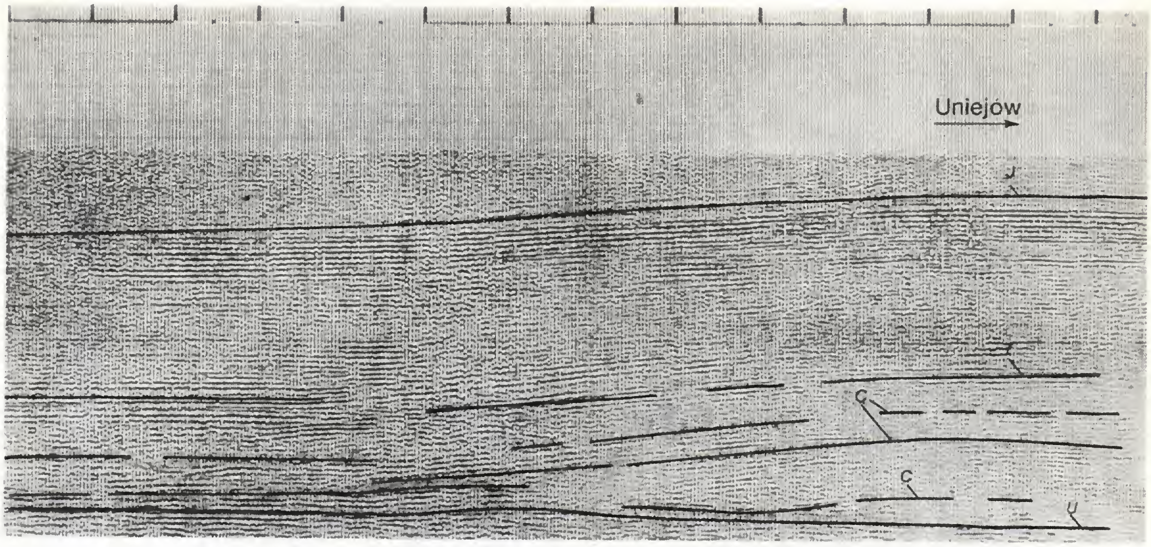
a
b

205. W czasie trzęsienia ziemi może nastąpić wyraźne przesunięcie powierzchniowych warstw skorupy ziemskiej: a) jedna z ulic San Francisco po słynnym trzęsieniu ziemi w 1906 r.; b) szyny kolejowe w pobliżu Puerto Banios (Gwatemala) po trzęsieniu ziemi w lutym 1976 r.

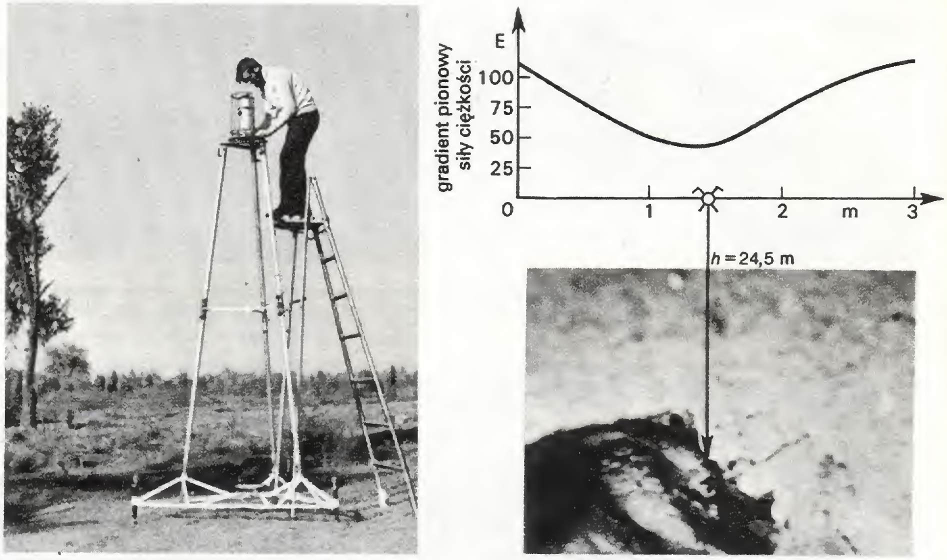


206. Zdjęcie lotnicze gaju cytrusowego zrobione po trzęsieniu ziemi w 1940 r. w Imperial Valley. Wyraźnie widać przesunięcie drzew w gaju.

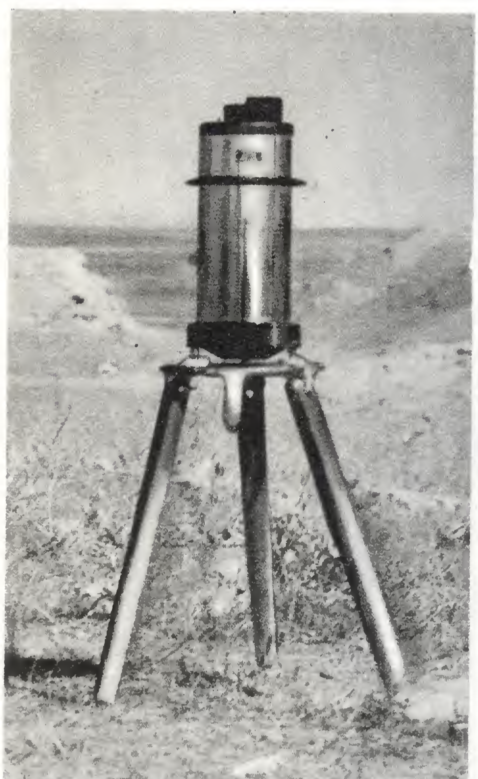
207. Sejsmiczny przekrój czasowy i jego związek z budową geologiczną synklinarium łódzkiego w rejonie Uniejowa; *J* jura, *T* trias, *C* cechsztyn, *U* utwory podsolne



208. Pomiary wartości gradientu pionowego siły ciężkości $\partial g/\partial z$ za pomocą grawimetru. Z lewej strony wieża pomiarowa, u dołu wykryta forma antropogeniczna (wyrobisko górnicze); *h* głębokość, na której znaleziono tę formę; wyniki pomiarów podane są na wykresie w etweszach E, 1 E = 1 (cm/s²)/cm



209. Grawimetr — przyrząd do pomiarów względnych siły ciężkości



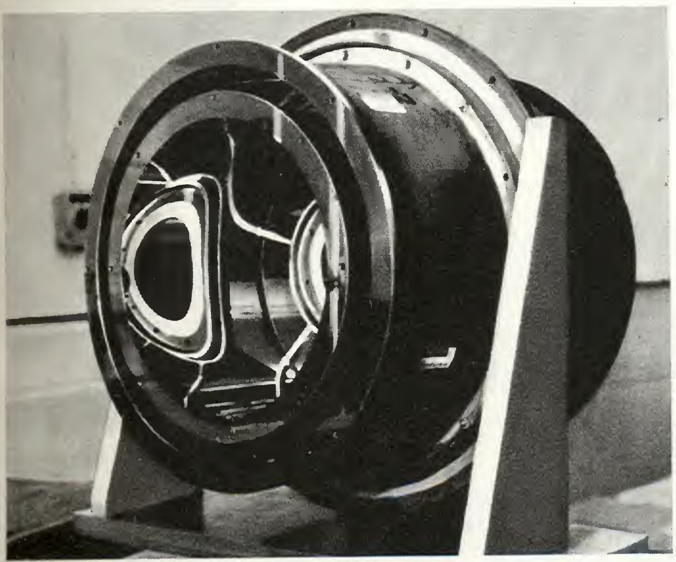
210. Przenośny magnetometr protonowy do pomiarów wartości całkowitego natężenia pola magnetycznego Ziemi, typ PMP-2





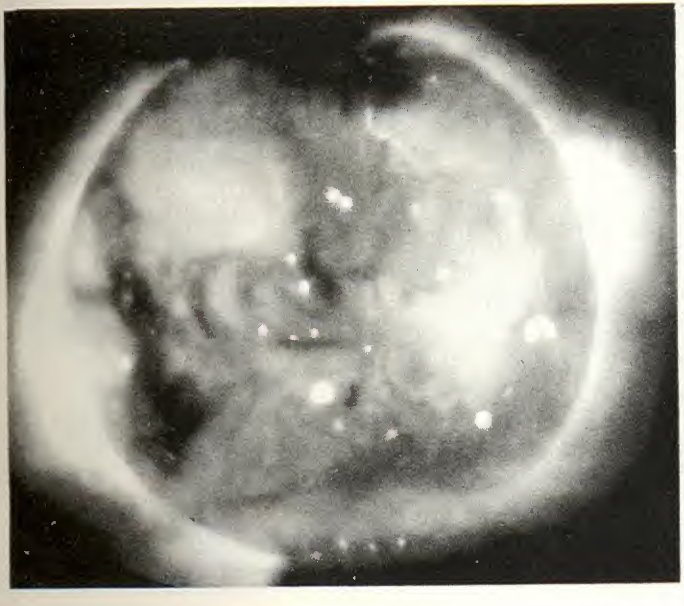


211. Fotografie Mgławicy Krab w świetle spolaryzowanym (strzałki ukazują kierunek wektora elektrycznego). Wyraźna różnica w obrazie świadczy o synchrotronowym mechanizmie emisji promieniowania



212. Ogniskujący teleskop rentgenowski. Dwa zwierciadła o średnicach 23 i 31 cm mają łączną powierzchnię zbierającą równą 44 cm²

213. Rentgenowskie zdjęcie korony słonecznej wykonane w miękkim promieniowaniu *X*





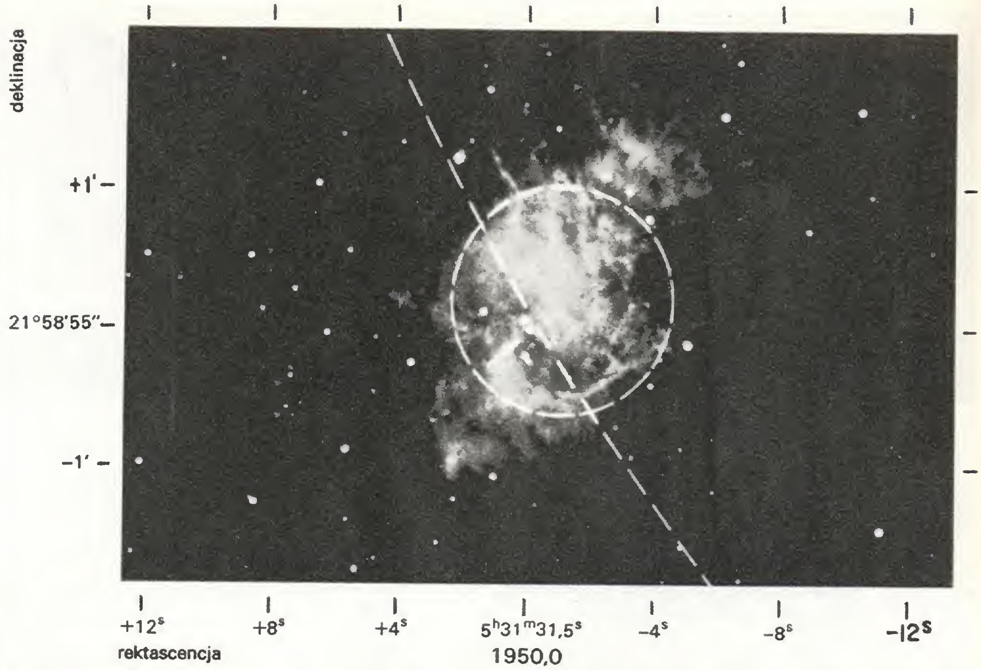
a

b

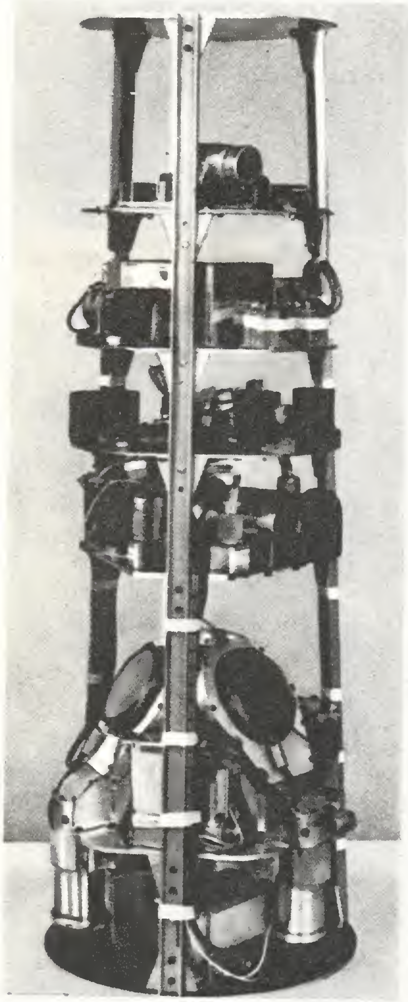


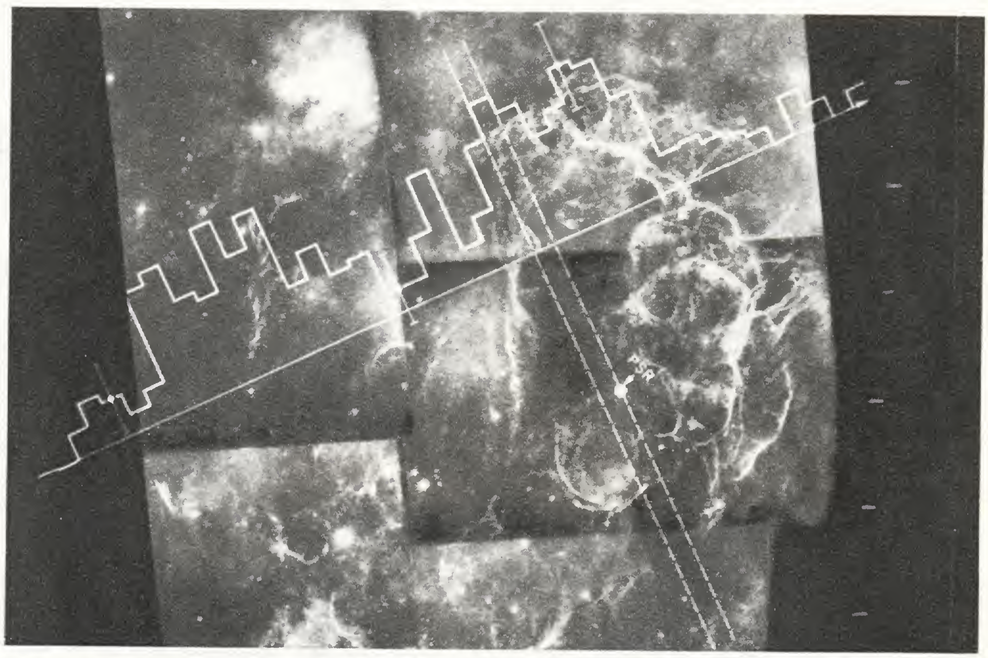
214. Teleskop do obserwacji błysków promieniowania Czerenkowa w atmosferze (u góry) i sfotografowany za jego pomocą obraz błysku. Krzyżykiem zaznaczony jest punkt największej jasności błysku

215. Mgławica Krab. Kółkiem zaznaczony jest obszar będący źródłem promieniowania *X*. Skośna linia przedstawia brzeg Księżyca w czasie zaćmienia pulsara oznaczonego strzałką



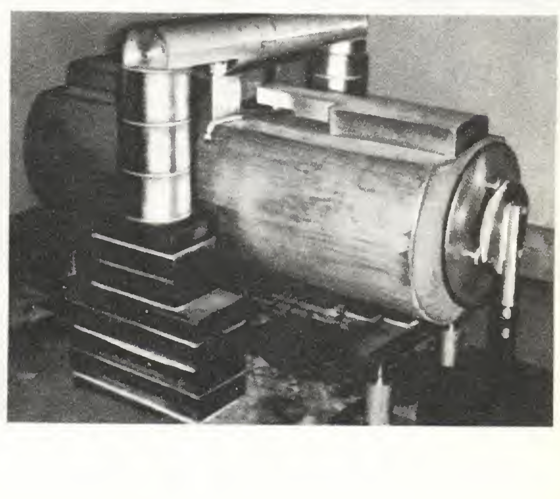
216. Odbiornik, za którego pomocą wykryto pierwsze kosmiczne źródło promieniowania *X*





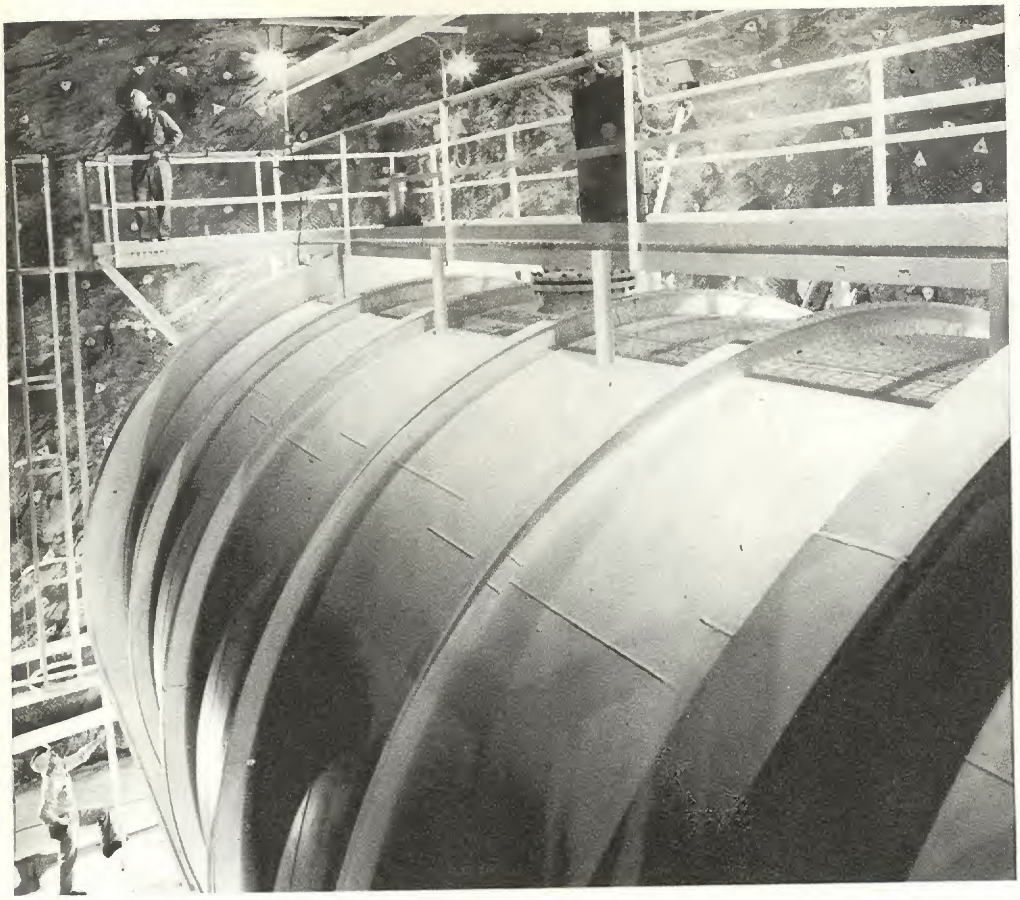
217. Jednowymiarowy rozkład promieniowania *X* mgławicy Vela X nałożony na jej zdjęcie wykonane w świetle nadfioletowym. Symbolem PSR zaznaczone jest położenie pulsara





218. Promieniowanie radiowe Wielkiej Mgławicy (M 31) w Andromedzie (względne natężenie promieniowania 1,415 MHz)

219. Antena grawitacyjna zbudowana przez Braginskiego w Moskwie

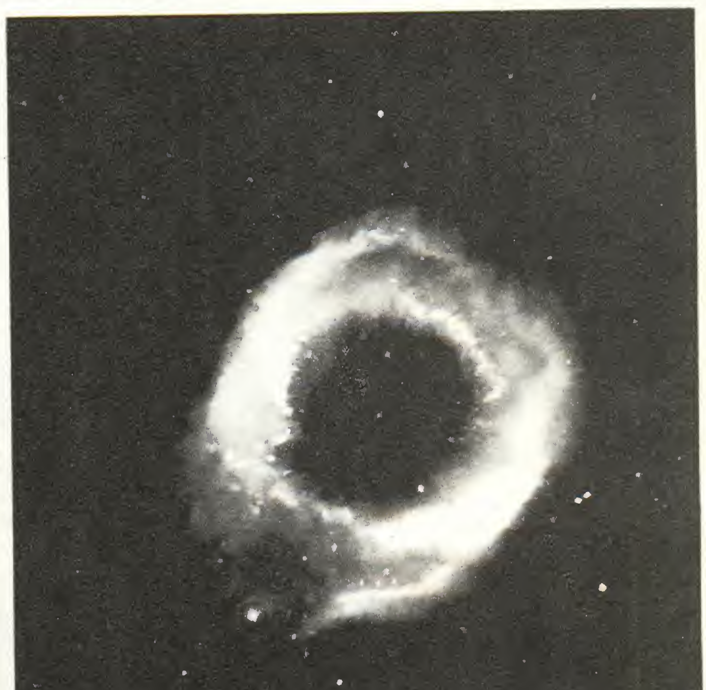


220. Podziemne obserwatorium neutrinowe Davisa — zbiornik z czterochloroetylenem zmontowany w szybie kopalni złota w Południowej Dakocie (USA)

221. Mgławica Oriona — zbiorowisko młodych gwiazd i materii międzygwiazdowej



222. Mgławica planetarna





223. Galaktyka spiralna widziana z góry



224. Galaktyka spiralna widziana z boku. Ciemna smuga biegnąca wzdłuż równika jest utworzona z obłoków pyłu



a b c

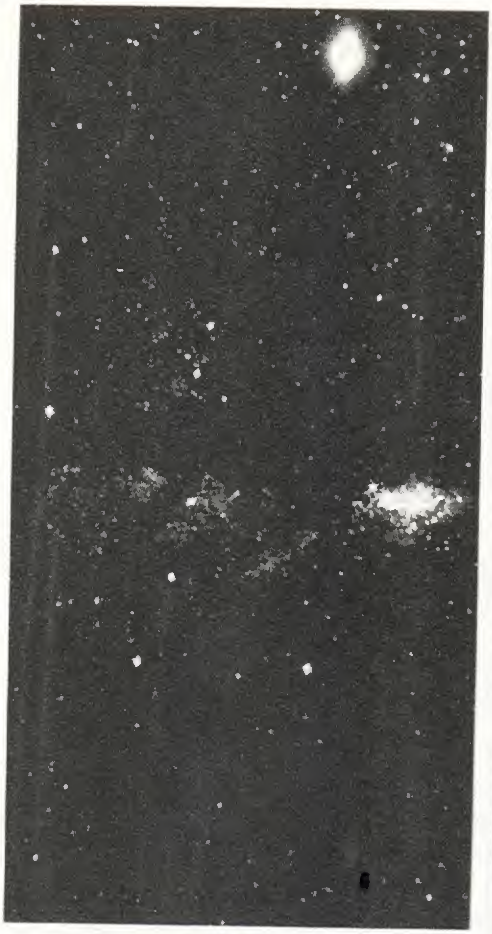
d e f

g h i

j k l

m n

225. Różne typy galatyk: nieregularna (a Mały Obłok Magellana), spiralne (b Mgławica Trójkąta, c Mgławica Andromedy, d M104); spiralne z poprzeczką (e, f, g), soczewkowate (h, i), eliptyczne (j M32, k NGC185, l M87), eliptyczne karłowate (m system Formax, n system Sekstans)



226. Zewnętrzna część Mgławicy Andromedy rozdzielona na gwiazdy. Aby określić odległość Mgławicy trzeba wśród gwiazd wyszukać cefeidy i zmierzyć ich okresy zmian blasku



a

227. Eksplozja w jądrze galaktyki M 87. Struga wypływającej materii widoczna jest przy odpowiednio dobranym czasie naświetlenia. Zdjęcie 225 I przedstawia tę samą galaktykę (naświetlano je wielokrotnie dłużej)





b

228. Jądro Mgławicy Andromedy. Zdjęcie b), które jest stukrotnie powiększonym wycinkiem środkowej częśc zdjęcia a), naświetlano wielokrotnie krócej. Maksymalne zagęszczenie gwiazd na rys. b ma średnicę zaledwie 3'' łuku



229. Centralna część gromady galaktyk Coma



230. Kwintet Stefana. Przykład ciasnej grupy wzajemnie oddziałujących galaktyk

231. Wybuch w jądrze galaktyki (galaktyka M 82). Chmury materii oddalają się od jądra z prędkościami dochodzącymi do 3000 km/s





232. Aktywna galaktyka NGC 5128 (Centaurus *A*). Sklasyfikowana jako *E2* ma nietypowy dla galaktyk eliptycznych równikowy pas ciemnej materii. W centrum galaktyki znajduje się zwarte źródło podczerwone